Reclassifying Yuè Chinese

A Dialectometric Approach

John Carlyle

2022-05-22 Sun

University of Washington
jtcarlyl@uw.edu

Outline

Preliminaries

Introduction to Dialectometry

Measuring Lexical Similarity

Linguistic and Geographic Distance

Multidimensional Scaling

Clustering

Wrap-up

Preliminaries

The Data

- 70 Yuè dialects:
  - Zhān, Cheung, et al., et al. (1987a)
  - Zhān, Cheung, et al., et al. (1987b)
  - Zhān, Cheung, et al., et al. (1994)
  - Zhān, Cheung, et al., et al. (1998)
  - Zhān et al. et al. (2000)
  - Yue-Hashimoto (2005)
  - Xiè (2007)
  - Lǐ (2014)
  - Xiǎn (2016)
- 77 word list
- See Carlyle (2020) for details.

What is a Yuè Dialect?

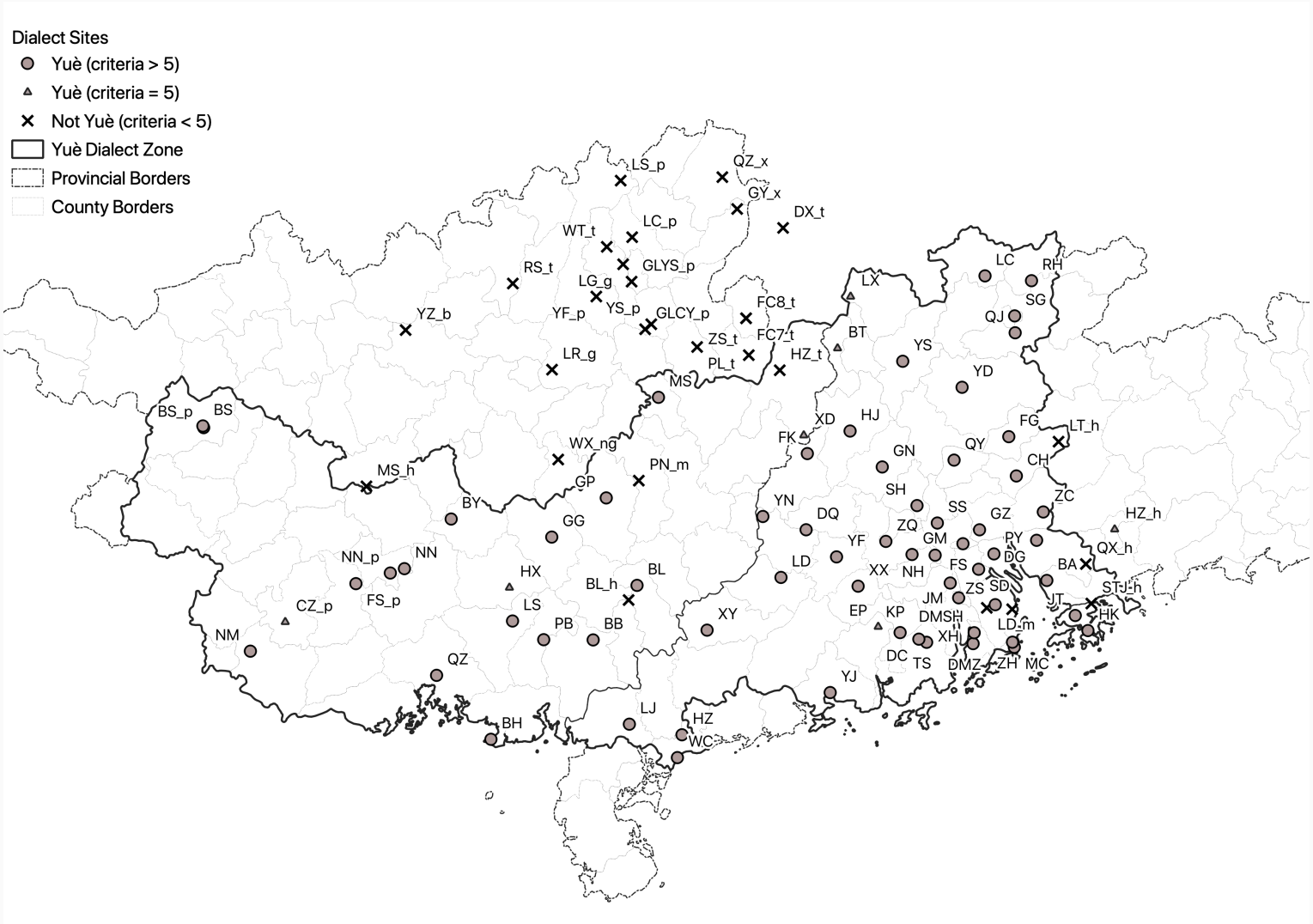
A simple diagnostic test (Carlyle 2020, pp. 33-34):

1. A phonemic lower *yīnrù* (陰入) tone
2. Long, open /a/ vs. short, centralized /ɐ/
3. “slaughter” /tʰɔŋ¹/ (劏)
4. “thing” /ɲɛ⁴/ (嘢)
5. “noon” involves /an⁵/ (晏)
6. Feminine suffix for animals /na³/ (嫲)
7. Person plural marker /ti⁶/ (哋)
8. “child” either /sɐj⁵ mɐn¹ tsɐj³/ (細民仔) or /sɐj⁵ lɔw³ kɔ¹/ (細佬哥)
9. “(early) morning” involves some combination /tʃiw¹/ (朝) and /tsɔw³/ (早)

What is a Yuè Dialect? (cont.)

- 5 or more → (probably) Yuè Chinese
- Convenient way to narrow focus to dialects most experts agree are Yuè. Not meant to be the final say.
- Predicts S. Pinghua dialects are Yuè, *but* N. Pinghua dialects are not.

The Diagnostic Test Applied to the Yuè Dialects and their Neighbors



Introduction to Dialectometry

Dialectometry

What is it?

- The use of computational and quantitative techniques in dialectology
- Measure the degree of linguistic similarity (or distance) between dialects
- Relate these measurements to geographic distance and plot them

Why use it?

- Visualize migration, contact, and cultural boundaries
- Obtain a synchronic classification
- Create high quality maps using GIS
- Useful for education, language planning, etc.

The Isogloss Method

Process

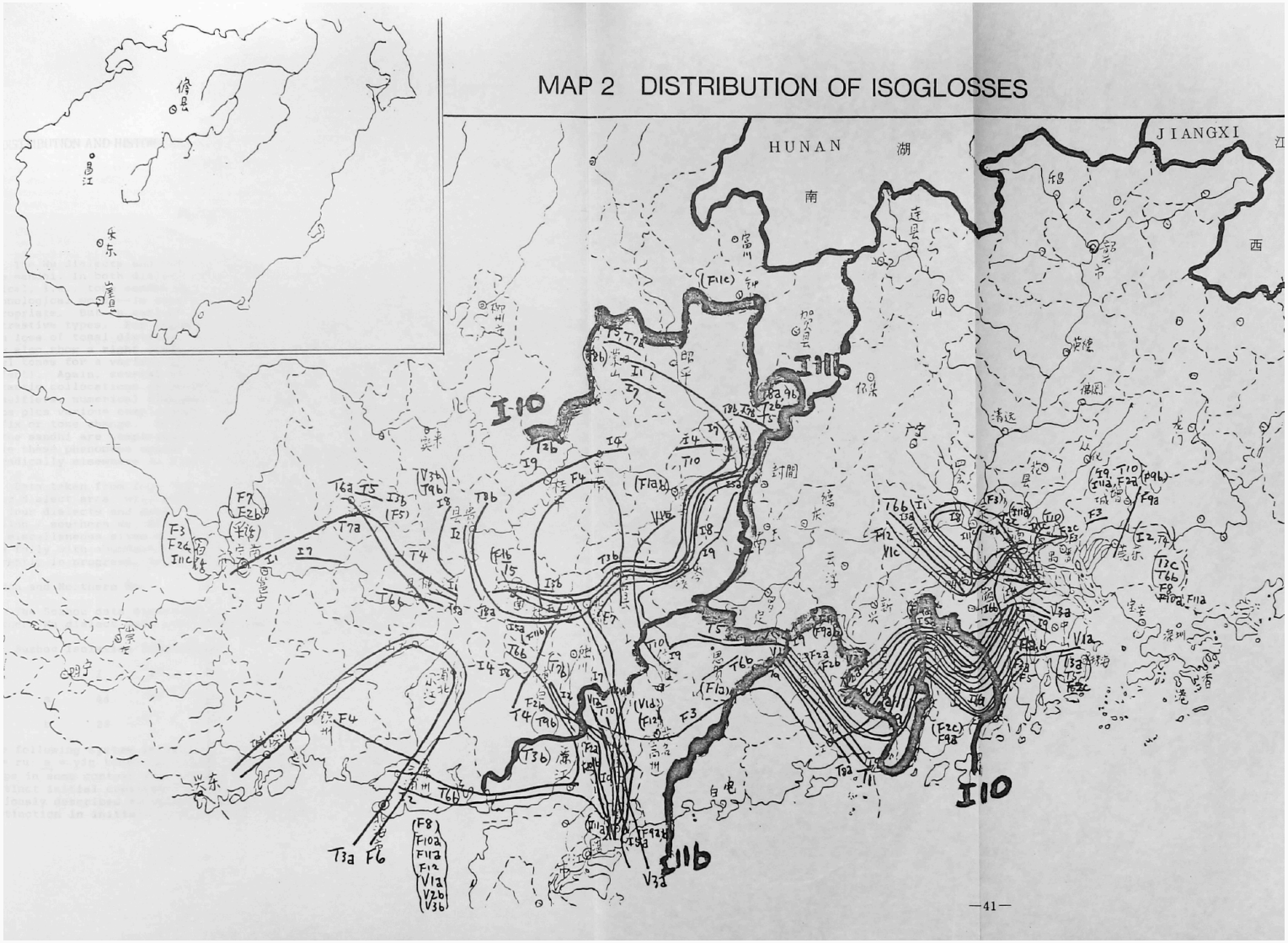
- Select linguistic differences
- Draw lines on map to mark boundaries of differences
- Seek out bundles of isoglosses

