Aggregate Variation in the South in LAMSAS

John Nerbonne
Rijksuniversiteit Groningen

Language Variety in the South III

University of Alabama, Tuscaloosa

Fri. April 16, 2004

Aggregation in Variation

Thesis: Language variation must be studied in the aggregate.

- Detailed studies of single features ([aı] vs. [a], [æ] vs. [æ$^{\text{ə}}$]) are at best inconclusive, at worst misleading.

- Coseriu warned against “atomism” in dialectology

- “aggregate” is not necessarily “global”
  —we can aggregate at different levels

The bulk of the talk demonstrates the second point.

Outline

- Rhetorical Question
- Aggregating Technique
- Experiment on Southern Vowels in LAMSAS
- Results
- Reflections

Rhetorical Question

What is the right level of aggregation for variant studies?

- Allophonic
  - [r] vs [ə], [aɪ] vs. [a], [æ] vs. [æ^ə], . . .
- Lexical choice
  - *dragonfly* vs *darning needle*, . . .
- Morphological *blew* vs. *blewed*, . . .
- Syntactic construction *needs washed*, *a hundred year/years*. . .
- Any of above, with frequency?

Aggregate Pronunciation Difference

Dialectometry (Séguy, Goebl) has focused on collecting lots of features, treated as qualitative variables.

Levenshtein distance (aka “edit distance”, “string distance”) provides a way to measure pronunciation difference.

- numerical, therefore additive
- collect random sample of pronunciations (of the same word)
- use sum of differences as **aggregate varietal distance**

Segment Distance

- Sum feature distances in feature vectors to obtain segment distances.

  Example: d([i],[e]) $\ll$ d([i],[u])

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid high)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
rounded	0(not rounded)	0(not rounded)	1(rounded)	0	1
				1	5

- Diacritics [ĩ,eː,ə^{r}] can also be taken into account
- Vieregge-Cucchiarini system used, also Almeida-Braun
- Chomsky-Halle (SPE) system less useful (clever features for making rules compact)

Levenshtein Distance

Cost of least costly set of operations mapping one string into another.

	Operation	Cost
æəf t ən ʉn		
æf t ən ʉn	delete ə	d(ə,[])=0.3
æf t ər n ʉn	insert r	d([],r)=0.2
æf t ər n u n	replace [ʉ] with u	d([ʉ],[u])=0.1
Total		0.6

