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

Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

Introduction

- 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].

Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

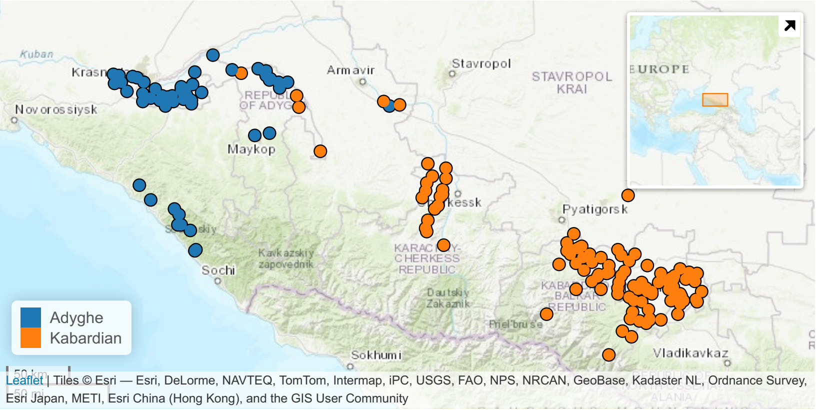
Circassian data example

Entropy

Conclusions

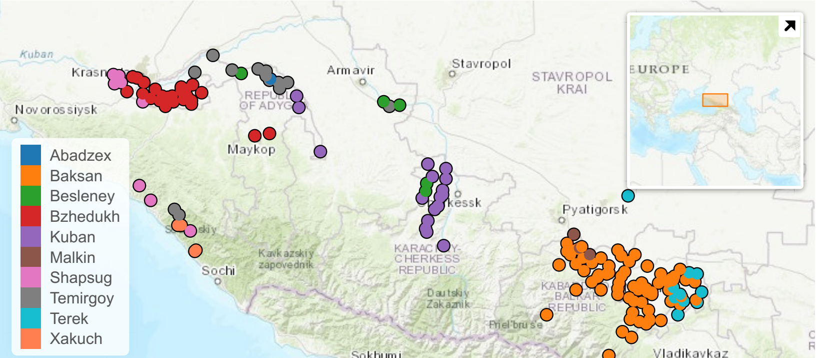
The problem

- 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. Given an amount of sampled villages, how to decide which ones are representative of our population?



Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

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)

Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

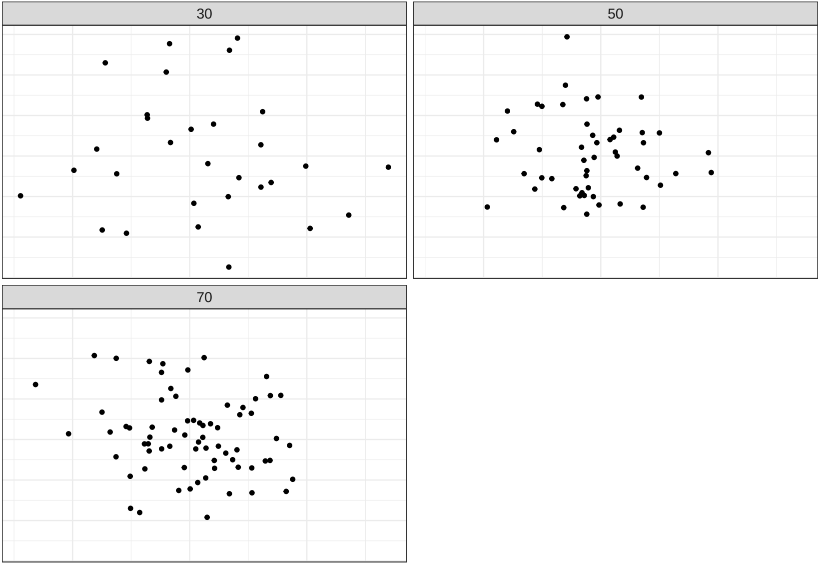
Simulated data

- total number of locations (N): $30, 50, 70$
- type of spatial relations:
  - uniformly distributed
  - equadistant – ...
  - central and periphery
- number of categories (???): $3, 4, 5$
- proportion of categories (???): e. g. $20 - 16 - 14, 9 - 8 - 7 - 5 - 1$
- proportion of variation in the explored variable (p): $0.05, 0.10 \dots 0.90$
- amount of clusters (k): $N \times 0.05, N \times 0.10, \dots N \times 0.90$