Limitations

- Possible bias in selecting differences
- Can’t directly compare non-contiguous regions
- Laborious
- Difficult to interpret results

Yuè Chinese Isoglosses

Yue-Hashimoto (1988, p. 41)



The Advantages of Dialectometry

- Can use data from all linguistic levels
- Data represents modern dialects directly
- Includes all data without biased selections
- Uses data maximally
- Can compare areas that are not close
- Clear results

Gabmap

Nerbonne et al. et al. (2011)

- Online Dialectometry Web App
- Based on earlier RuG/L04 program
- http://www.let.rug.nl/~kleiweg/L04/webapp/

Measuring Lexical Similarity

Categorical Data

Site	to rain	morning	salt	…
GZ	落雨	聽日	鹽	…
TS	落水	天早	上味	…

Categorical Distance
Séguy (1971)

For distance between two dialects:

- same words as 0 (no distance)
- different words as 1
- take average for word list

Weighted Difference Value
Goebl (1984)

- *gewichteter Identitätswert*
- weight words by the frequency they appear answer to word list item
- emphasize less common responses

Limitations of Categorical Yuè Data

- Orthography not standard across regions or surveys
- No accepted *zi* for some morphemes
- Even broad transcriptions not very similar
- Judging which words are the “same” not always trivial

String Edit (Levenshtein) Distance

The method comes from Levenshtein (1966). Applied to gauge lexical similarity in Nerbonne and Kleiweg (2003).

- Smallest set of operations to transform one string (of segments) to another
- Insert, delete, substitute
- Normalize by length of compared strings
- Follow Yang and Castro (2008) to handle tone

GZ to BA

s	i	k	L	E	
s	e	ʔ	L	E	
	1	1			2

Local Incoherence

Nerbonne and Kleiweg (2007)

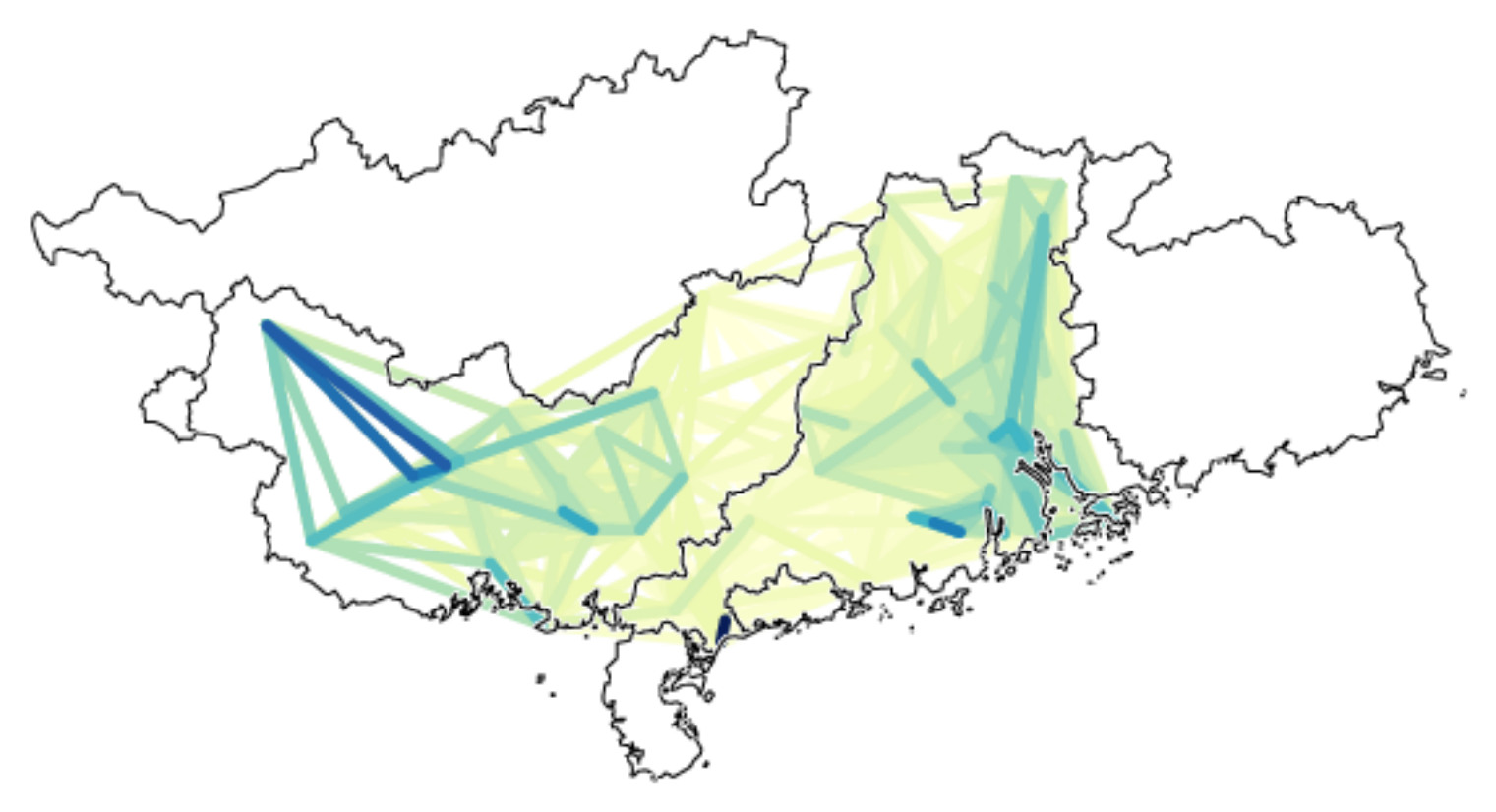
$$I_l = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i^L - D_i^G}{D_i^G}$$

- Average of the geographic distance between the most linguistically similar site for each site normalized by the distance of the actual closest site

