Projecting Dialect Distances to Geography

Bootstrapping Clustering vs. Clustering with Noise

John Nerbonne[1] Peter Kleiweg[1] Franz Manni[2]

[1]Alfa-informatica
University of Groningen

[2]Musée de l'Homme
Paris

Data Analysis, Machine Learning, and Applications:
31st Meeting, *Gesellschaft für Klassifikation*,
Freiburg, 9 March 2007

Background

Pronunciation Distance

- obtained by a modified edit distance
  - roughly equally to the minimal number of basic changes needed to obtain one phonetic transcription from another
- results for Dutch, Norwegian, German, Bulgarian, Sardinian, Gabon Bantu, ...
- validated wrt dialect speakers' judgements ($r = 0.7$)

k	ɔ	r	s	t		
k	œ		s	t	ə	
	1	1			1	3

Introduction

Dialect/Language Groups/Families
—Language varieties (sometimes) organized in groups. Even elsewhere (e.g., continua, islands), we still examine groupings to compare new results to older scholarship, which assumes groups.

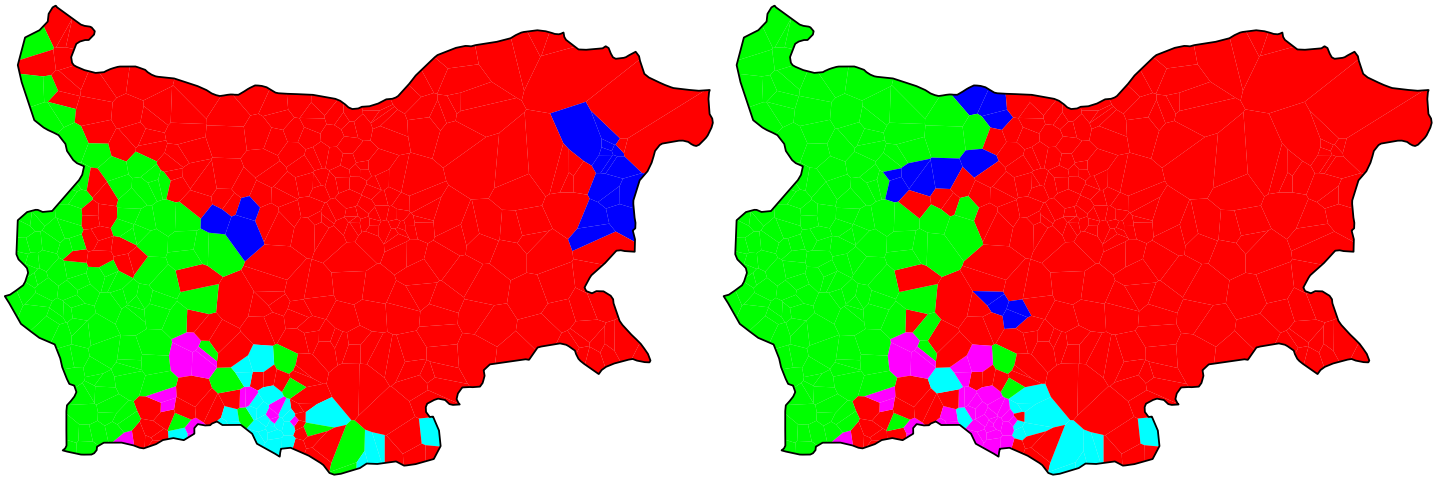
- organized hierarchically
  - Kaiserstuhl is South Badish, is Badish, is Alemannic, is Southern German, is Western Germanic, ...
- therefore we apply HIERARCHICAL CLUSTERING (not k-means, etc.)

Stability, borders, certainty

Problems

- Stability
  - slight input differences can change cluster results massively
- Hard and soft borders
- Certainty

Two Bulgarian Datasets ($r = 0.97$)