Computing Levenshtein Distance

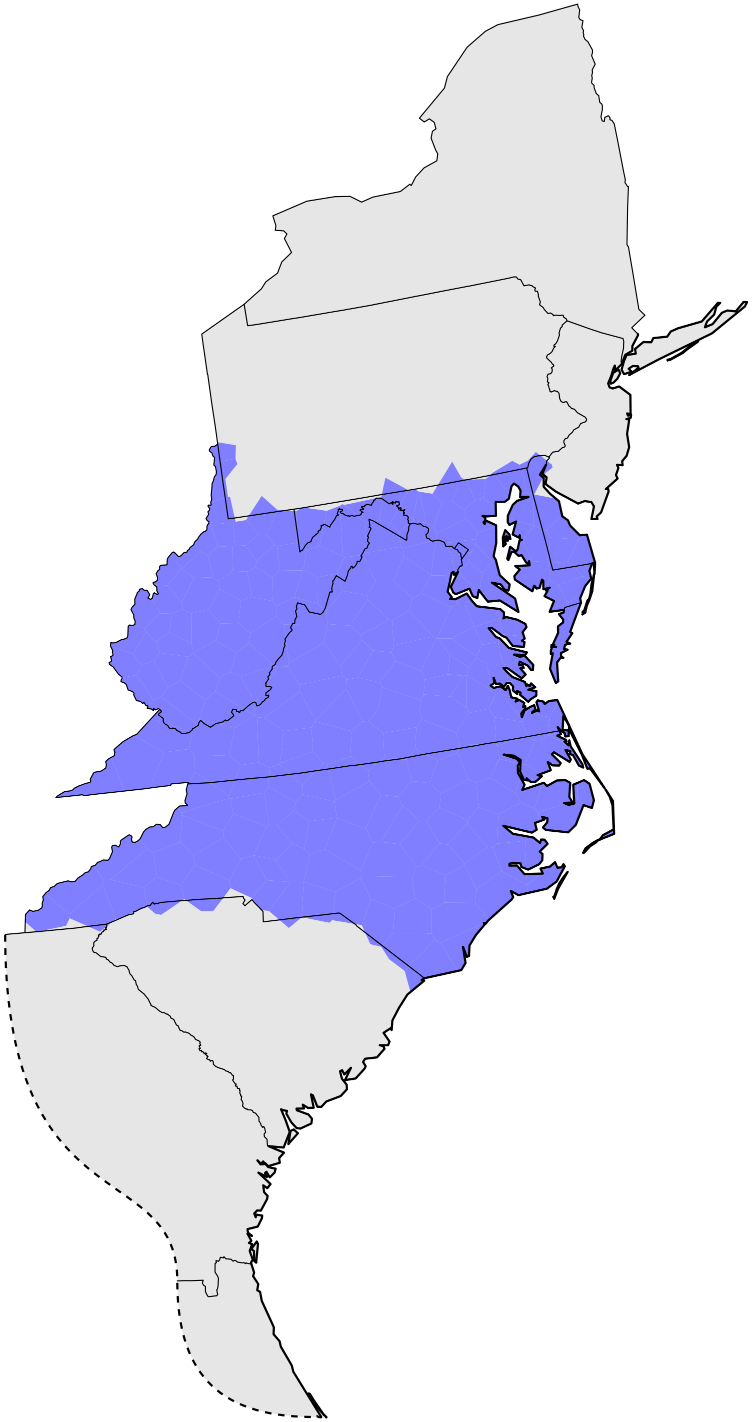
		æ	f	t	ə	r	n	u	n
	0	1	2	3	4	5	6	7	8
æ	1	0	1	2	3	4	5	6	7
ə	2	1	2	3	2	3	4	5	6
f	3	2	1	2	3	4	5	6	7
t	4	3	2	1	2	3	4	5	6
ə	5	4	3	2	1	2	3	4	5
n	6	5	4	3	2	3	2	3	4
ʉ	7	6	5	4	3	4	3	4	5
n	8	7	6	5	4	5	4	5	4

We simplify costs (everything $= 1$) for illustration.

Experimental material

LAMSAS: Linguistic Atlas of the Middle and South Atlantic States

- data collection 1933–1974
- 63.7% of data collected by Guy Lowman 1933–1941
- most analyses based on lexical overlap
- now maintained by Bill Kretzschmar
- restriction to part of Lowman data: NC VA WV DC DE MD
  - apparent transcriber effects



Data Preparation

- Parser recognizes a segment as a sequence of:

1. zero or more pre-modifiers
2. one head
3. zero or more post-modifiers

- Results:
  - 57803 strings (99.95%) parsed correctly
  - 30 strings (0.05%) rejected

Restrict attention to vowels (aggregation below global level).