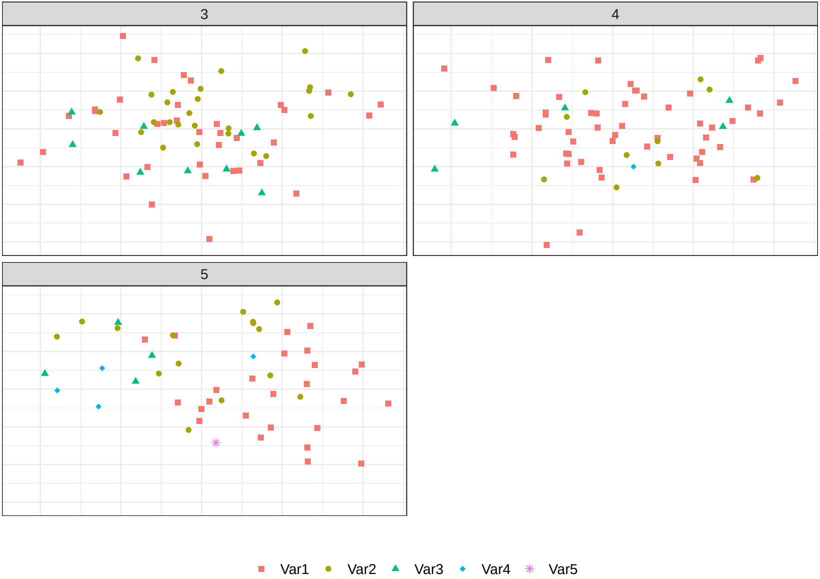
From those values we could derive a number of sampled locations (n):

$$n = N \times r$$

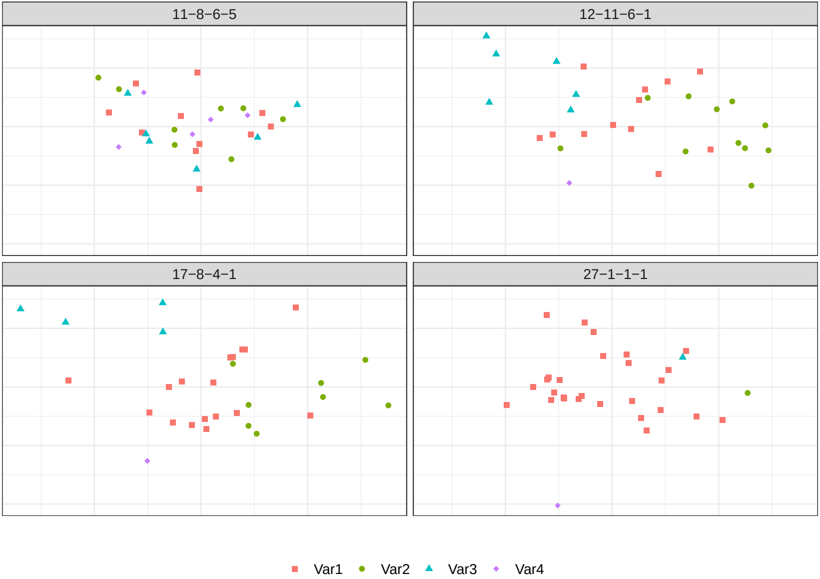
Example of different number of locations (N)



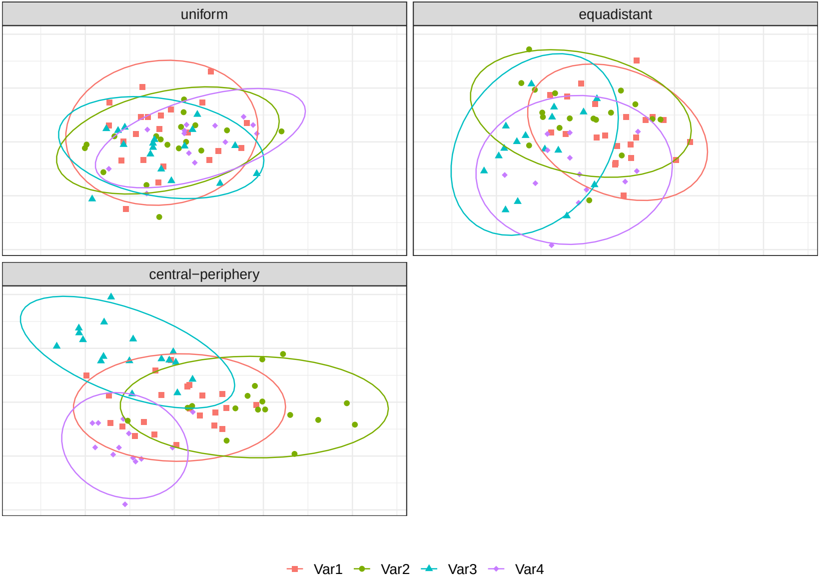
Example of different number of categories