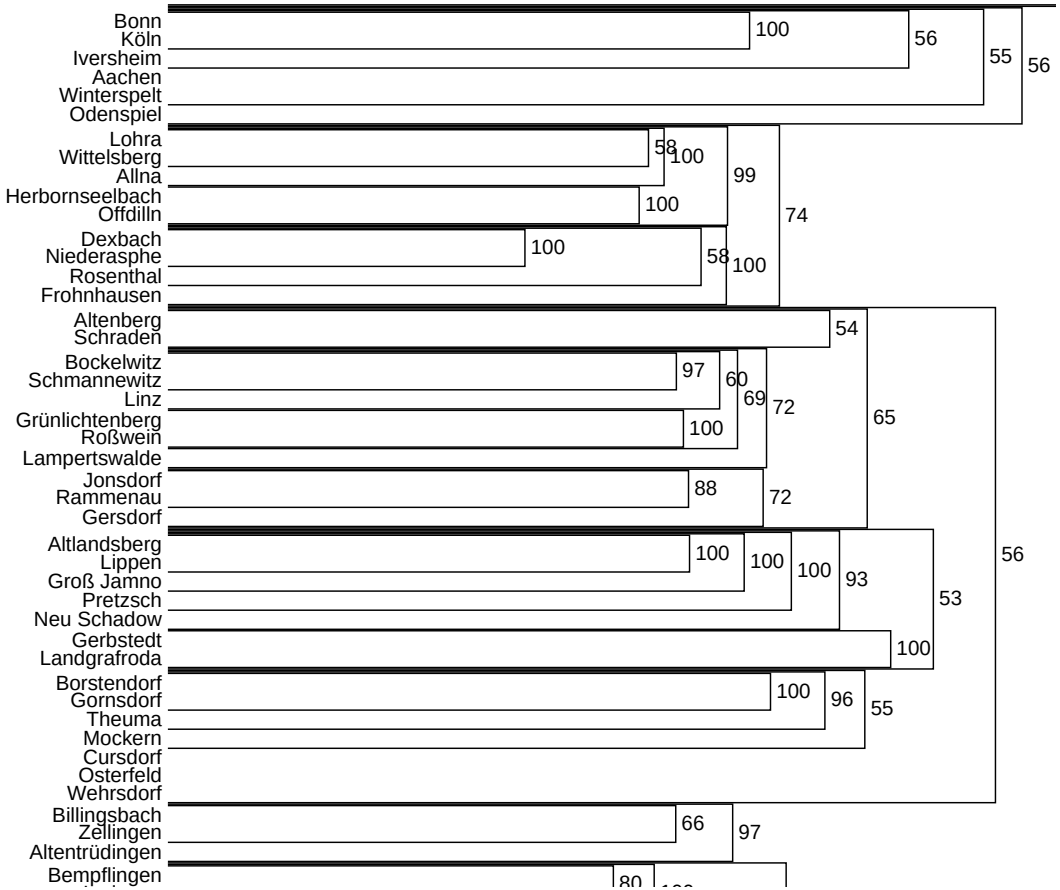


Stability is a real problem!

Bootstrapping Clustering

- assume n (linguistic) distance matrices, $M_{1 \leq i \leq n}$, e.g. one matrix/word
- choose clustering technique, e.g. WPGMA
- repeat, e.g. 100 times
  - select $m \leq n$ matrices, allowing replacement
    - option 1: use repeated selection as weight (Mucha & Haimerl, GfKl 2006)
    - option 2: ignore repetition
  - cluster sum of matrices obtaining dendrogram, recording groups
  - "composite matrix" $M' \Leftarrow$ mean cophenetic distances
  - collect groups that appear a majority of times into a "composite dendrogram"
  - (new!) project dendrogram borders to map, reflecting cophenetic distance in darkness