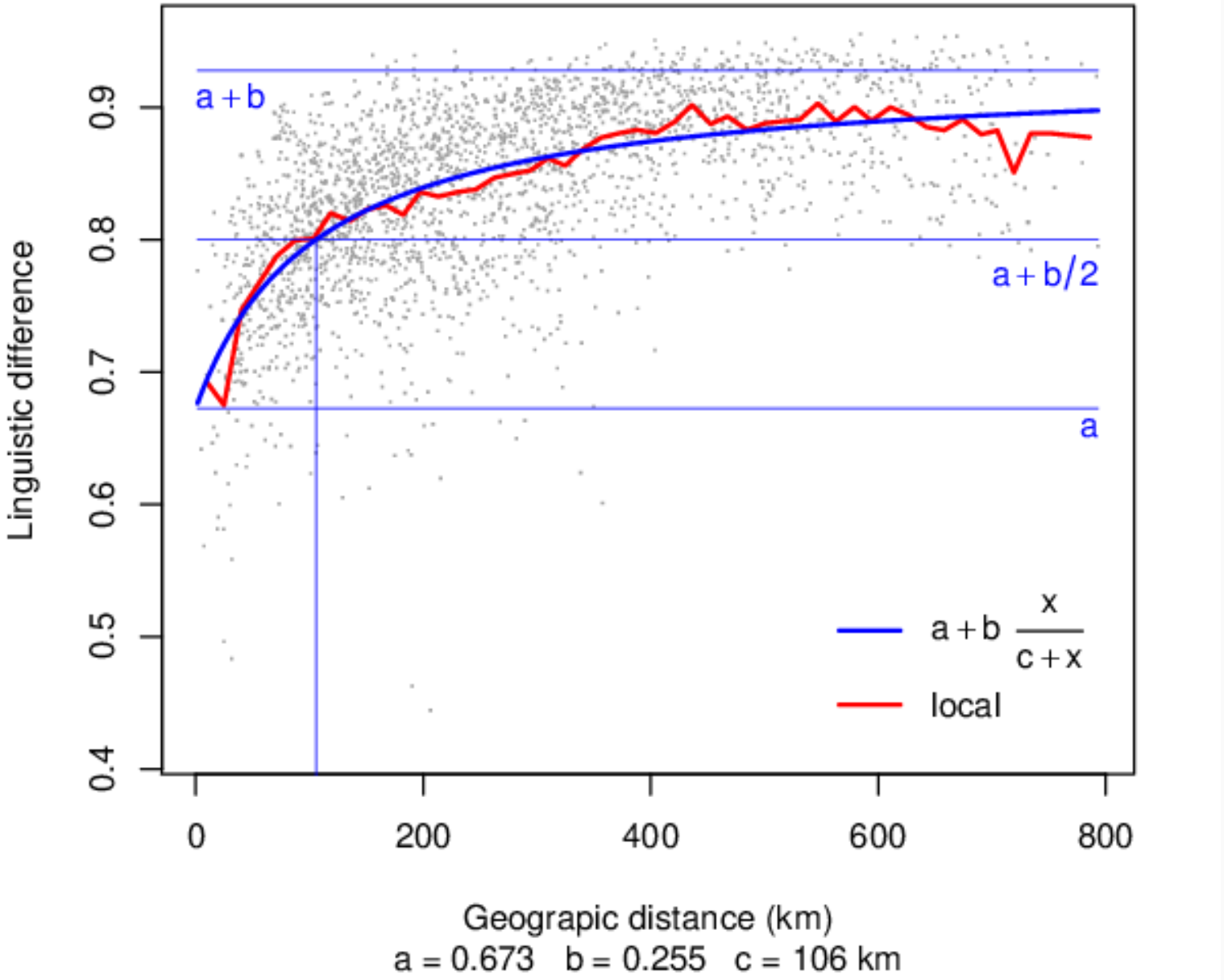
Method	Local Incoherence
Binary	1.10
Weighted	0.95
Levenshtein	1.34

