# Detecting linguistic variation with geographic sampling

Ezequiel Koile, George Moroz
Linguistic Convergence Laboratory, NRU HSE

14 December 2020

# Outline of the talk

Introduction

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

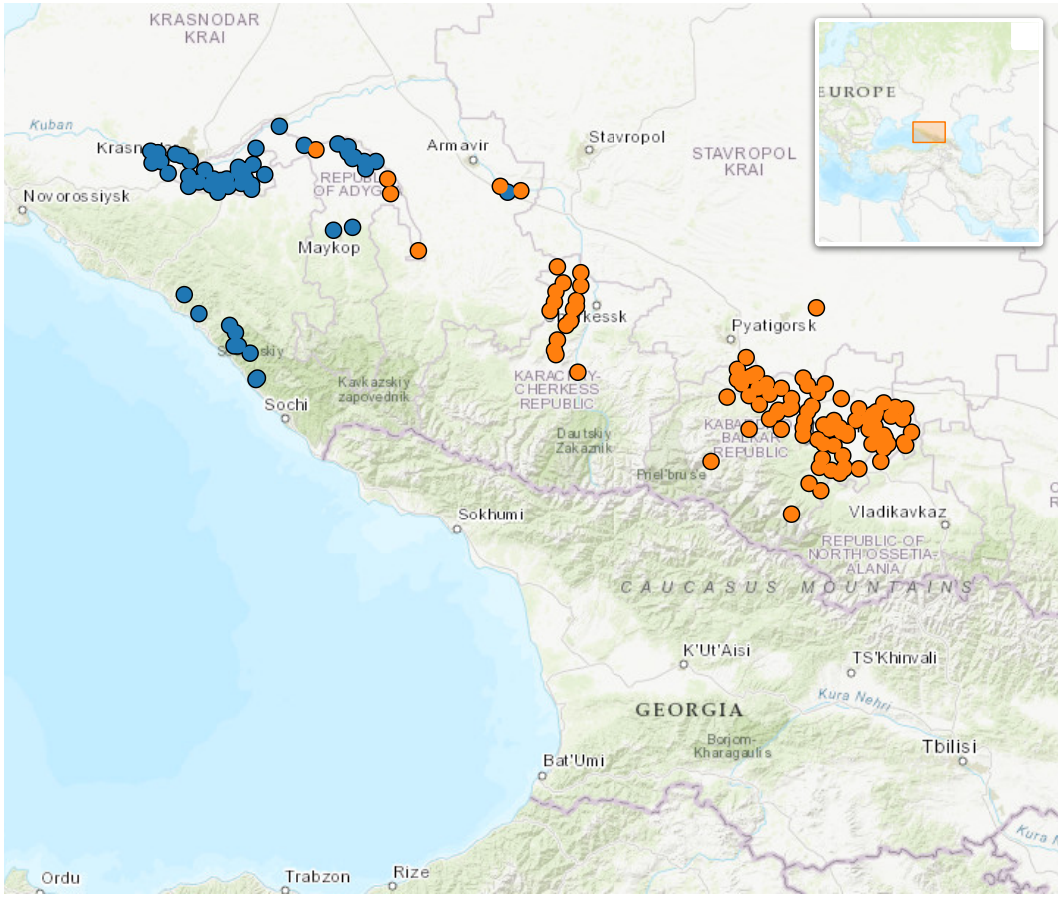
Conclusions

# 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.
- It could be overlooked by linguists [Dorian 2010].