Composite Dendrograms

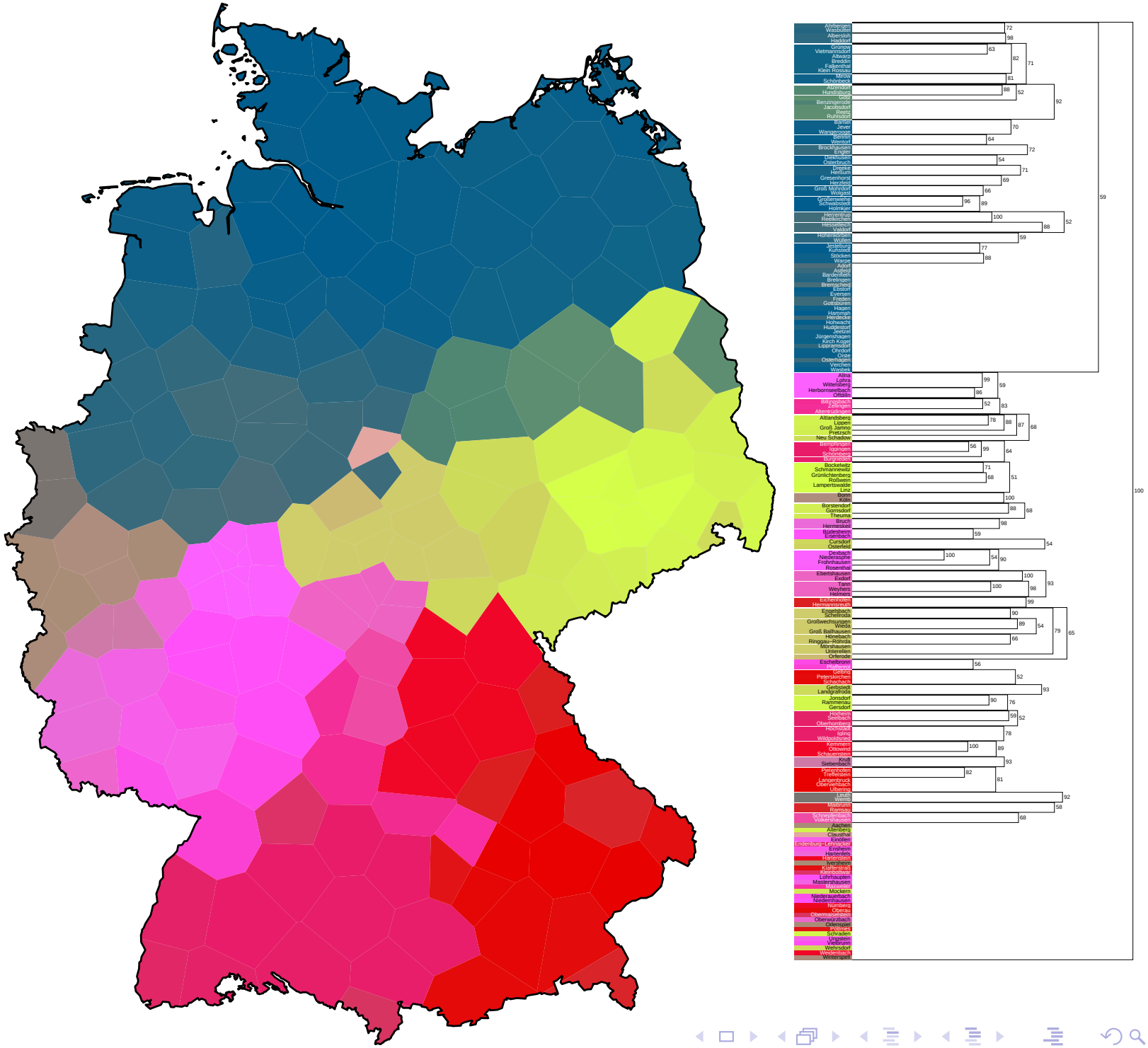


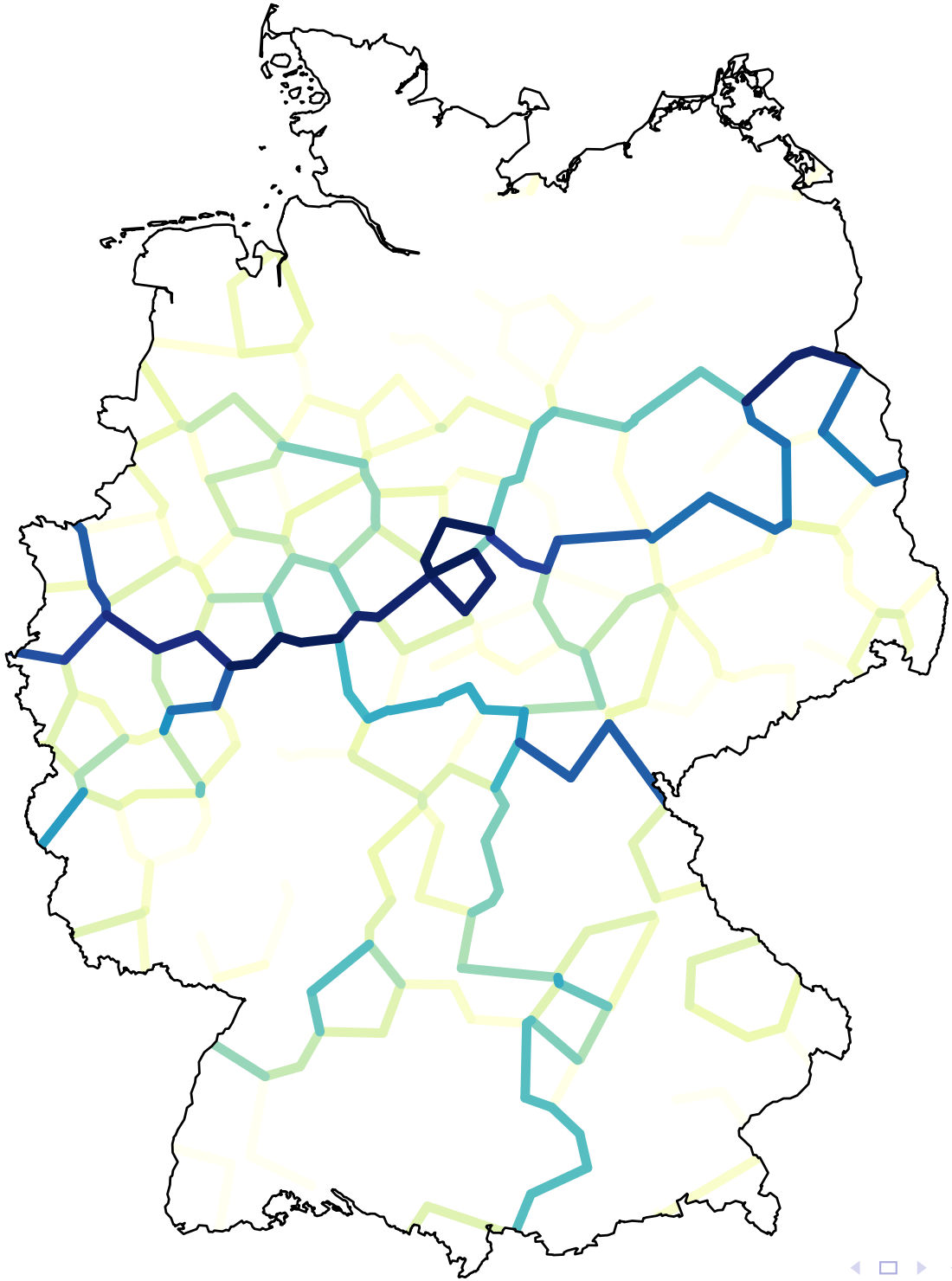
Composite dendrograms shows groups which appear in more than 50% of the repeated (bootstrapped) clusterings.