Some numbers

57833 strings

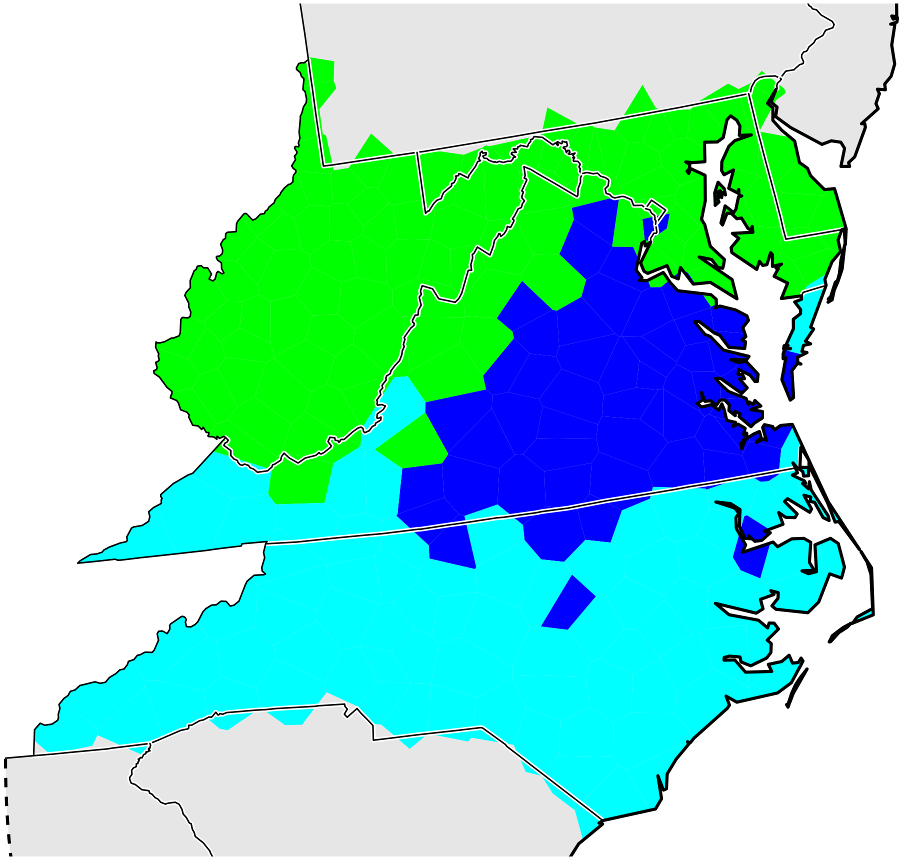
238 locations

On average, 11.0 characters per string, parsed into 6.9 sound tokens per string

1132 unique vowel sounds (combinations of heads and modifiers)

Ignoring all consonants including [j,w,r,l]

Clustering Vowel Differences (Ward’s Method)



Recall that geographic coherence is not assumed!