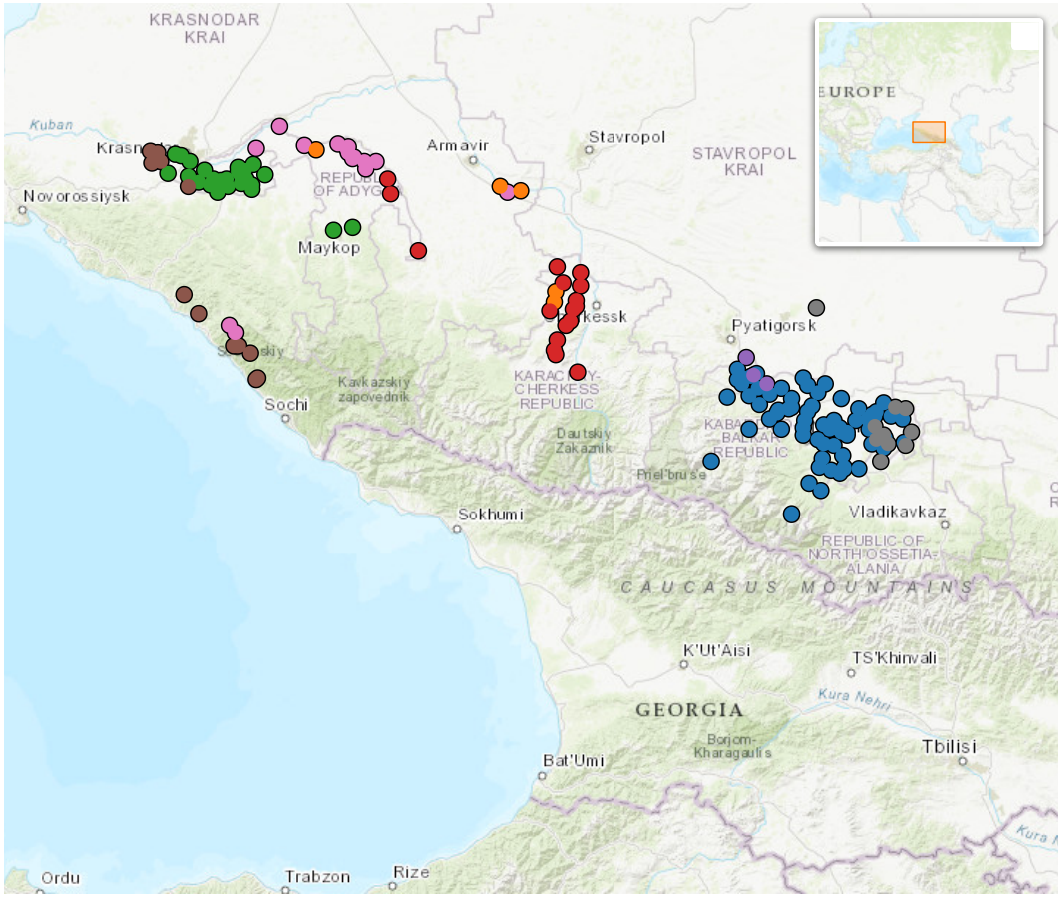
# The problem

- 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 in these villages



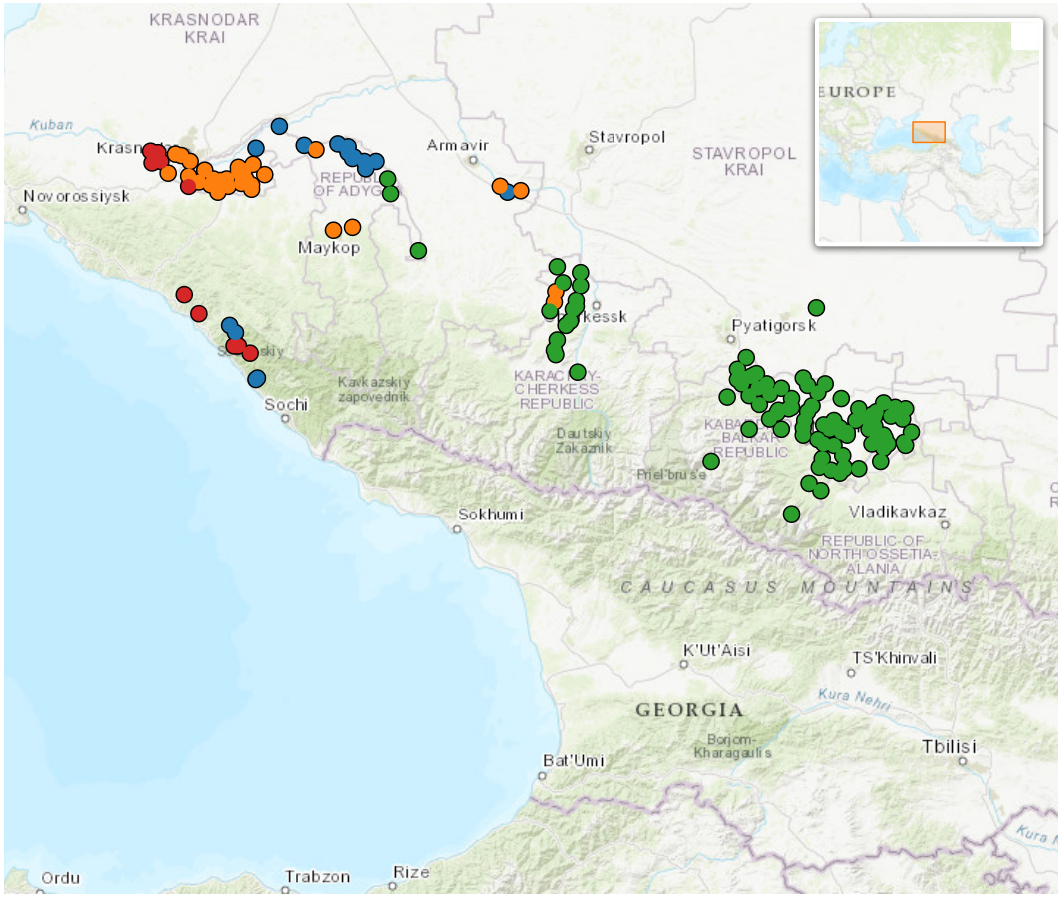
# The problem

- It is very impractical 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?



# The problem

- It is very impractical 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?



# Outline of the talk

Introduction

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

# Our approach

- We want to find the amount of variation present for a given feature. Therefore, we try different ways of choosing the villages to sample for detecting this variation.
- As we assume we do not 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 geographic sampling for multiple categorical data, in two different scenarios:
  - Simulated data
  - Dialects of Circassian languages