Linguistic and Geographic Distance

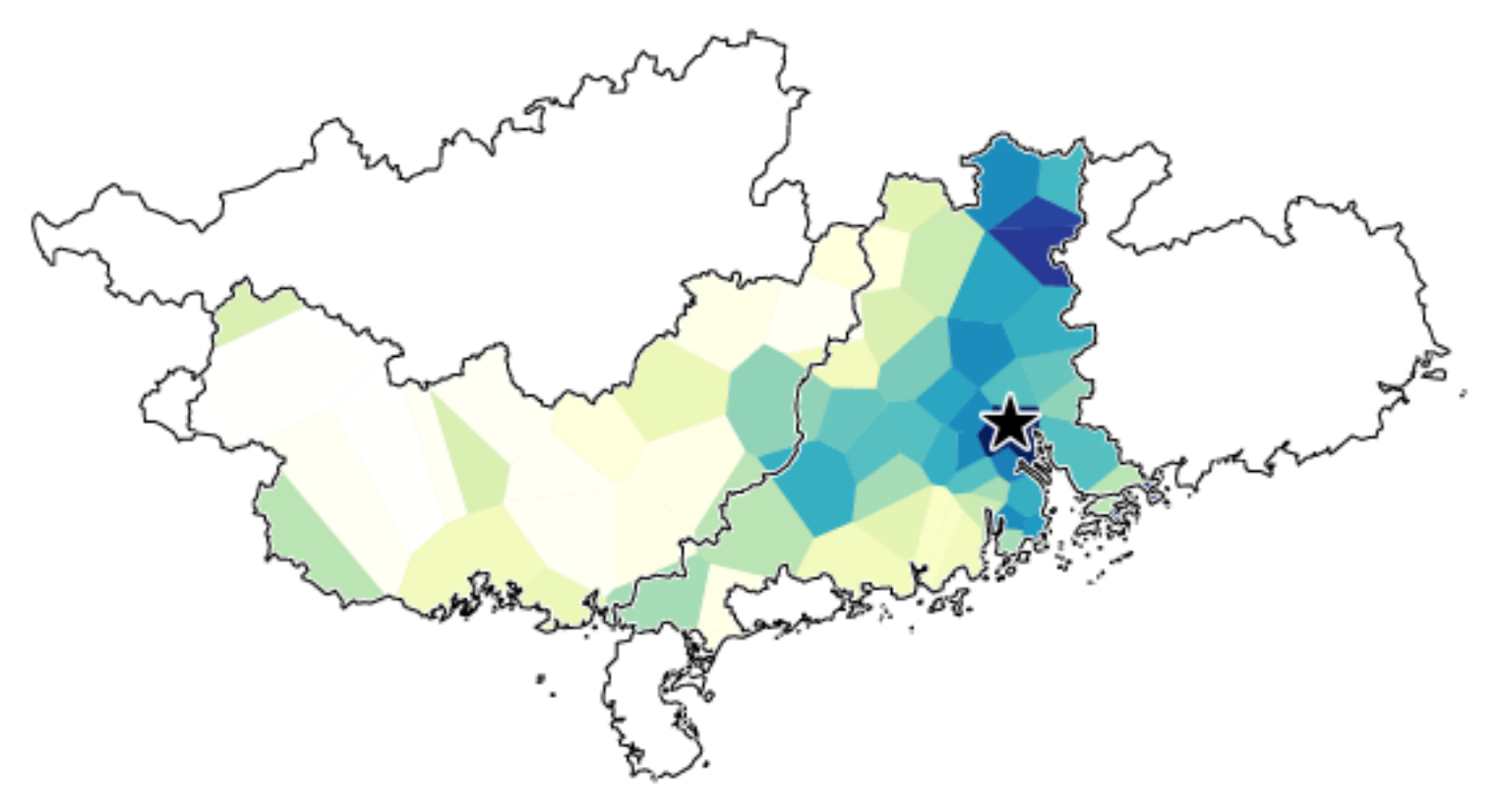
Difference Map



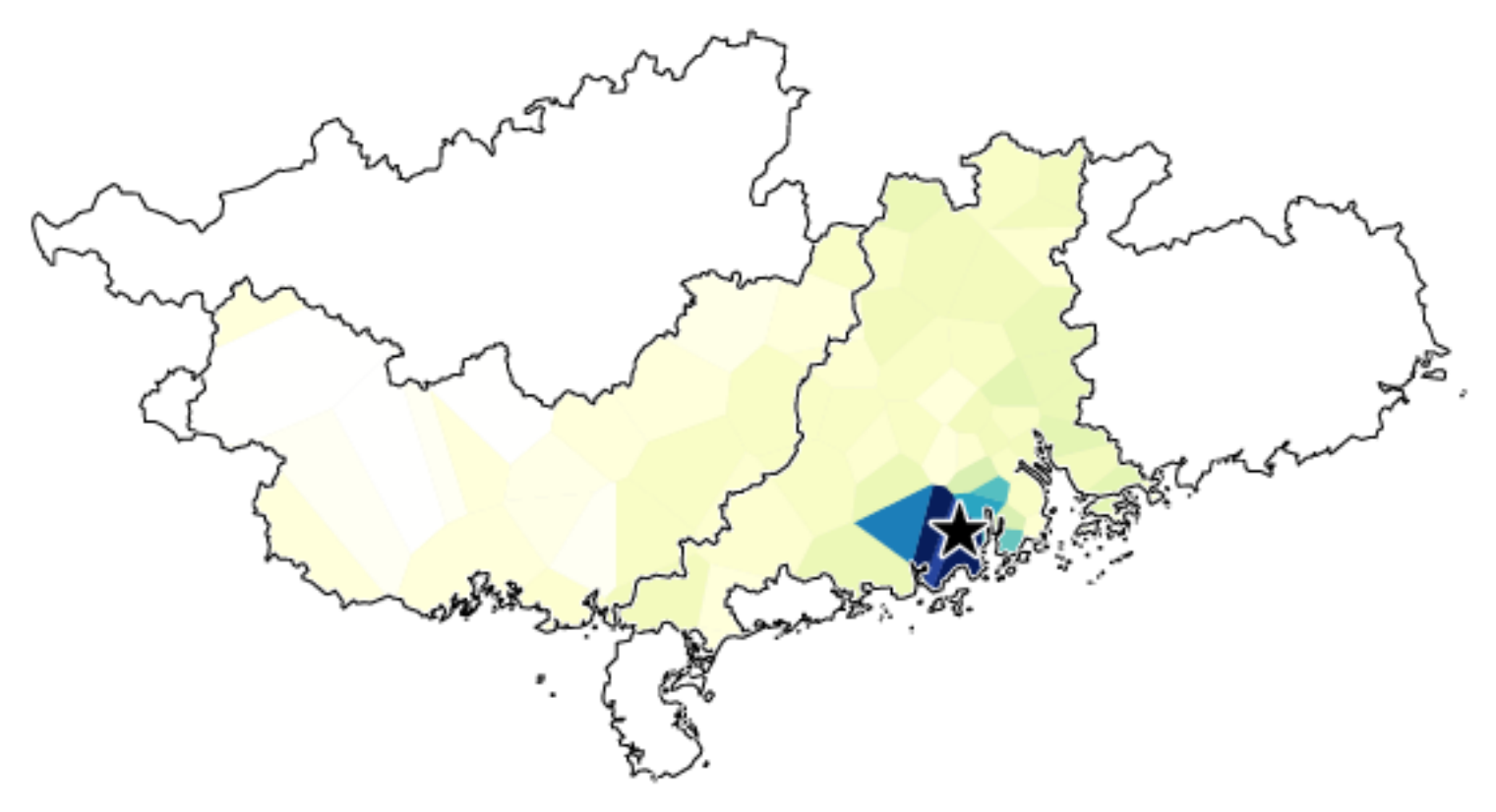
Linguistic Difference ↔ Geographic Distance