Cophenetic distance

Mean cophenetic distances M' obtained from bootstrapping $M_{1 \leq i \leq n}$:

- Apply (classical) multi-dimensional scaling to M', obtaining 3-dimensional solutions
  - remaining stress $\approx 10\%$
  - correlation with original M' very high, $r = 0.9$
- Interpret dimensions as red, green, blue intensities



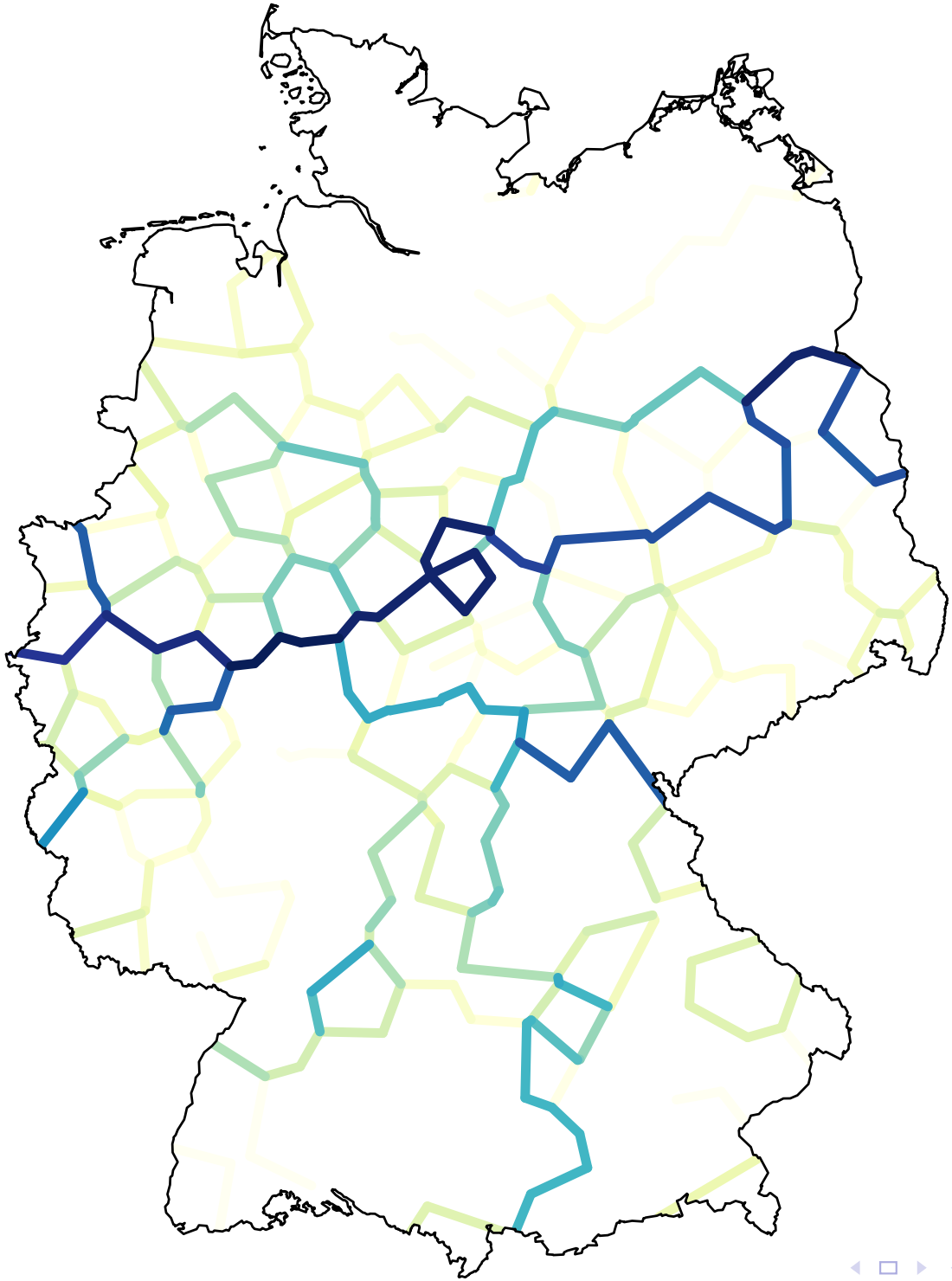


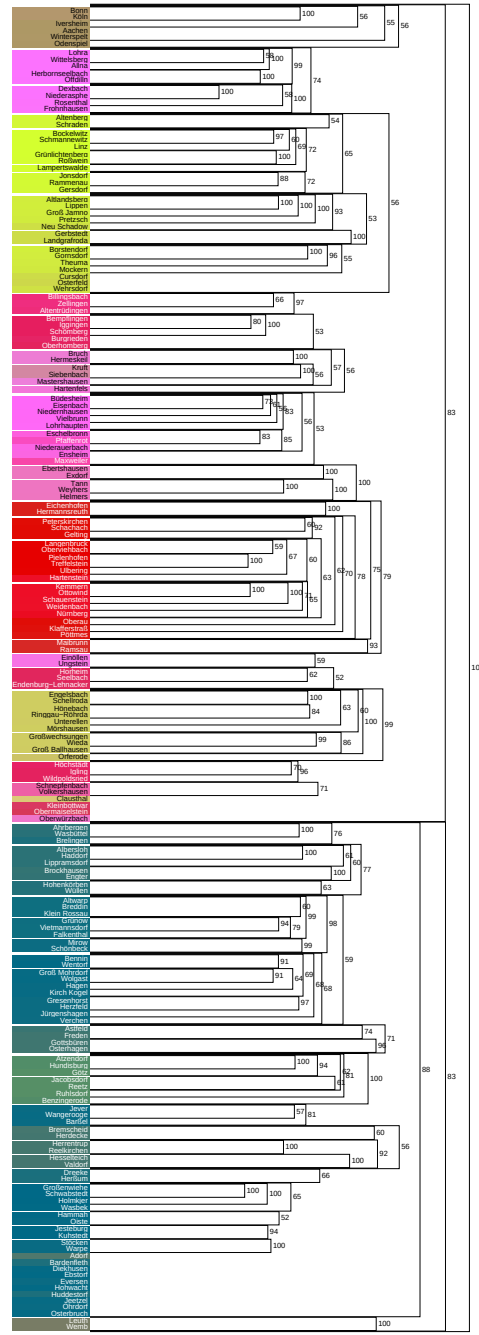
Composite Cluster Maps

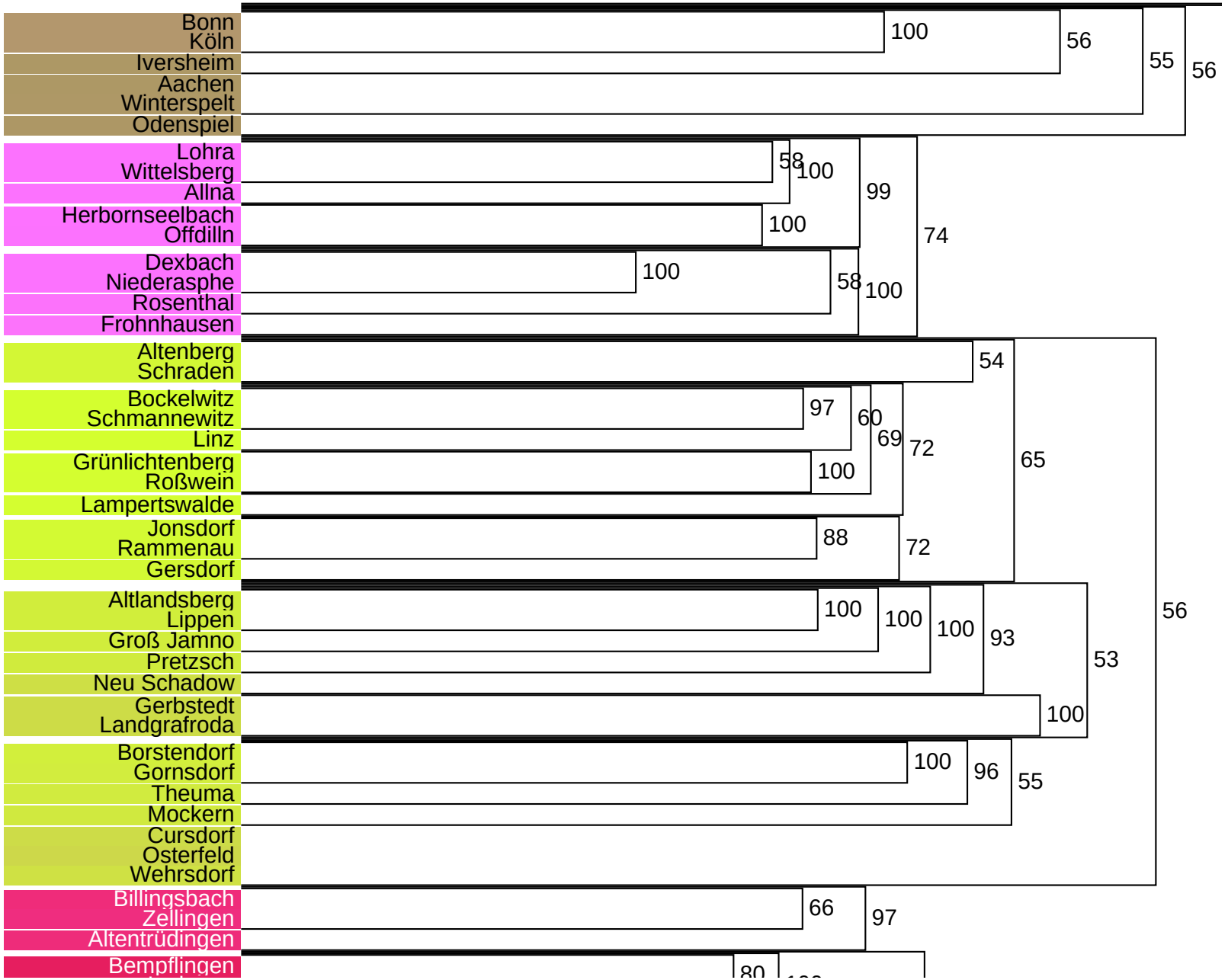
- instability of clustering countermanded by bootstrapping (or by iteration with noise)
- darkness of borders reflects certainty
- “hard” and “soft”borders