# Outline of the talk

# Simulated data

## Data generation

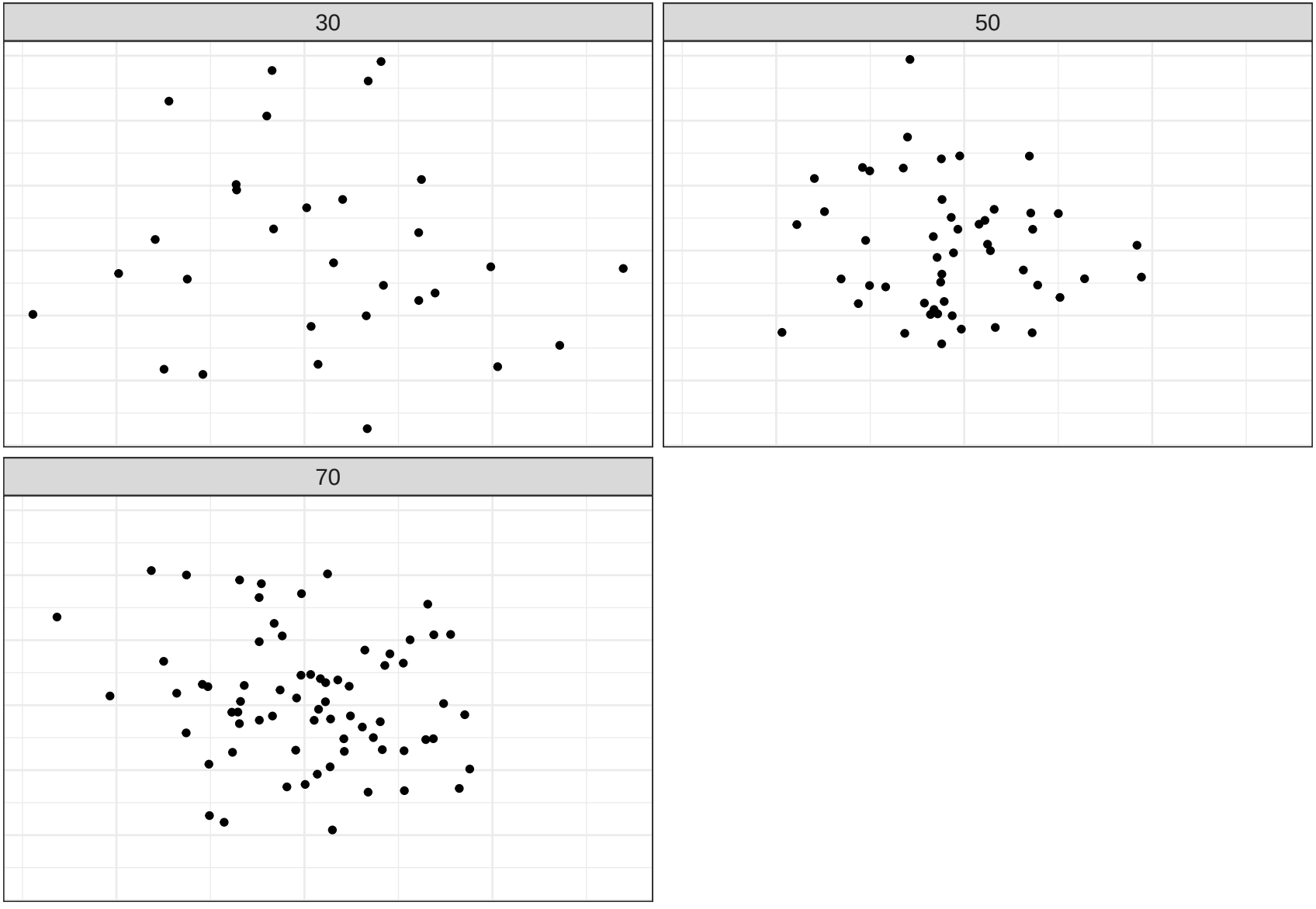
- total number of locations ($N$): $30, 50, 70$
- number of categories ($n$): $3, 4, 5$
- type of spatial relation:
  - uniform: variation is uniformly distributed across space
  - equadistant: $n$ groups with unique values, partially overlaping
  - central-periphery: one main group in the center, the rest around it
- count configuration ($c$): how the $n$ categories are distributed across the $N$ locations (e.g., for $N$=30, $n$=3, the count configuration could be $c$=10-10-10, $c$=20-8-2, etc.)

## Sampling

- clustering method: hierarchical clustering, $k$-means, random sampling
- proportion of villages sampled: $p = 0.05, 0.10, \ldots, 0.90$