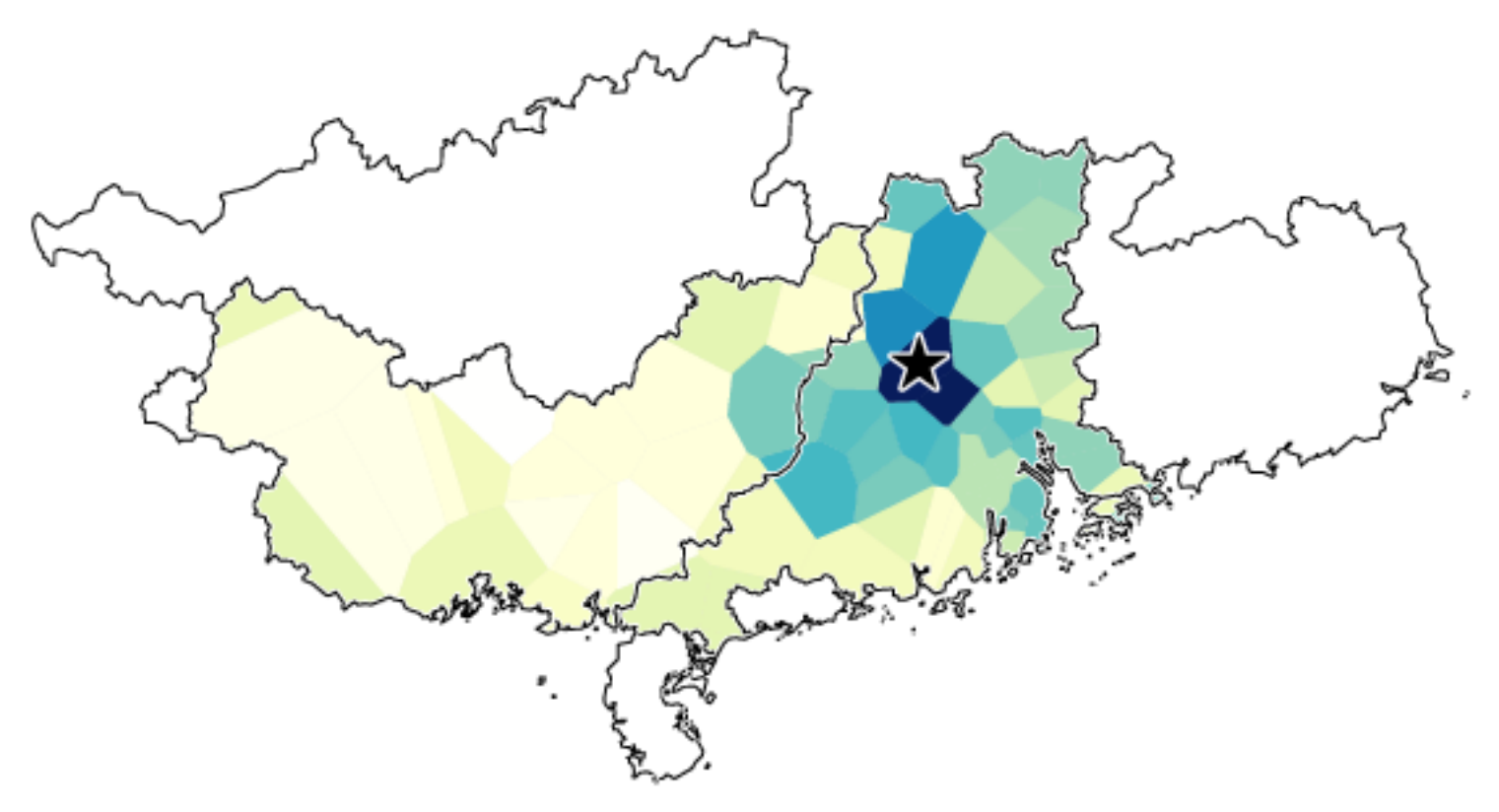
Reference Point: Guangzhou



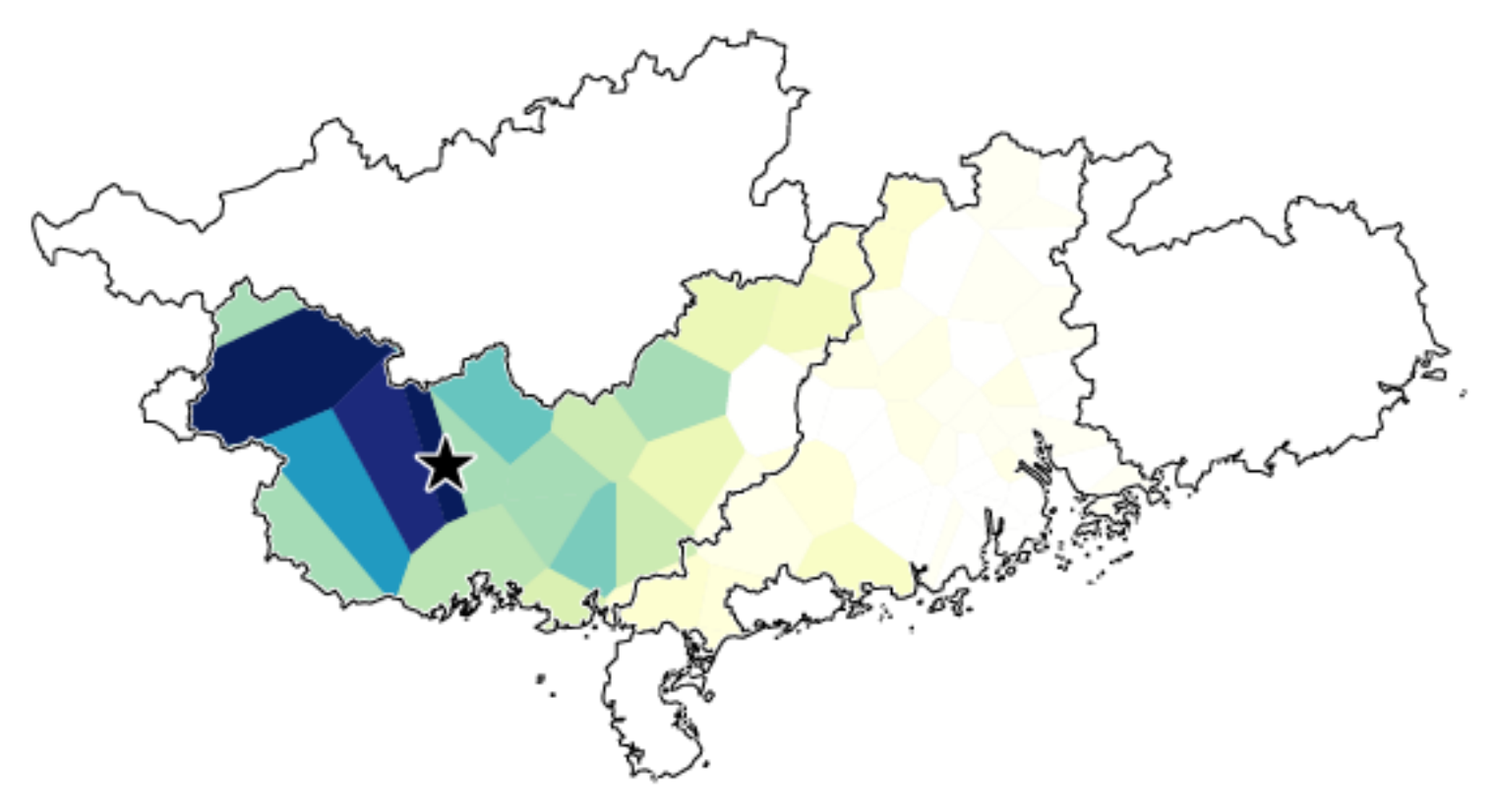
Reference Point: Taishan



Reference Point: Guangning



Reference Point: Nanning (Pinghua)



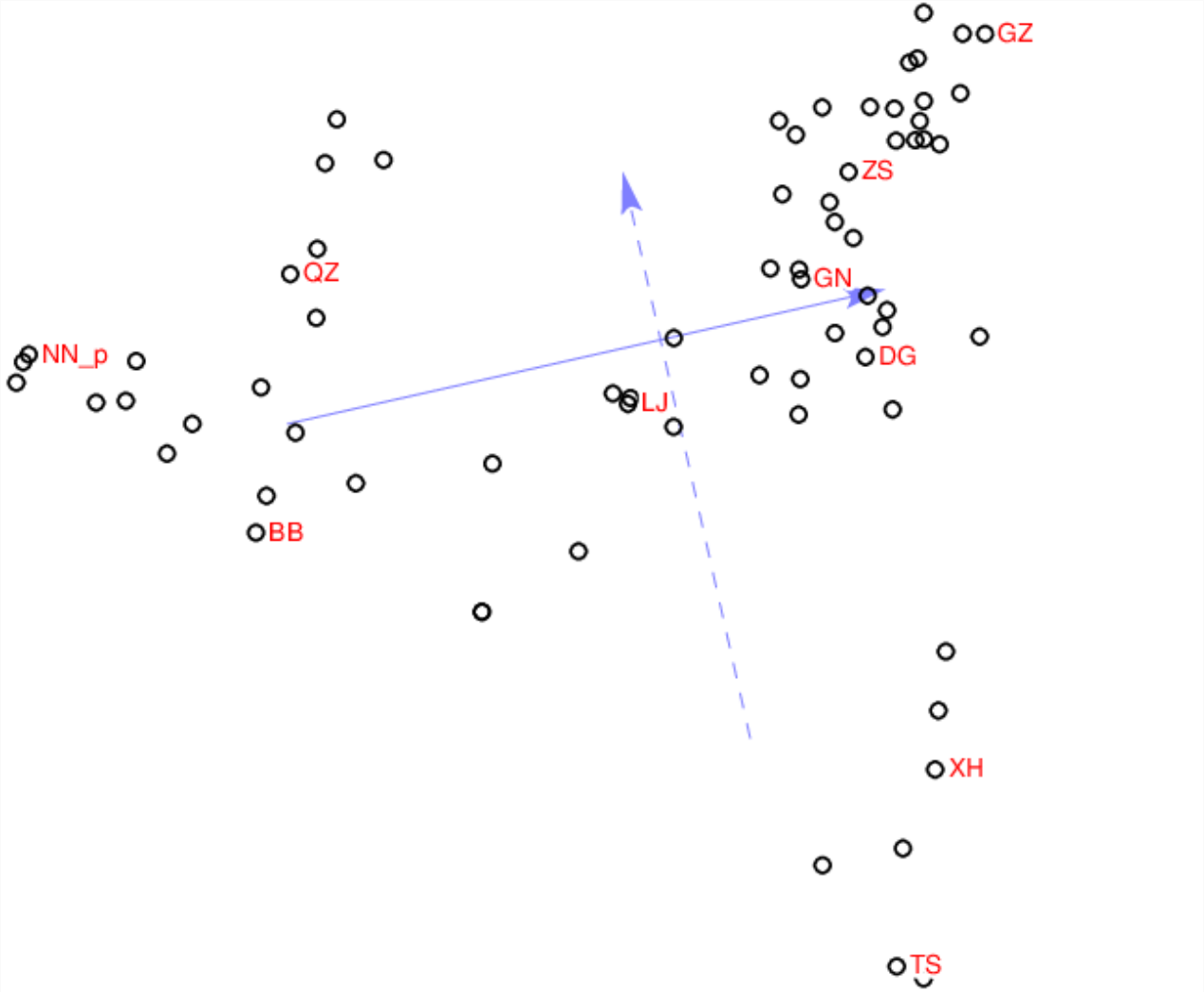
Multidimensional Scaling

Multidimensional Scaling

Kruskal (1964)

