

Detecting linguistic variation with geographic sampling

Ezequiel Koile, George Moroz

Linguistic Convergence Laboratory, NRU HSE

... 2020



Outline of the talk

- introduction
- blar

- 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.
- Often overlooked by linguists ([Dorian \(2010\)](#)).

ADD SOME MAPS

The problem

- Let us consider a geolectal continuum formed by a group of small villages ([Chambers and Trudgill \(1998\)](#))

FLAT MAP OF VILLAGES

- We are interested in spotting variation of a certain 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?

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, [Team et al. \(2013\)](#)).
- 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

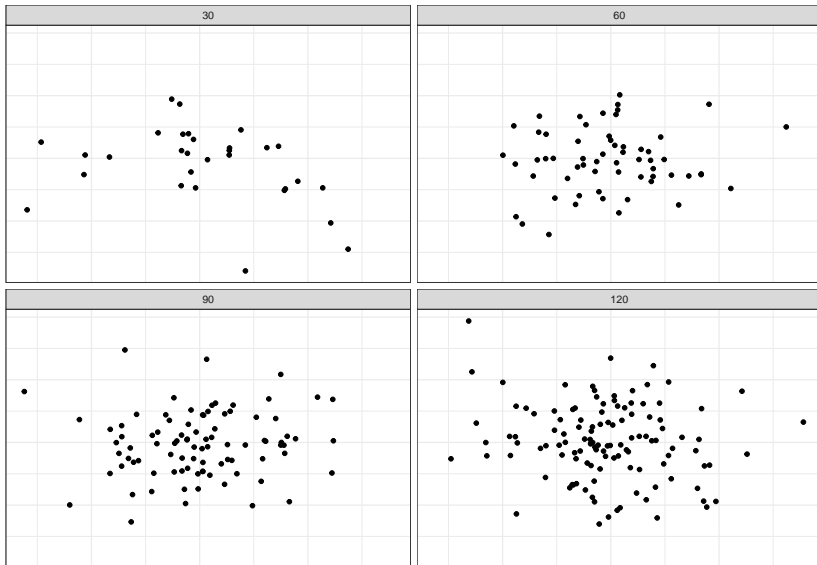
Simulated data

- 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, ...

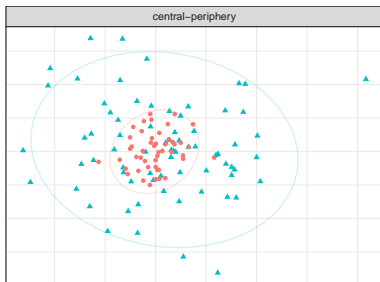
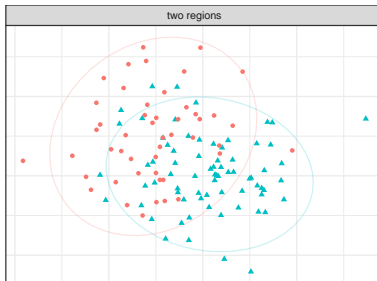
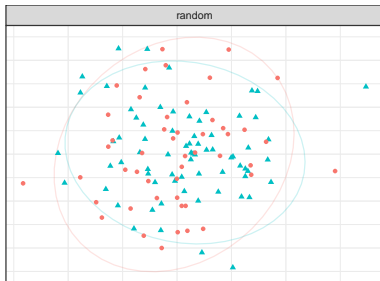
From those values we could derive a number of sampled locations (n):

$$n = N \times r$$

Example of different number of locations (N)



Example of different type of spatial relations

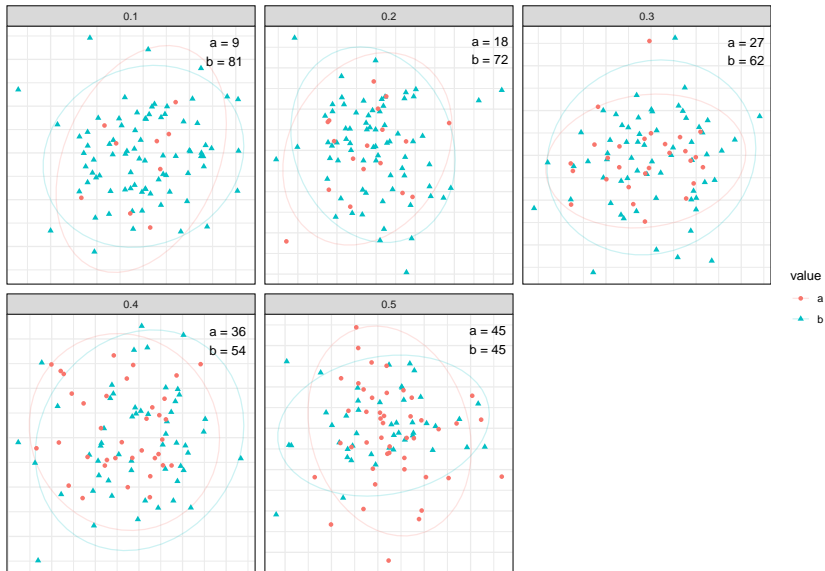


value

• a

▲ b

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



Conclusion

- Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. 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.
- Team, R. C. et al. (2013). R: A language and environment for statistical computing.