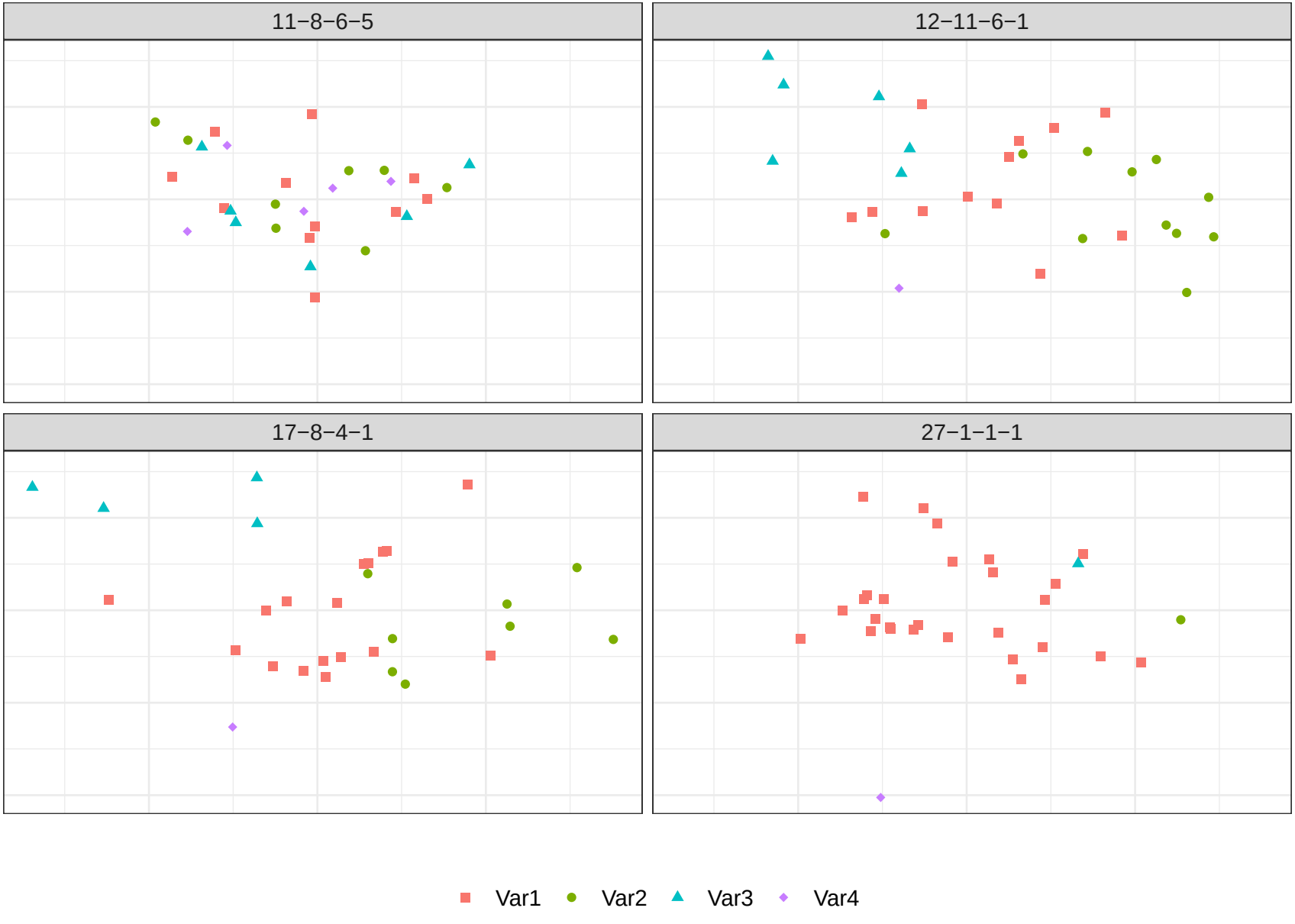
From those values we could derive the number of sampled locations, or number of clusters ($k$): $k = p \times N$