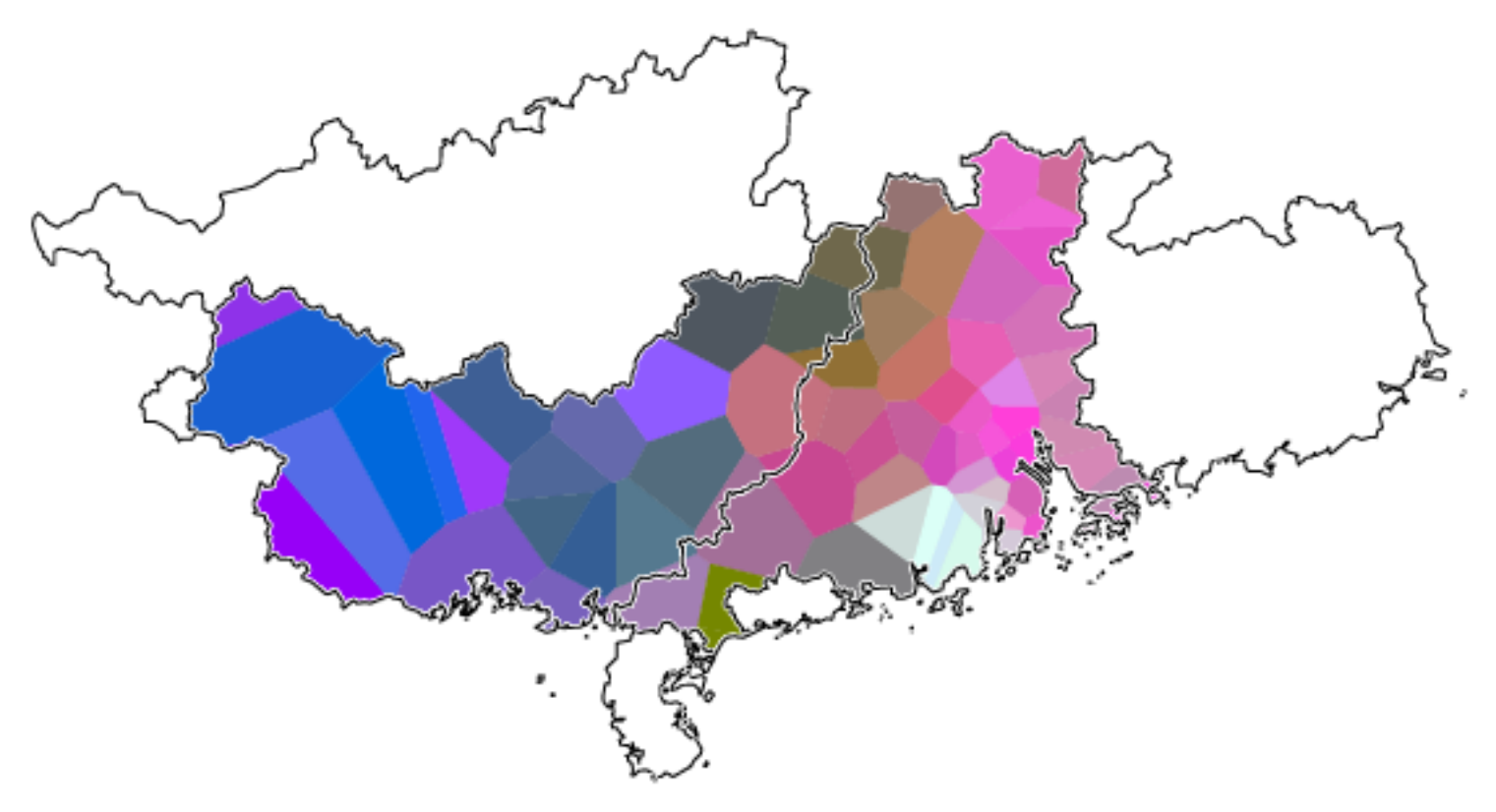
- Multidimensional Scaling = MDS
- Technique to estimate relative positions of points in an arbitrary multidimensional space using relative distances as input
- Useful for understanding the gradual nature of boundaries, but is a bit of sensory overload
- Not always precise

In Two Dimensions



r = 0.76

In Three Dimensions



r = 0.81

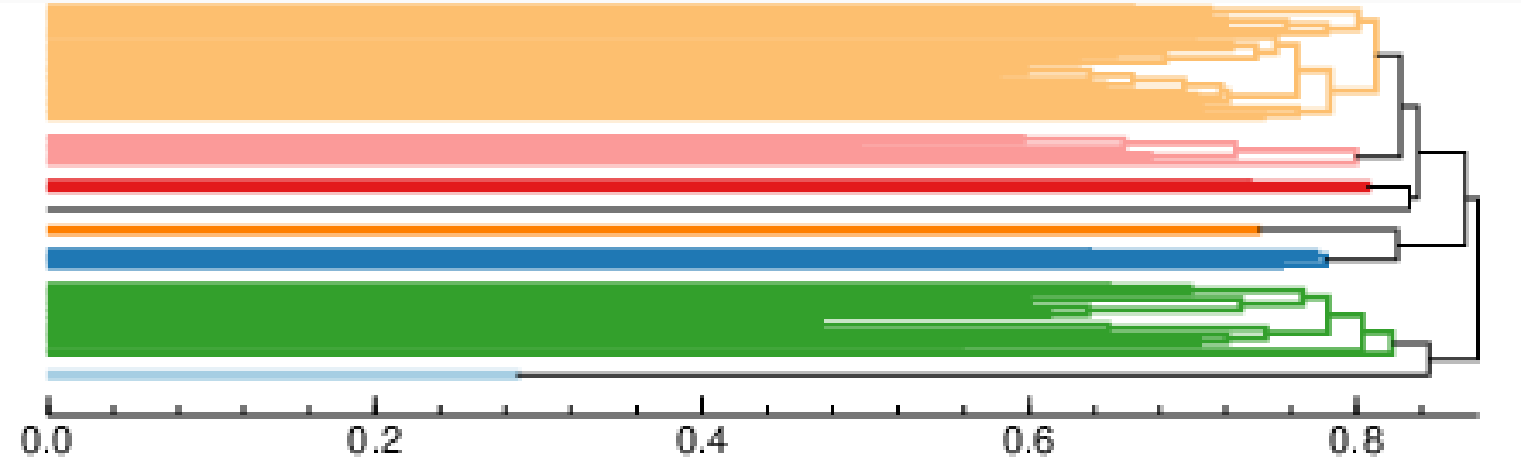
Clustering

Discrete Clustering

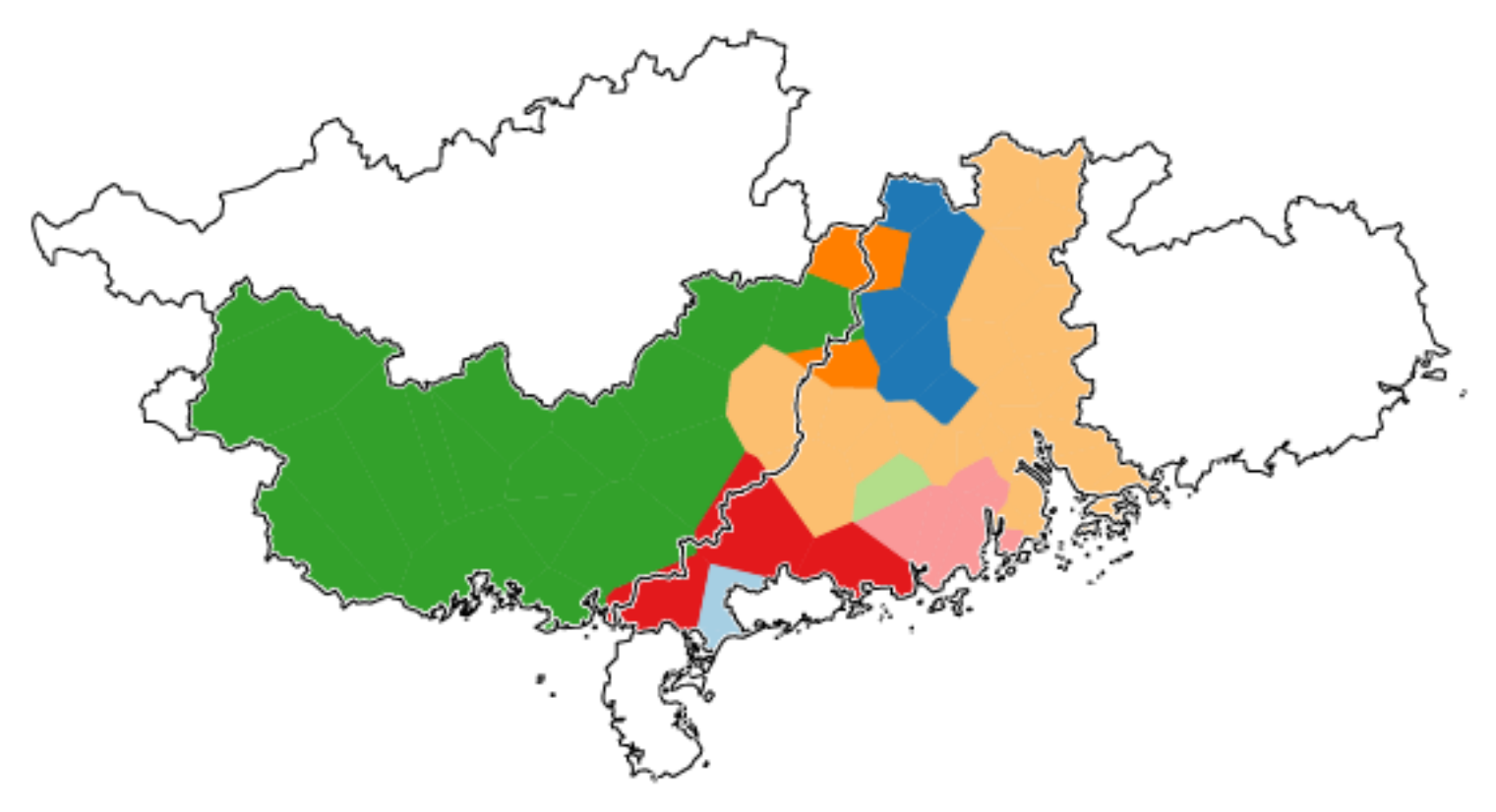
- Given the distances between sites, cluster sites so that sites in the same cluster are more linguistically similar to each other than to those in other clusters.
- Prone to produce very different results based on even small fluctuations in the data.

