

Detecting linguistic variation with geographic sampling

Ezequiel Koile, George Moroz

Linguistic Convergence Laboratory, NRU HSE

... 2020

Presentation is available here: tinyurl.com/y7kjsp67



Outline of the talk

Introduction

Our approach

Simulated data

Conclusion

- Geolectal variation is often present in settings where one language is spoken across a vast geographic area [[Labov 1963](#)].
- It can be found in phonological, morphosyntactic, and lexical features.
- Could be overlooked by linguists [[Dorian 2010](#)].

The problem

- Let us consider a geographical dialect continuum formed by a group of small villages [Chambers and Trudgill 2004: 5–7]
- We are interested in spotting variation of a discrete parameter among the lects spoken on these villages

MAP OF FEATURE ON VILLAGES

- We will very unlikely be able to conduct fieldwork in each single village. Therefore, we need to choose a *sample* of locations.
- *Research Question:* How to choose the sample of villages to survey?
 - 1 How many villages is enough for spotting variation?
 - 2 Given an amount of sampled villages, how to decide which ones are representative of our population?

Outline of the talk

Introduction

Our approach

Simulated data

Conclusion

Our approach

- We assume that we want to find the distribution of variation for one feature, and we try different ways of choosing the sampled villages for finding it:
- As we assume we don't have any data beyond the geographic location of each village, we use these locations for building our sample
- We generate clusters with different algorithms (k-means, hierarchical clustering) and pick our sampled locations based on them (package stats, [[R Core Team 2020](#)]).
- We compare our results with random sampling for two different scenarios:
 - Binary categories for simulated data with different distributions
 - Multiple categorical data for Circassian languages

Outline of the talk

Introduction

Our approach

Simulated data

Conclusion

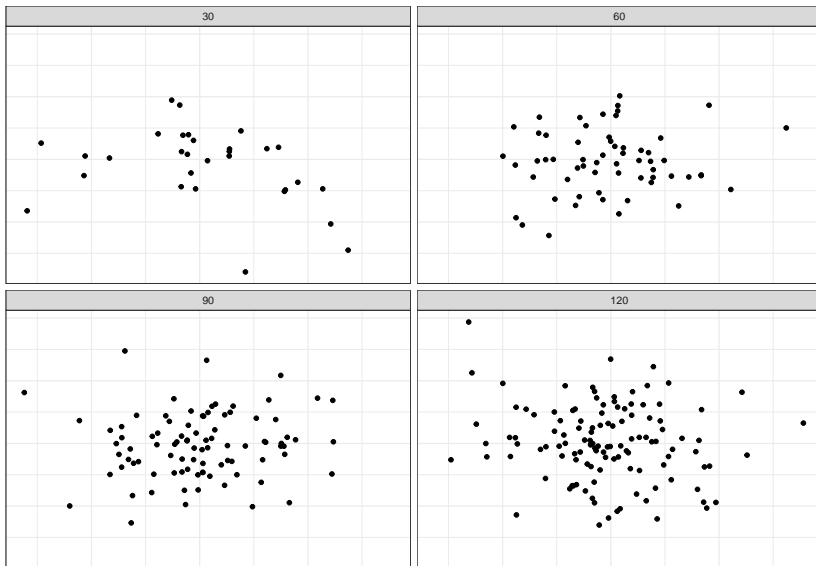
Simulated data

- total number of locations (N): 30, 60, 90, 120
- type of spatial relations:
 - random
 - two more or less separable regions
 - central and periphery
- proportion of variation in the explored variable (p): 0.1, 0.2, 0.3, 0.4, 0.5
- amount of clusters (k): 2, ... $N/2$
- percentage of observations taken from each cluster (r): 0.1, 0.2, ...

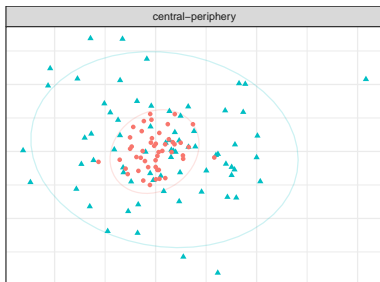
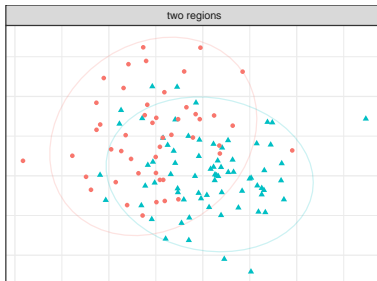
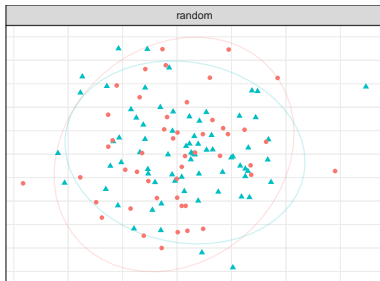
From those values we could derive a number of sampled locations (n):

$$n = N \times r$$

Example of different number of locations (N)



Example of different type of spatial relations

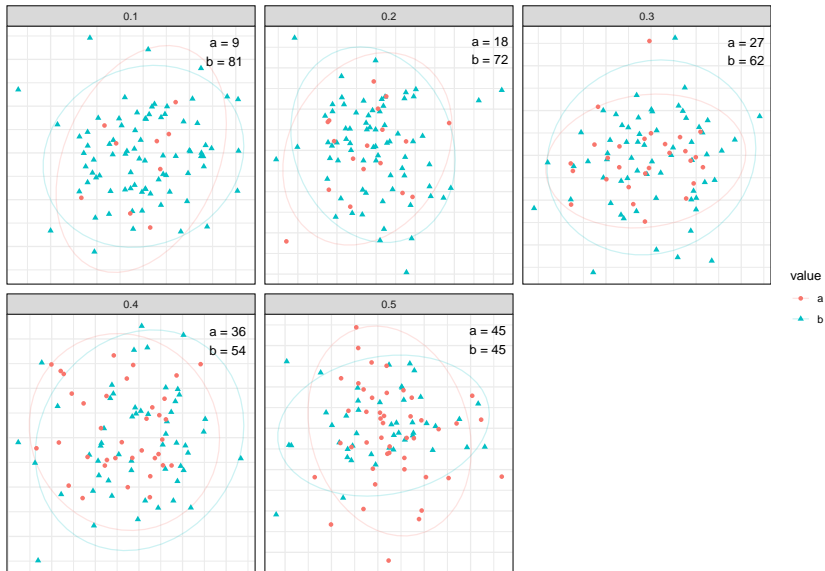


value

● a

▲ b

Example of different proportions of variation in the explored variable (p)



Outline of the talk

Introduction

Our approach

Simulated data

Conclusion

- Chambers, J. K. and Trudgill, P. (2004). *Dialectology, 2nd edition*. Cambridge University Press.
- Dorian, N. C. (2010). *Investigating variation: The effects of social organization and social setting*. Oxford University Press.
- Labov, W. (1963). The social motivation of a sound change. *Word*, 19(3):273–309.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.