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

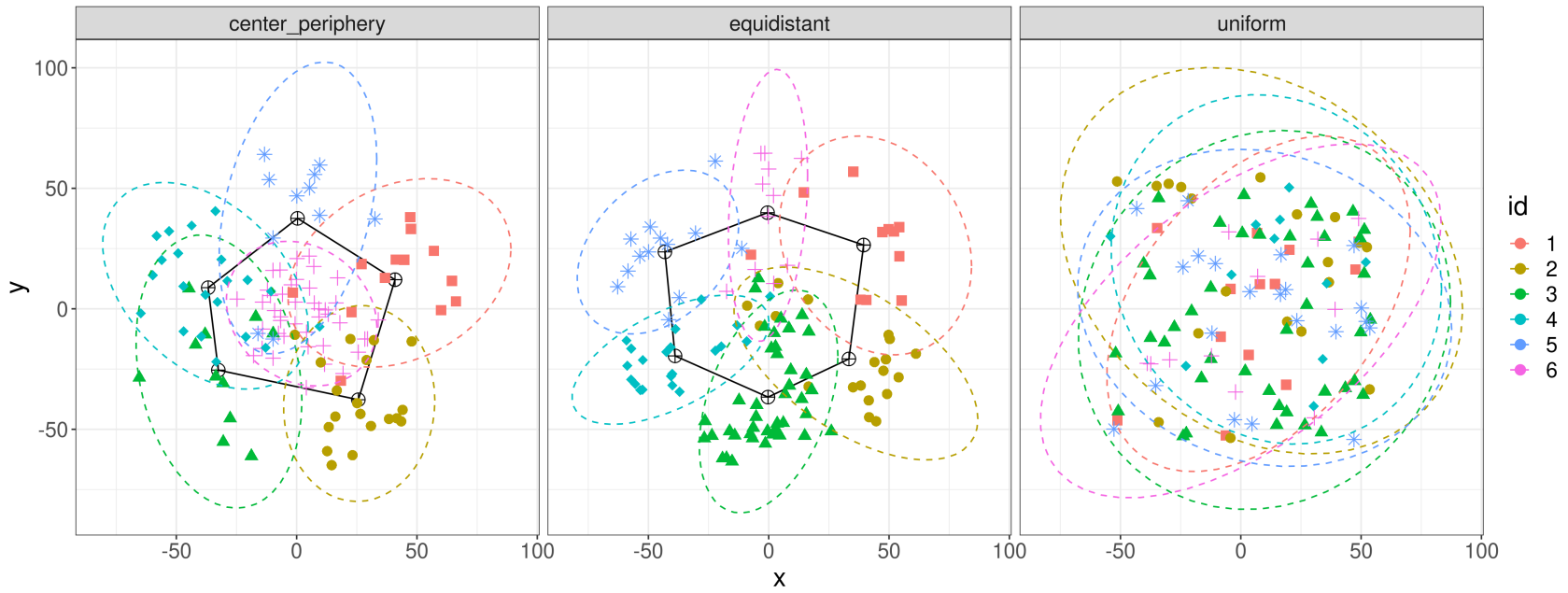


# Example of different number of categories ($n$)

# Example of different count configurations ($c$)



# Example of different types of spatial configurations




Equidistant (a), center-periphery (b), and uniform (c) distributions, for N_s = 117 settlements distributed in N_c = 6 categories, with a count configuration Q = (42, 21, 19, 13, 11, 11), and r = 26. In (a), the distributions' centers form a regular hexagon. In (b), the most populated category lies in the origin, while the other centers form a regular pentagon around it. In (c) all centers coincide at the origin.