Clustering with Noise

- assume a (linguistic) distance matrix, M
- choose clustering technique, e.g. WPGMA
- fix noise parameter, e.g. $a = 0.5\sigma$, where σ is standard deviation among the varietal distances
- repeat, e.g. 100 times
  - add uniform noise r to M, choosing randomly from $0 \leq r \leq a$
  - cluster, noting cophenetic distances obtained
  - collect mean cophenetic distances in "composite matrix" M'
  - project mean cophenetic distances from M' to map
  - collect groups that appear a majority of times into a "composite dendrogram"

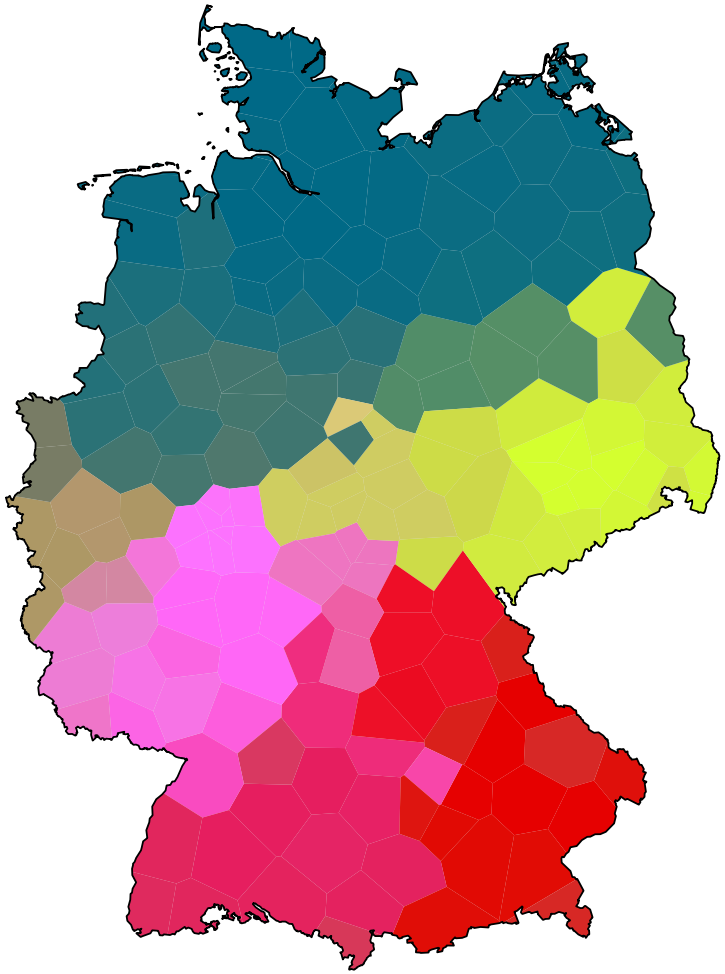




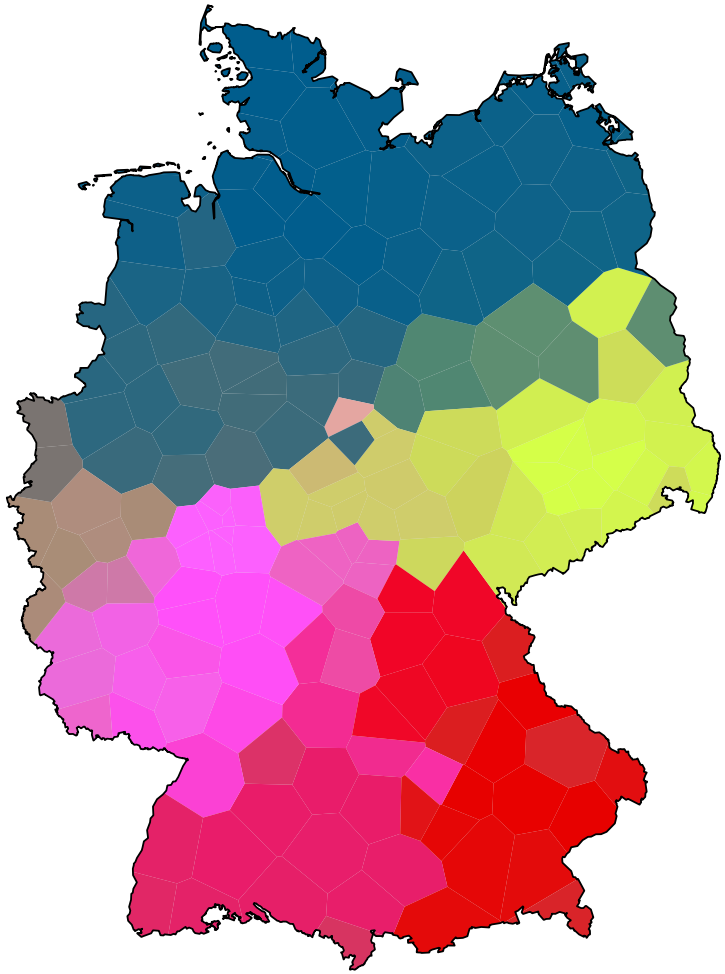

Bonn
Köln
Iversheim
Aachen
Winterspelt
Odenspiel
Lohra
Wittelsberg
Allna
Herbornseelbach
Offdilln
Dexbach
Niederasphe
Rosenthal
Frohnhausen
Altenberg
Schraden
Bockelwitz
Schmannewitz
Linz
Grünlichtenberg
Roßwein
Lampertswalde
Jonsdorf
Rammenau
Gersdorf
Altlandsberg
Lippen
Groß Jamno
Pretzsch
Neu Schadow
Gerbstedt
Landgrafroda
Borstendorf
Gornsdorf
Theuma
Mockern
Cursdorf
Osterfeld
Wehrsdorf
Billingsbach
Zellingen
Altentrüdingen
Bempflingen
100
56
55
56
58
100
99
74
100
100
58
100
54
97
60
69
72
65
100
88
72
100
100
100
93
53
56
100
100
96
55
66
97
80