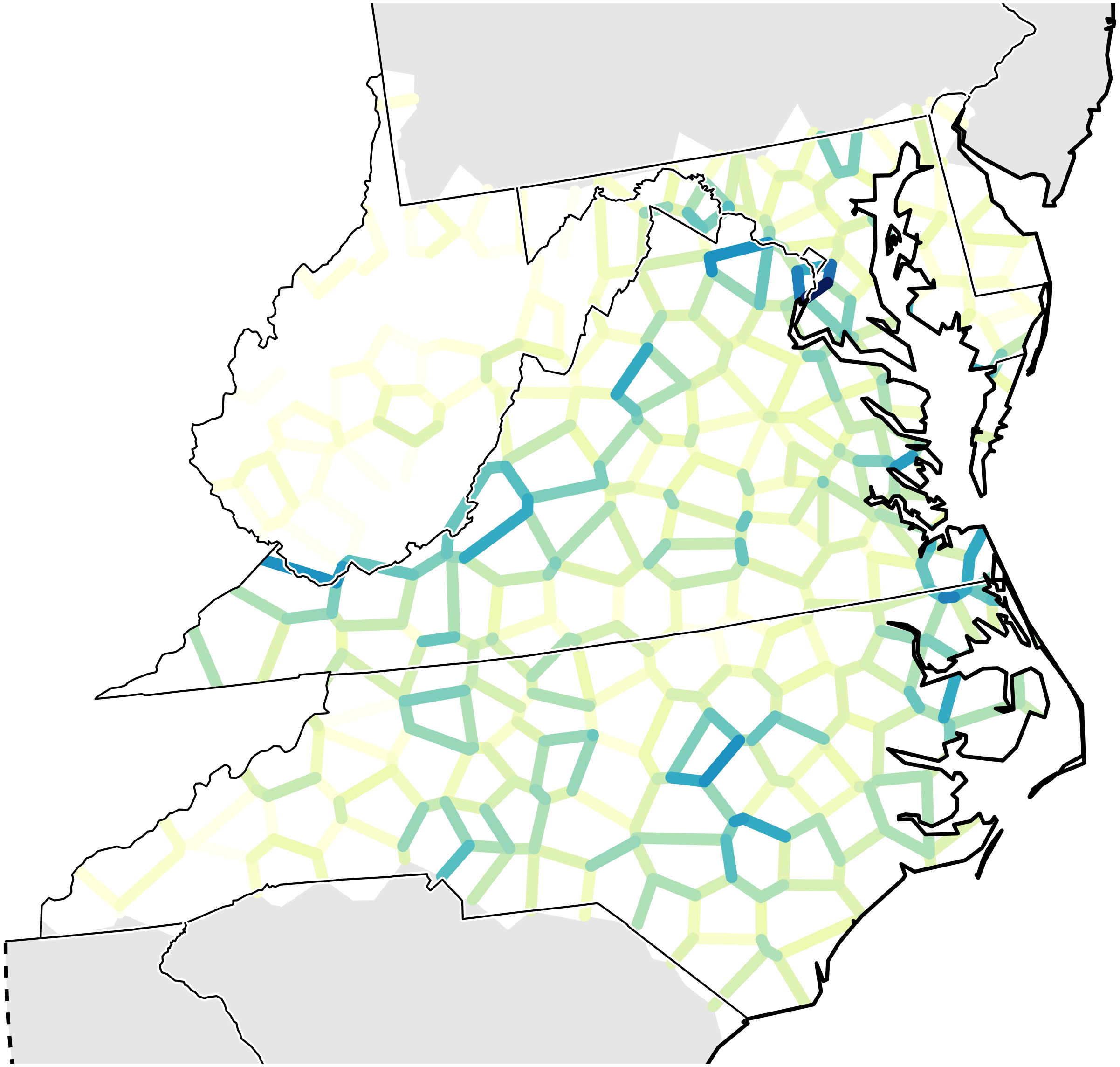
Composite Clustering

- exploratory “grouping”