# Outline of the talk

Introduction

Our approach

Simulated data

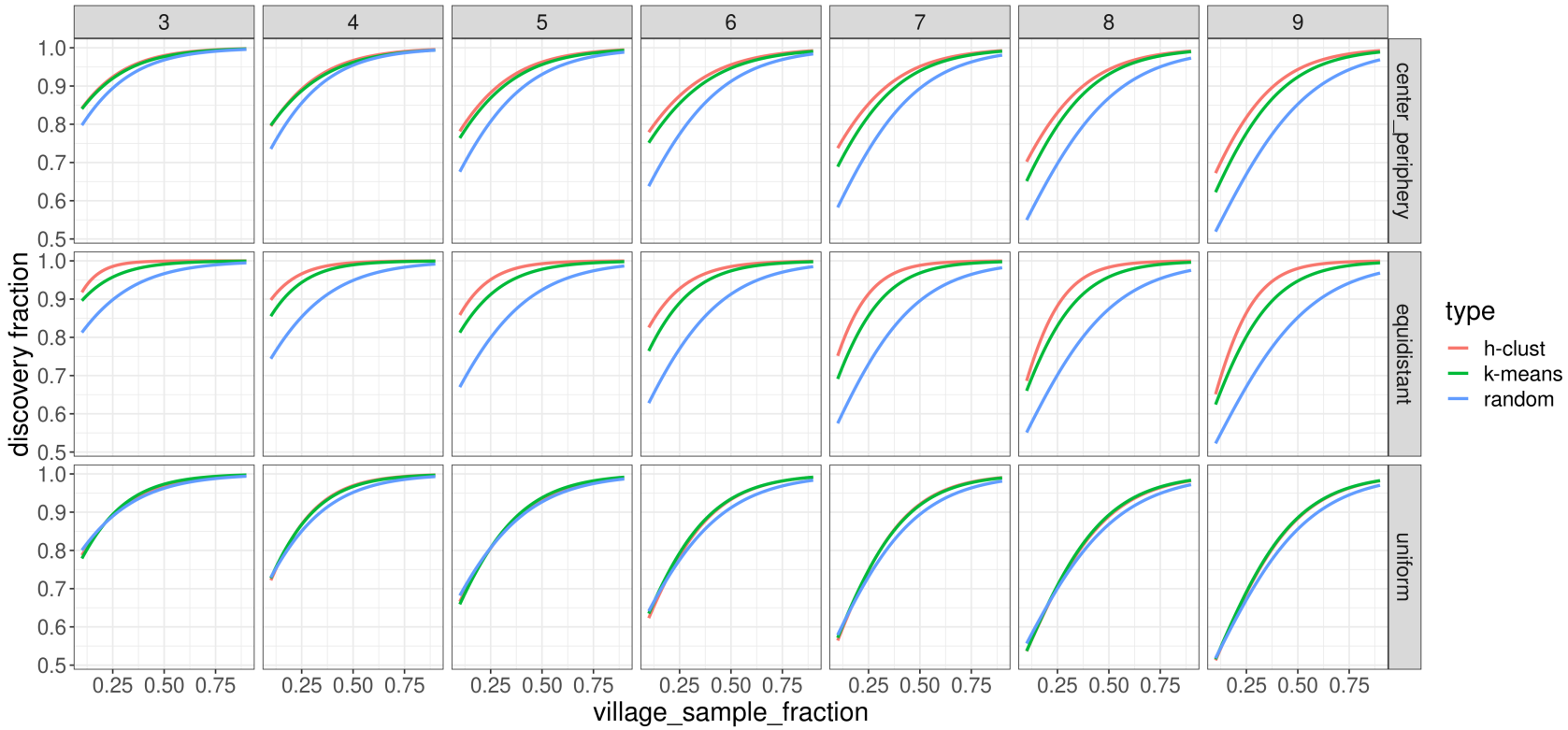
Results and modelling

Circassian data example

Entropy

Conclusions

# Results




Results for the discovery fraction as a function of the settlement sample fraction for differnt values of the parameters. We can see a clear improvement in the discovery fraction when using clustering algorithms for the data with spatial structure (first and second rows), and no improvement or depreciation in performance for the cases with no spatial structure (last row).

# Modelling the variation

- From the previous slides, we can see that
  - *k*-means and hierarchical clustering are signifficantly better than random sampling in non-uniform spatial relations;
  - *k*-means and hierarchical clustering are as good as random sampling with uniform spatial relations.

# Modelling the variation

- From the previous slides, we can see that
  - *k*-means and hierarchical clustering are signifficantly better than random sampling in non-uniform spatial relations;
  - *k*-means and hierarchical clustering are as good as random sampling with uniform spatial relations.