UPGMA	Unweighted Pair Group Method using Arithmetic averages cophenetic distances (distances in the clusters) match original distances most closely
WPGMA	Weighted Pair Group Method using Arithmetic averages for irregular distribution
s Method	Minimum Variance Gives clusters of roughly even size

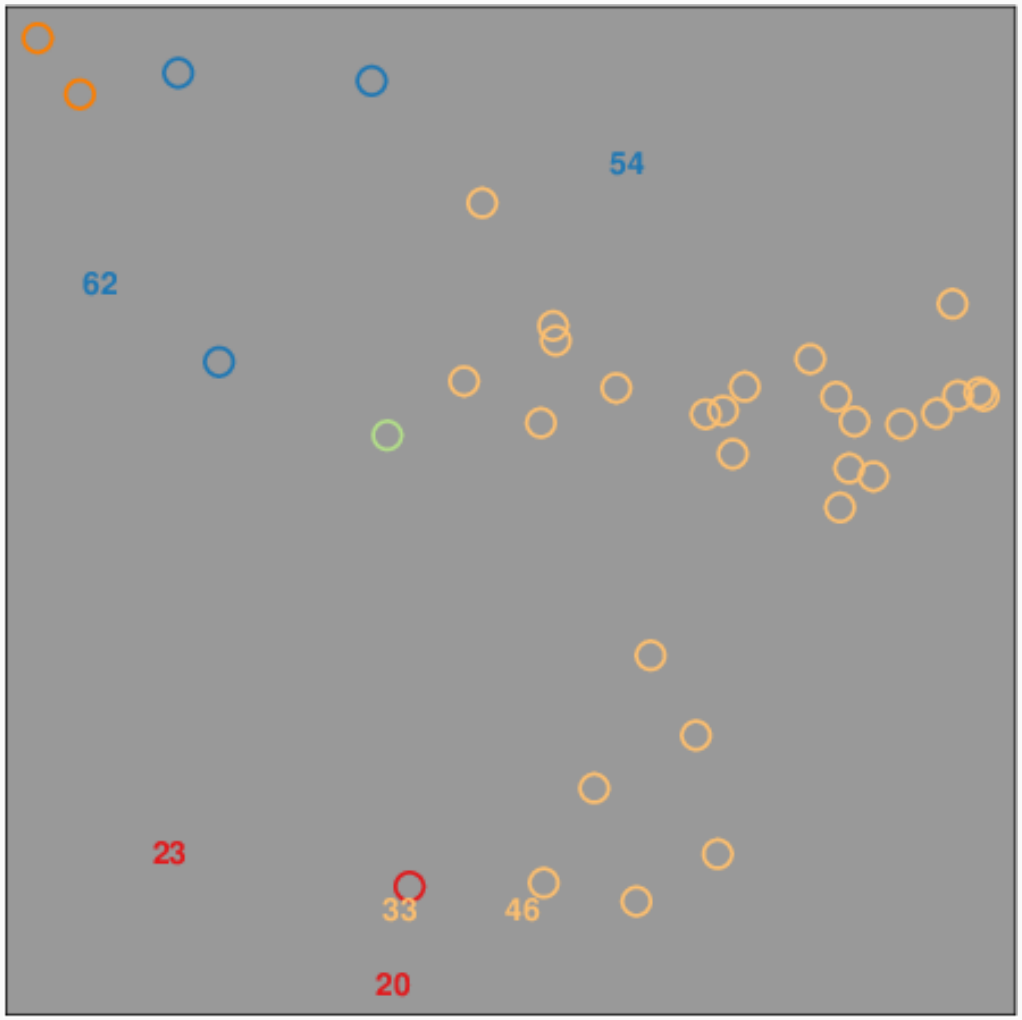
Weighted Average Dendrogram



Weighted Average Map



Cluster Verification

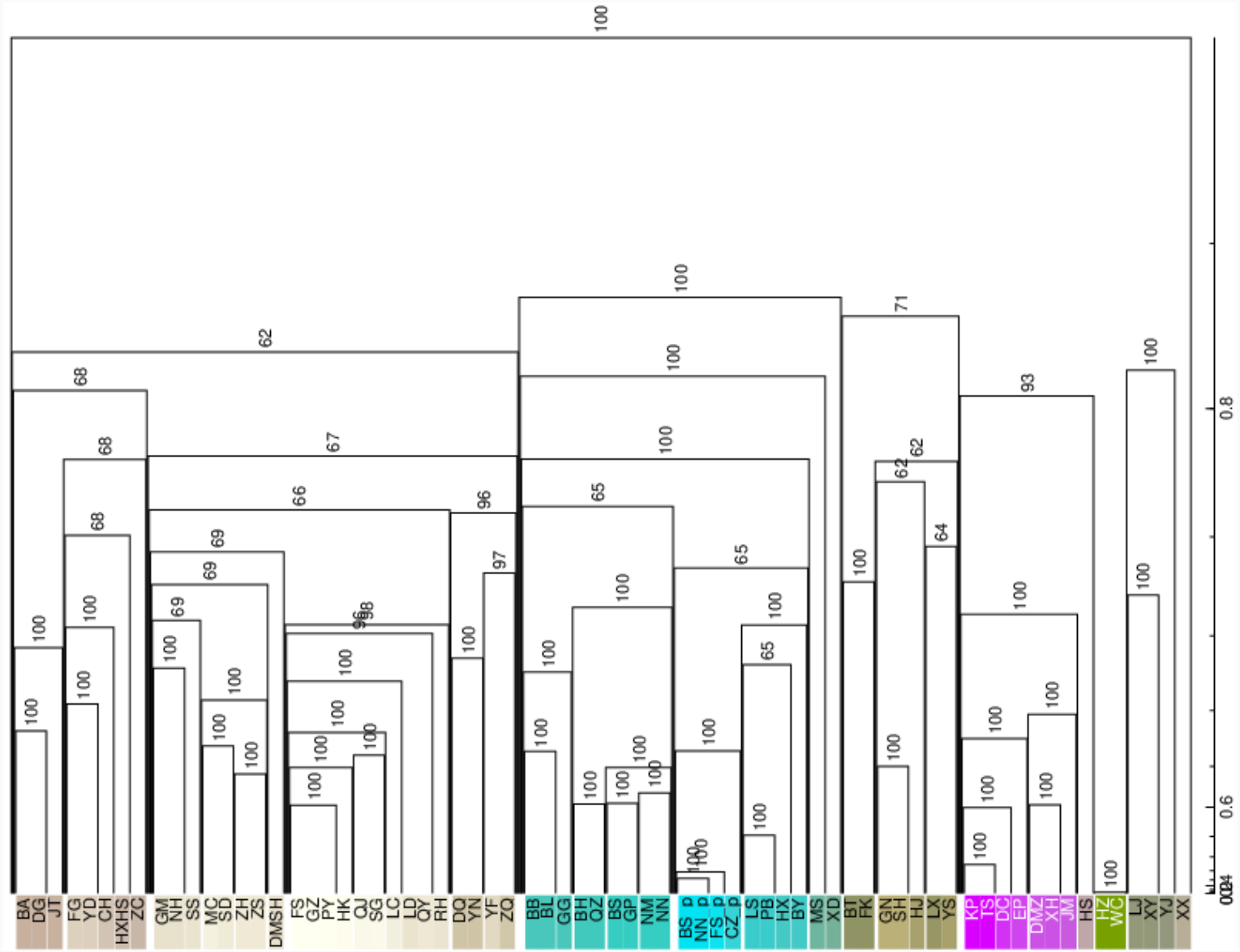


Fuzzy Clustering

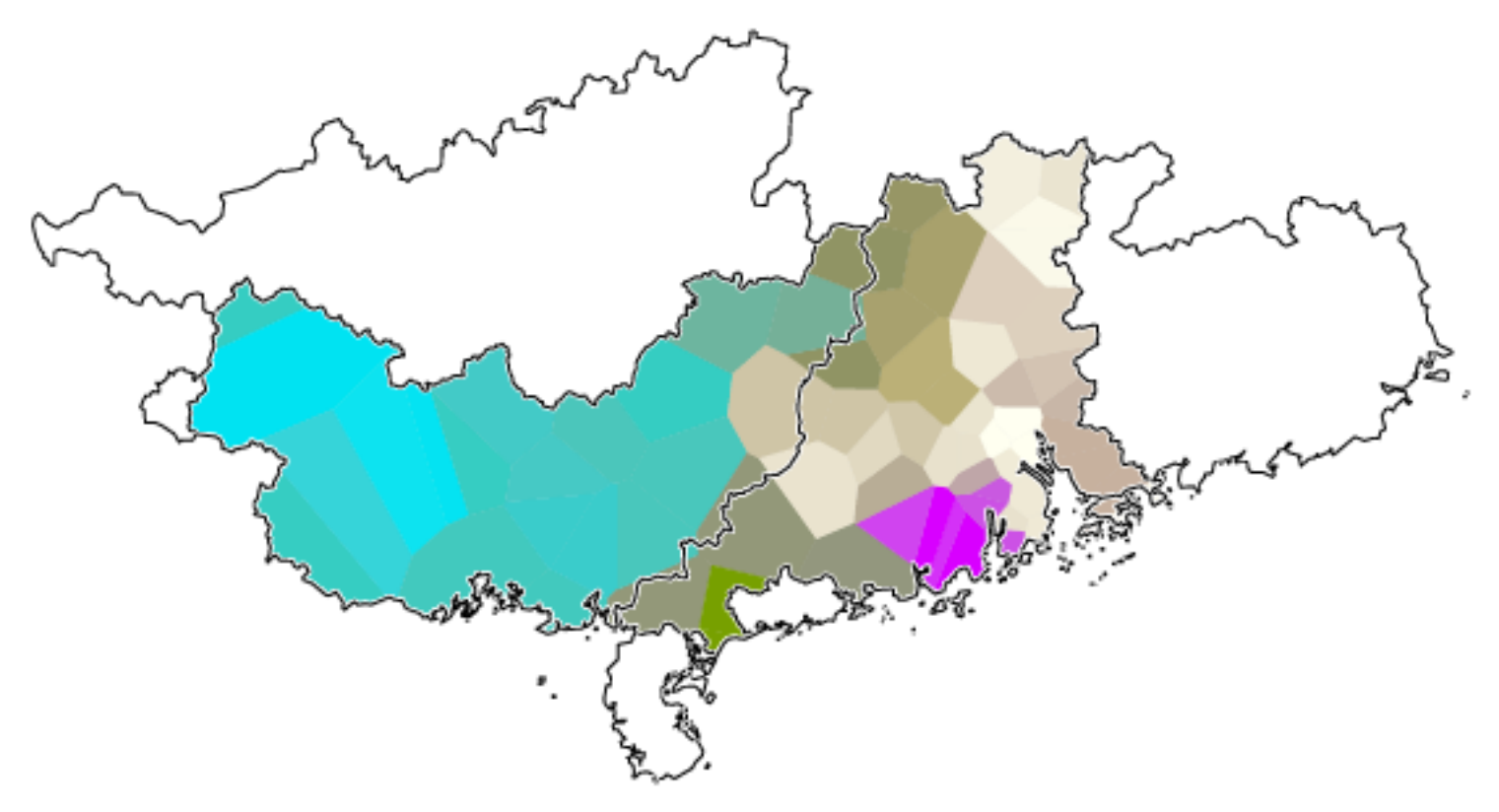
Nerbonne et al. et al. (2008)

- Prevent instability by repeatedly clustering while introducing noise. Clusters that occur the most often are the stablest.
- Process:
  - Cluster repeatedly (n=100), randomly adding noise to the distance matrix ($0 \leq r \leq 0.2$) each iteration
  - Count how many times certain clusters form in these repeated clusterings to approximate certainty of clustering
  - Combine analysis into composite cluster
- Can project results to geography using cophenetic distances

Probabilistic Dendrogram



Fuzzy Clustering Map



Wrap-up

Bibliography i

Carlyle, John. 2020. “Common Yue: A Comparative Study of Yue Dialect Historical Phonology.” Master’s thesis, University of Washington.

Goebl, Hans. 1984. *Dialektometrische Studien: anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. 3 vols. Tübingen: M. Niemeyer. ISBN: 978-3-484-52191-9 978-3-484-50200-0 978-3-484-50201-7 978-3-484-50202-4.

Bibliography ii

- Kruskal, Joseph B. 1964. “Nonmetric Multidimensional Scaling: A Numerical Method.” *Psychometrika* 29 (2): 115–129.
- Levenshtein, Vladimir I. 1966. “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.” In *Soviet Physics Doklady*, 10:707–710. 8. Soviet Union.
- Lǐ, Jiàn. 2014. *Wú Huà yuèyǔ yánjiū 吳化粵語研究*. Zhānjiāng shīfàndàxué zhōngguó yǔyánxué xuékē xīn shìyě xuéshù wéncóng. Zhōngguó shèhuìkēxué chūbǎnshè.

Bibliography iii

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. “Gabmap-a Web Application for Dialectology.” *Dialectologia: revista electrònica,* 65–89.