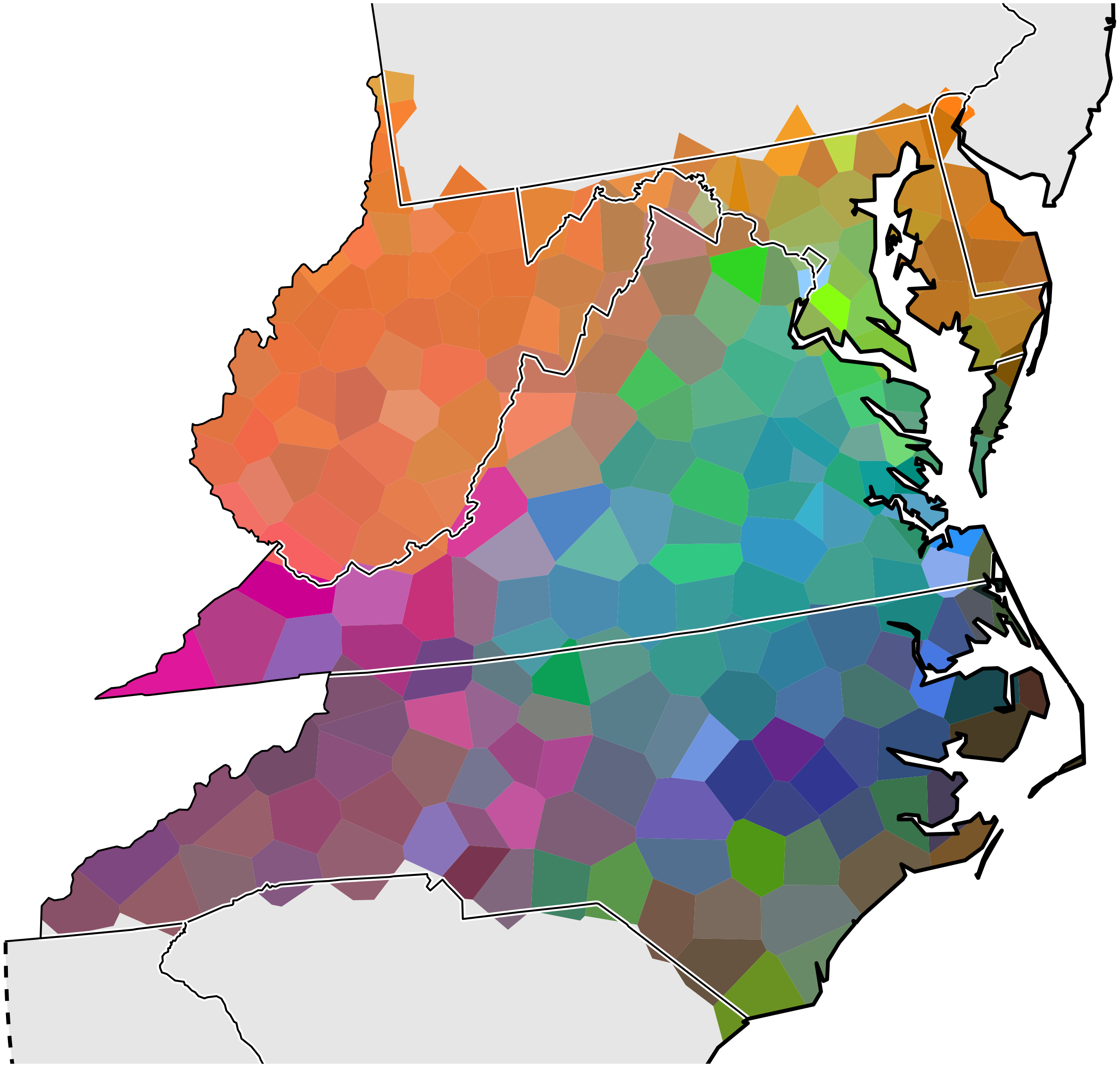
- group average $+$ weighted average

- repeated 50 times with random noise:
  - corrects for instability in clustering algorithm caused by sensitivity to small changes in input data
  - “smooths” differences in results of various clustering algorithms

