

# Detecting linguistic variation with geographic sampling

Ezequiel Koile, George Moroz

Linguistic Convergence Laboratory, NRU HSE

26 August 2020

Presentation is available here: [tinyurl.com/y7kjsp67](https://tinyurl.com/y7kjsp67)



# Outline of the talk

Introduction

Our approach

Simulated data

Circassian data example

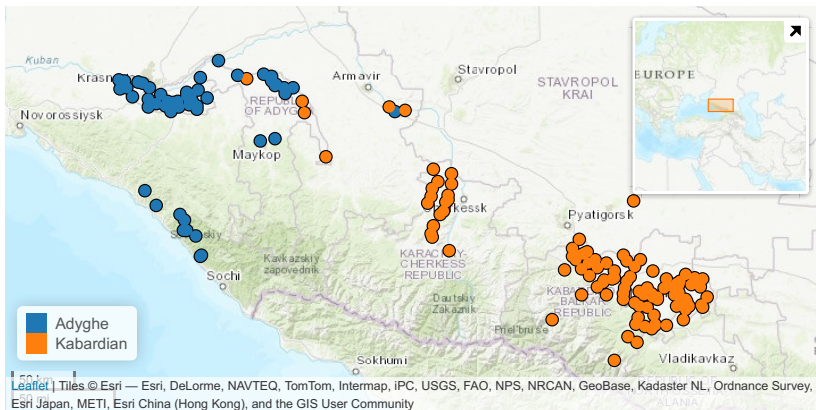
Conclusion

- 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

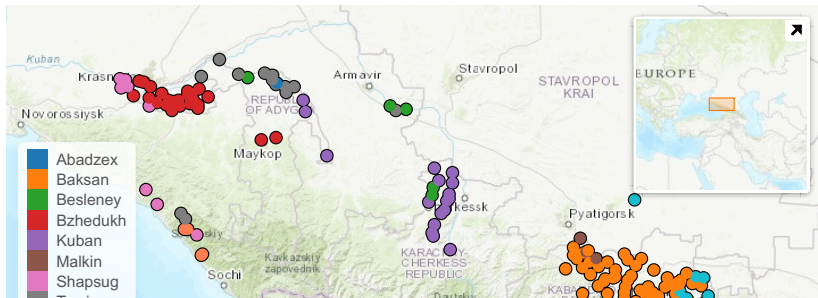
# The problem

- 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



# The problem

- 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 detecting all variation present? (number of categories)
  - 2 How many villages is enough for estimating the variation? (proportion of each category)
  - 3 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

Circassian data example

Conclusion

## Our approach

- 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 against random sampling in two different scenarios, both for simulated and for real Circassian data:
  - Multiple categorical data (detect variation)
  - Binary categorical data (estimate variation)

## Information entropy

In order to measure the diversity of the questions we used the easiest measure — information entropy, introduced in [Shannon 1948]:

$$H(X) = - \sum_{i=1}^n P(x_i) \times \log_2 P(x_i)$$

The range of the information entropy is  $H(X) \in [0, +\infty]$ :

data	entropy
A-A-A-A-A	0.00
A-A-A-A-B	0.72
A-A-A-B-B	0.97
A-A-B-B-B	0.97
A-A-B-B-C	1.52
A-B-C-A-B	1.52
A-B-C-D-E	2.32