Nerbonne, John, and Peter Kleiweg. 2003. “Lexical Distance in LAMSAS.” *Computers and the Humanities* 37 (3): 339–357.

———. 2007. “Toward a Dialectological Yardstick.” *Journal of Quantitative Linguistics* 14, nos. 2-3 (August): 148–166. ISSN: 0929-6174, 1744-5035. https://doi.org/10.1080/09296170701379260.

Bibliography iv

- Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. “Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering.” In *Data Analysis, Machine Learning and Applications*, 647–654. Springer.
- Séguy, Jean. 1971. “La Relation Entre La Distance Spatiale et La Distance Lexicale.” *Revue de Linguistique Romane* 35:335–357.
- Xiǎn, Wéntíng. 2016. “Guǎngdōng yángjiānghuà yánjiū 廣東陽江話研究.” Master’s thesis, Guǎngxī dàxué.

Bibliography v

- Xiè, Jiànyóu. 2007. *Guǎngxī Hànyǔ fāngyán yánjiū 廣西漢語方言研究*. 2 vols. Guǎngxī rénmín chūbǎnshè.
- Yang, Cathryn, and Andy Castro. 2008. “Representing Tone in Levenshtein Distance.” *International Journal of Humanities and Arts Computing* 2 (1-2): 205–219.
- Yue-Hashimoto, Anne. 1988. “A Preliminary Investigation into the Subclassification Problem of the Yue Dialects.” *Ajia-Afurikago no keisū kenkyū* (Tōkyō) 30:7–42.
- ———. 2005. *The Dancun Dialect of Taishan*. Language Information Sciences Research Centre, City University of Hong Kong.

Bibliography vi

- Zhān, Bóhuì, et al. 2000. *Guǎngdōng Yuè fāngyán gàiyào 廣東粵方言概要*. Jìnán dàxué chūbǎnshè.
- Zhān, Bóhuì, Yat-shing Cheung, et al. 1987a. *Zhūjiāng sānjiǎozhōu fāngyán cíhuì duìzhào 珠江三角洲方言詞彙對照*. Guǎngdōng Rénmín Chūbǎnshè.
- ———. 1987b. *Zhūjiāng sānjiǎozhōu fāngyán zìyīn duìzhào 珠江三角洲方言字音對照*. Guǎngdōng Rénmín Chūbǎnshè.
- ———. 1994. *Yuèběi shí xiànshì Yuè fāngyán diàochá bàogào 粵北十縣市粵方言調查報告*. Jìnán Dàxué Chūbǎnshè.

Bibliography vii

Zhān, Bóhuì, Yat-shing Cheung, et al. 1998. *Yuèxī shí xiànshì Yuè fāngyán diàochá bàogào 粵西十縣市粵方言調查報告*. Jìnán Dàxué Chūbǎnshè.

Questions

Thank you