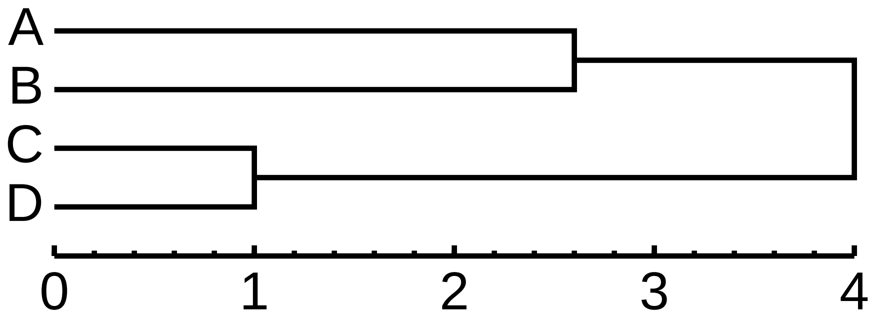
Phonetic differences



Multidimension scaling



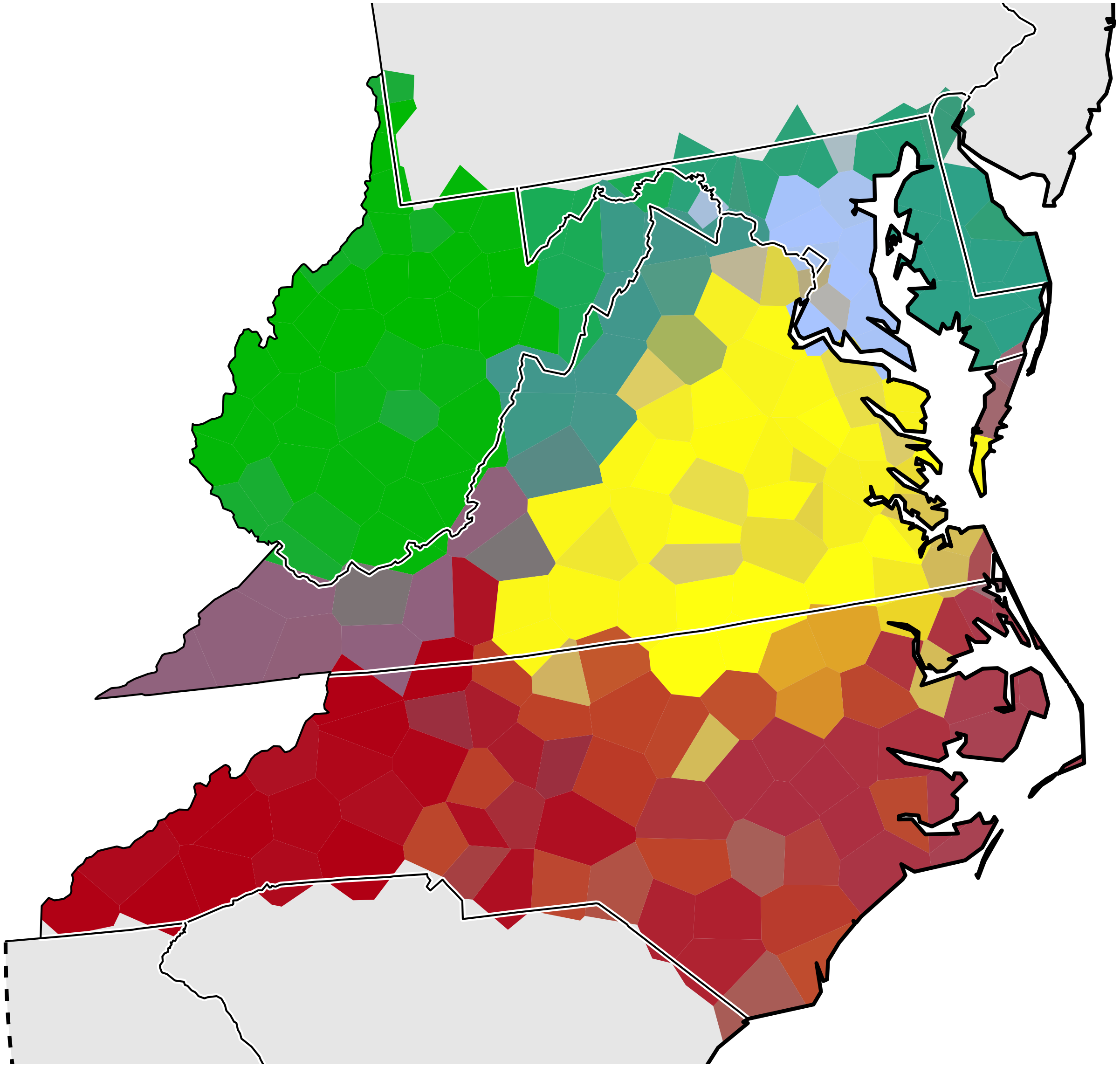