# Outline of the talk

Introduction

Our approach

**Simulated data**

Circassian data example

Conclusion

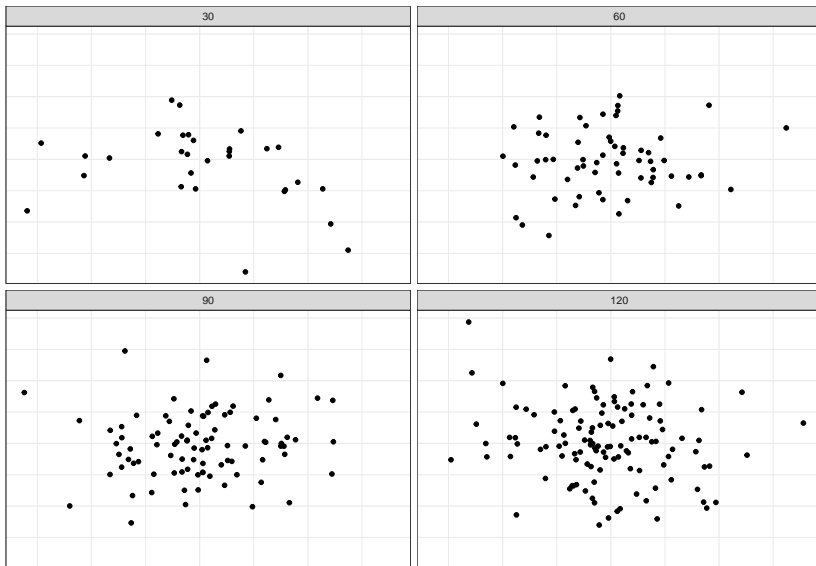
## Simulated data

- 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, ...

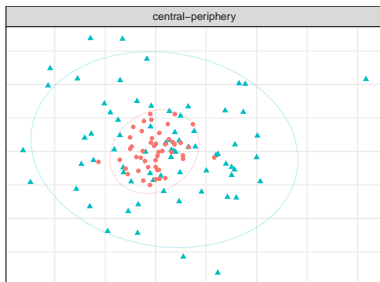
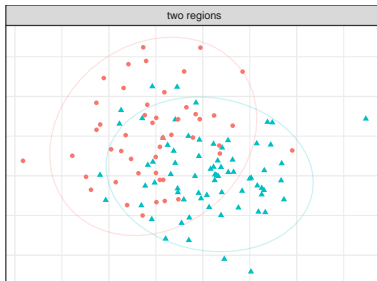
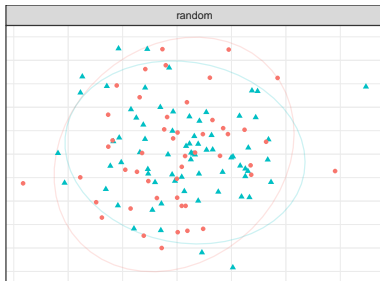
From those values we could derive a number of sampled locations ( $n$ ):

$$n = N \times r$$

## Example of different number of locations ( $N$ )



# Example of different types of spatial relations

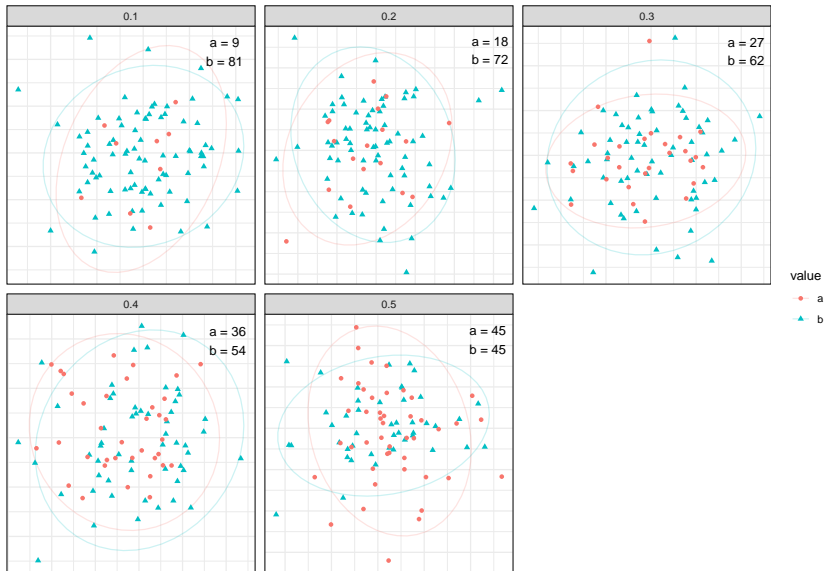


value

• a

▲ b

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



# Outline of the talk

Introduction

Our approach

Simulated data

Circassian data example

Conclusion

# Outline of the talk

Introduction

Our approach

Simulated data

Circassian data example

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.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.