Example of different proportion of categories



Example of different types of spatial relations



Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

Results

- Plots of fitting with different parameters

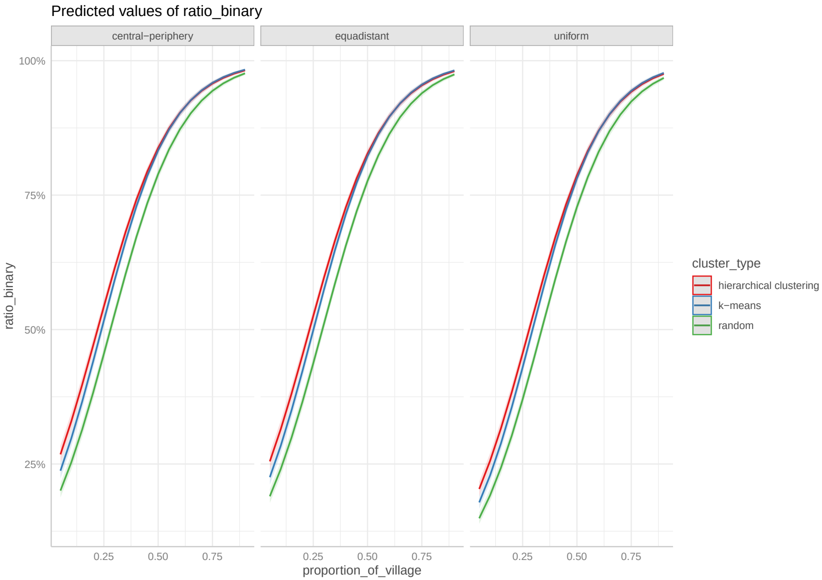
Modelling the variation

- We want to account quantitatively for the improvement of our fitting according to the different parameters present
- We model a logistic regression (values: all variation discovered / not all variation discovered), with the following parameters modifying the independent variable
    - Type of clustering (hirarchical, K-means, random)
    - Amount of categories (numeric: 3,4,5)
    - Type of geographic distribution (central-periphery, separated clusters, random)
    - Amount of total villages (numeric: 30, 50, 70)
    - Entropy (numeric)

Results

term	estimate	std.error	statistic	p.value
(Intercept)	-2.04	0.05	-42.96	0.00
typeequadistant	0.30	0.05	5.74	0.00
typecentral-periphery	0.36	0.05	7.04	0.00
cluster_typek-means	0.21	0.05	4.04	0.00
cluster_typehierarchical clustering	0.38	0.05	7.51	0.00
proportion_of_village	6.06	0.11	54.26	0.00
typeequadistant:proportion_of_village	-0.07	0.13	-0.55	0.59
typecentral-periphery:proportion_of_village	-0.05	0.13	-0.40	0.69
cluster_typek-means:proportion_of_village	0.16	0.13	1.26	0.21
cluster_typehierarchical clustering:proportion_of_village	-0.12	0.12	-0.97	0.33

Results



Predicted values of ratio_binary
central−periphery
equadistant
uniform
100%
75%
50%
25%
ratio_binary
0.25
0.50
0.75
proportion_of_village
cluster_type
hierarchical clustering
k−means
random

Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

Information entropy

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

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

Information entropy

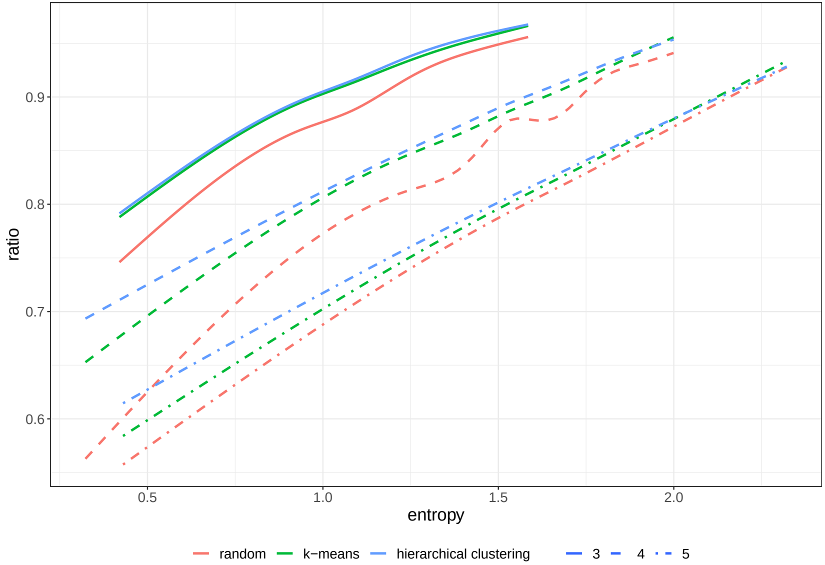
In order to measure the diversity of the questions we used the easiest measure — information entropy, introduced in [Shannon 1948]:

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

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

Information entropy: simulated data



Outline of the talk

Introduction

The problem

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

Conclusions

References

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.