Cophenetic distances



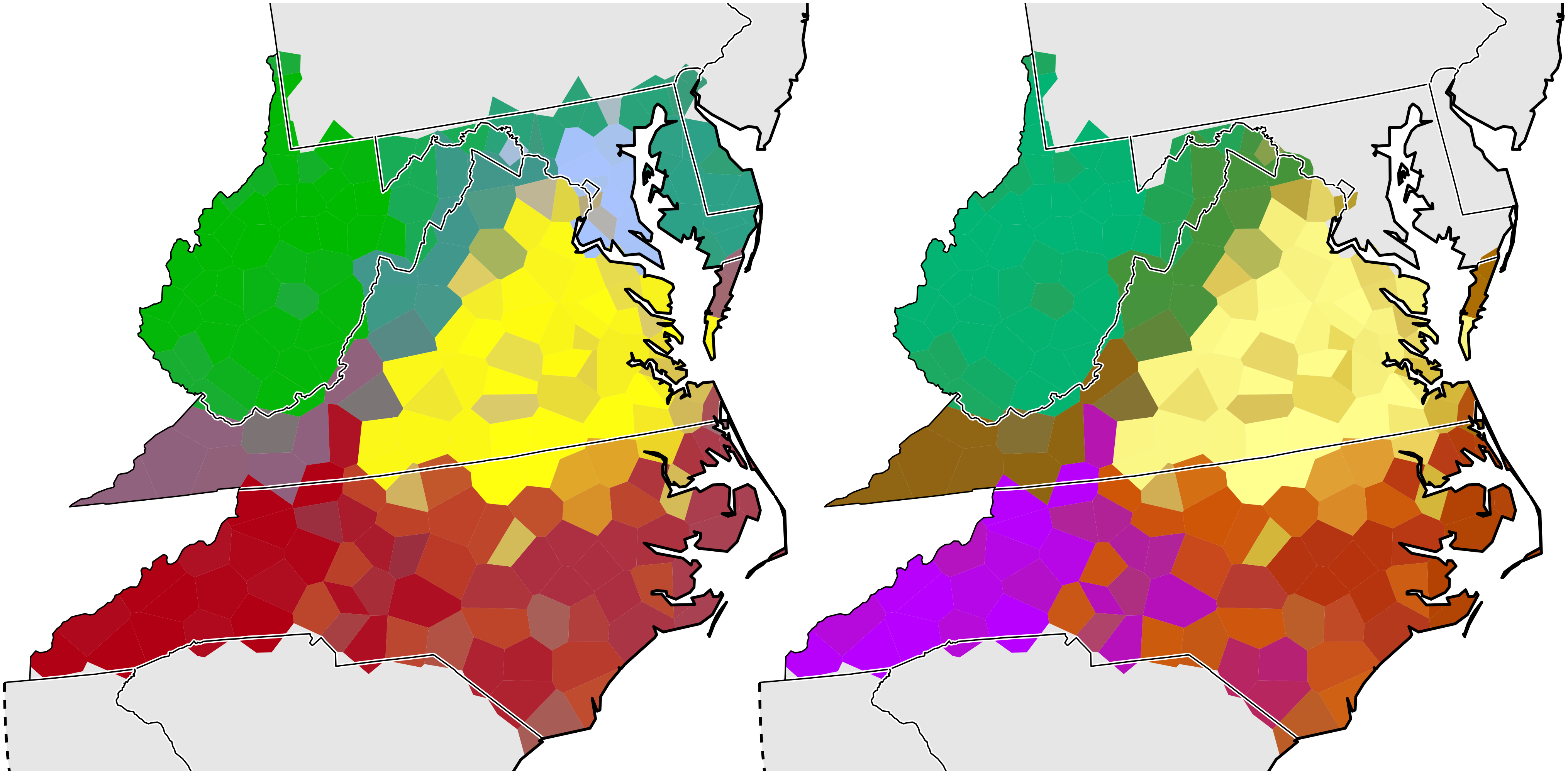
	A	B	C	D
A	0.0	2.6	4.0	4.0
B	2.6	0.0	4.0	4.0
C	4.0	4.0	0.0	1.0
D	4.0	4.0	1.0	0.0