Correlation

$r = 0.997$

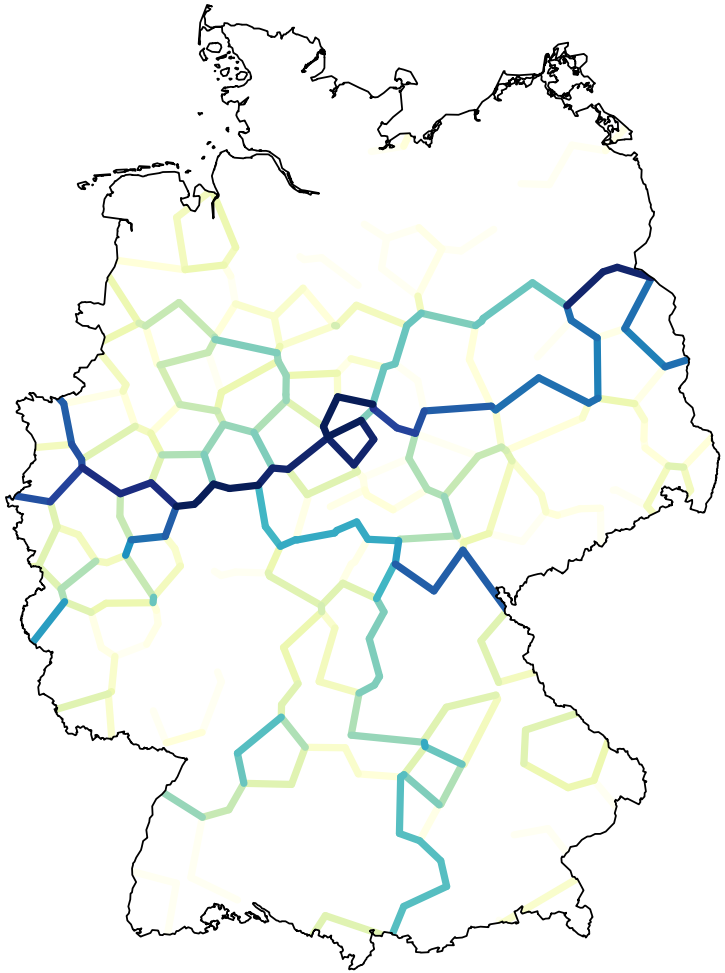


clustering with noise

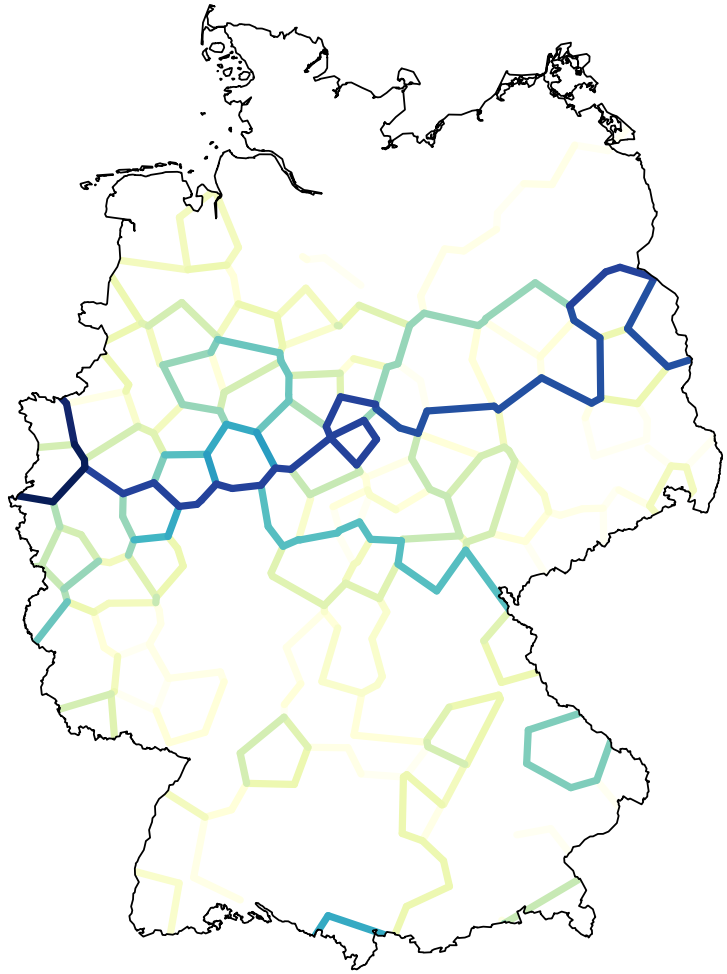


bootstrap clustering

Different Clustering Algorithms

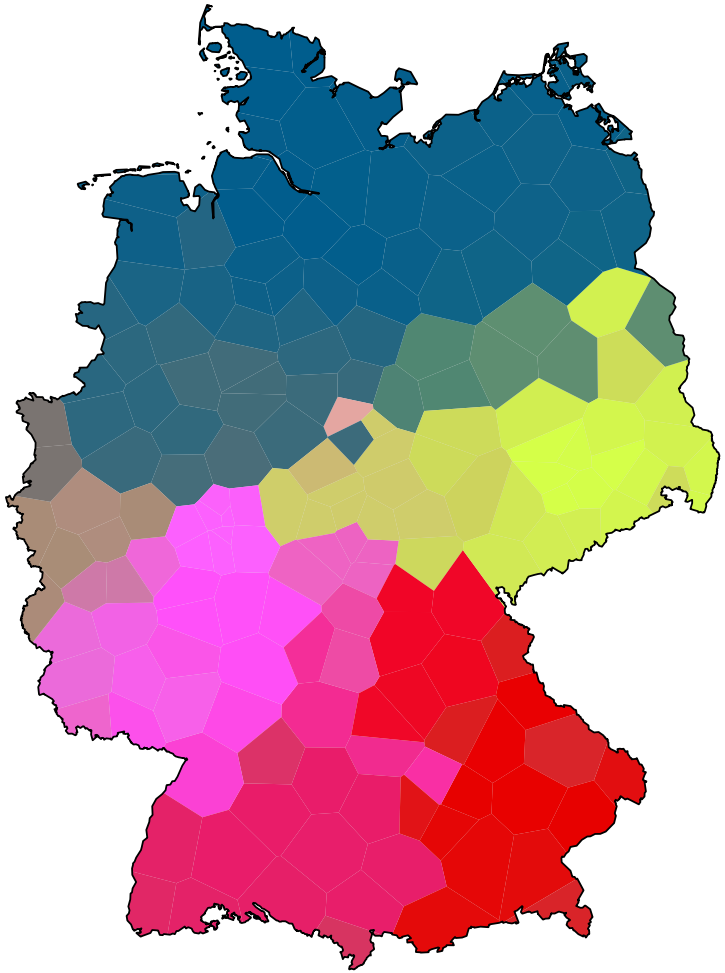


Weighted Ave. (WPGMA)

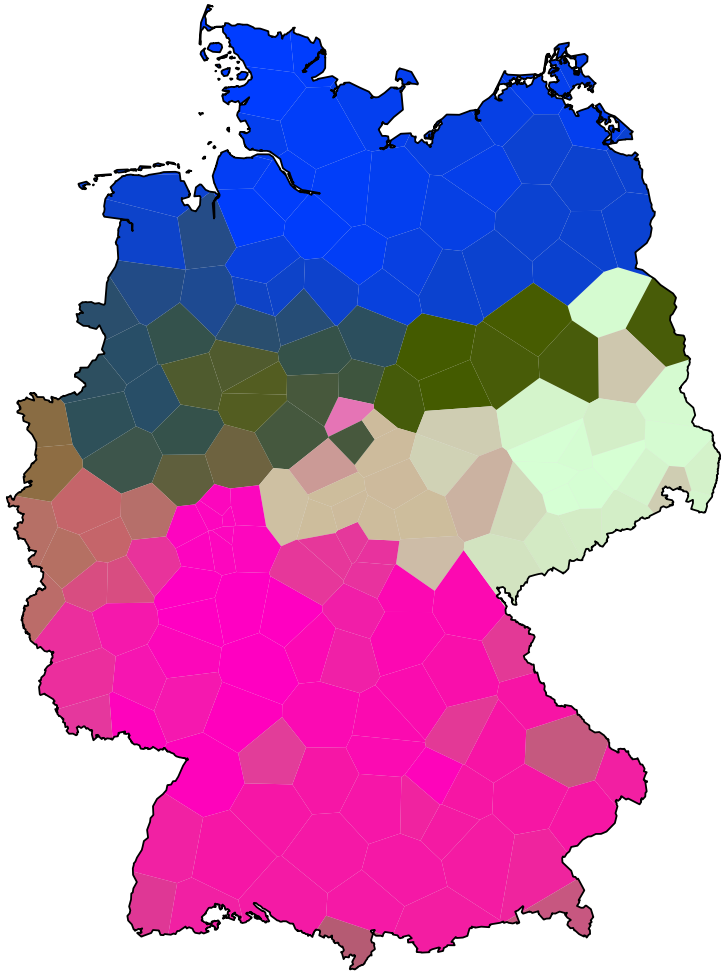


Unweighted Ave.(UPGMA)

Clustering Technique



Weighted Ave.



Unweighted Group Ave.

Conclusions

- Stability obtained—both through bootstrapping and through iteration with random small amounts of noise
- Noise-adding procedure needs a noise parameter, bootstrapping number of submatrices to use.
- Noise-adding procedure applicable to single matrices, bootstrapping requires that many be present.
- Choise of clustering technique still important
  But see: `http://www.let.rug.nl/~kleiweg/kaarten/MDS-clusters.html`