We average cophenetic distances from many clusterings, reanalysing through MDS.

MDS of Cophenetic Distances



Effect of Borders

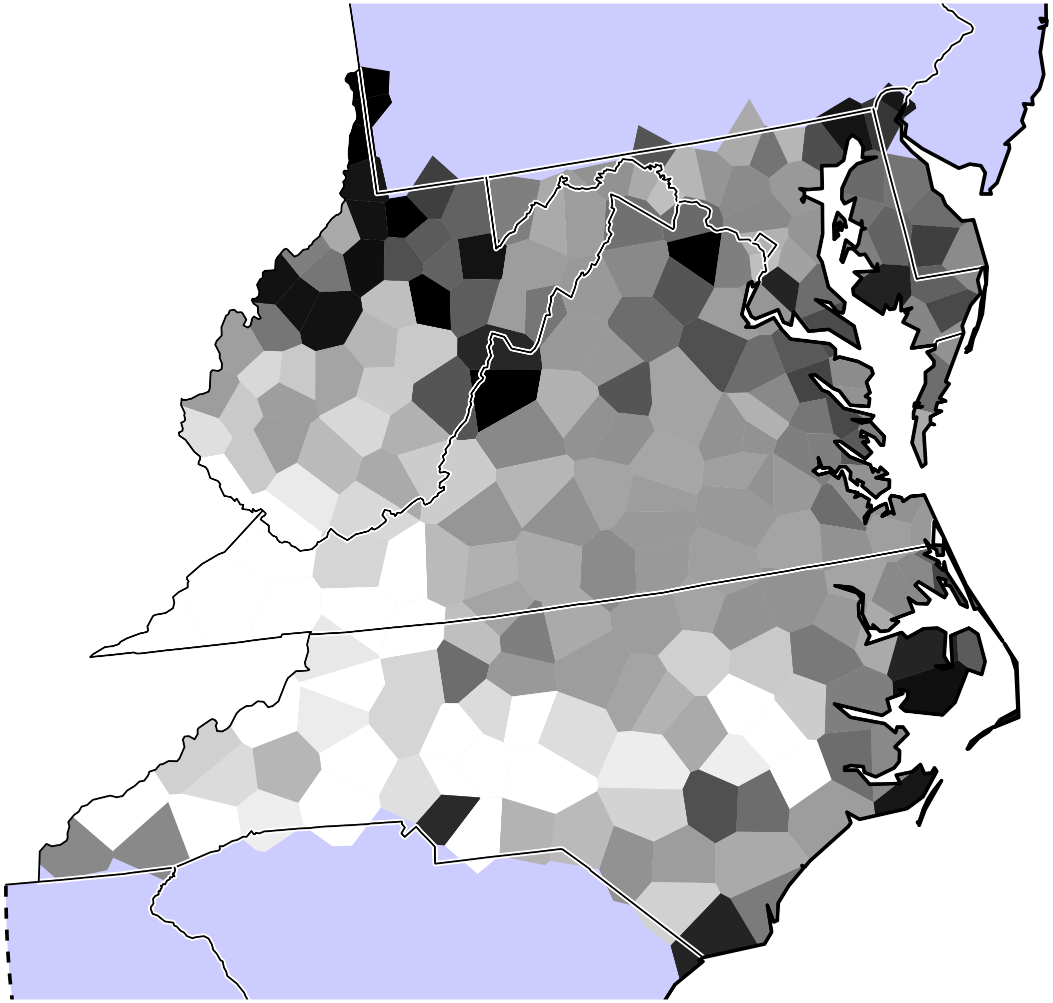


Conclusions

Aggregate views of pronunciation now possible

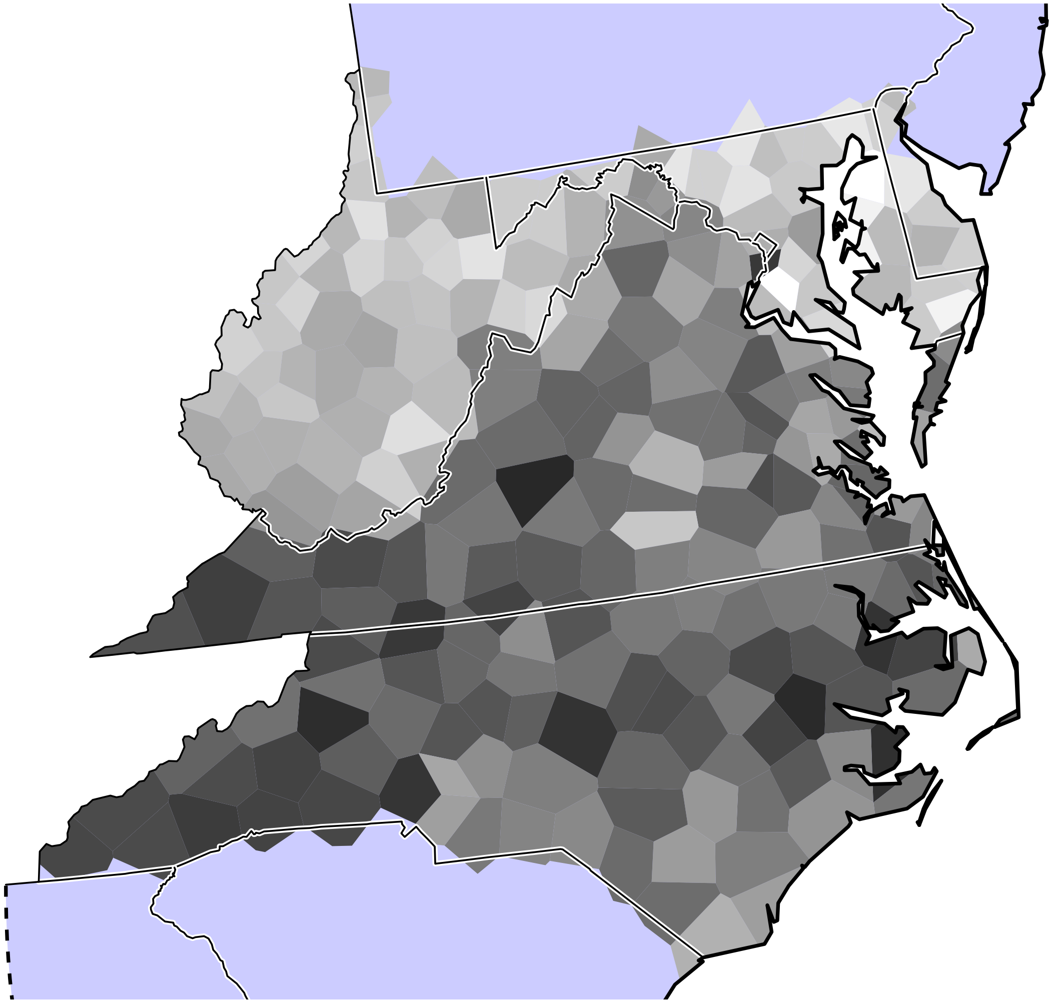
Various levels of aggregation, e.g., vowels.

Comparison to Individual Features [a/aɨ]



light [a] — dark [aɨ]

Comparison to Individual Features [æ/æ^{ə}]



light [æ] — dark [æ^{ə}]

Reflection

Single features vs aggregate

- How many single features?
  - segments, allophonic rules, frequencies, . . .
  - choice?
- Counter-indicating features
  - non-coinciding isoglosses
  - choice?
- Status of the linguistic *variety*
  - Sum of linguistic conventions characteristic of a group