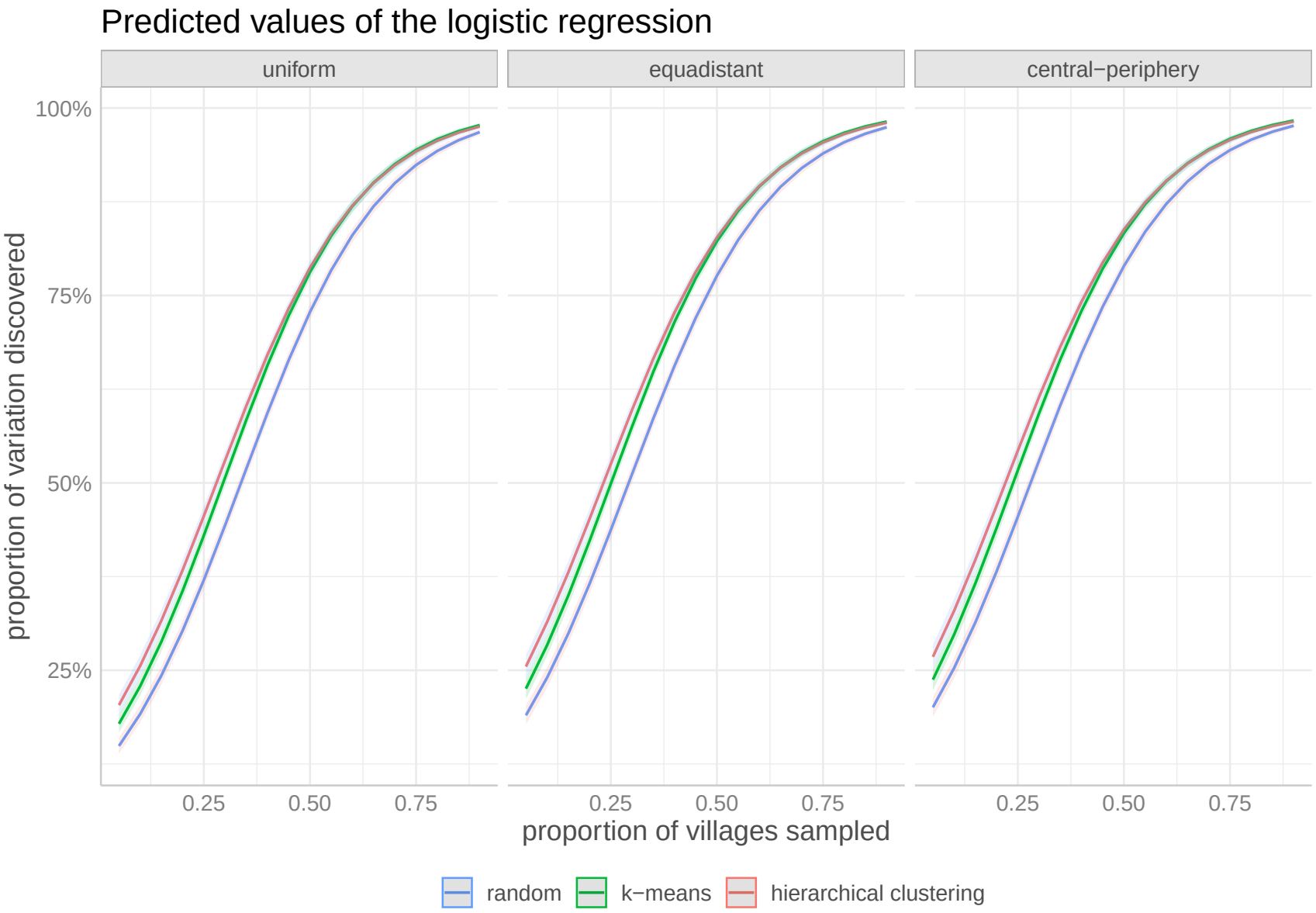
- We run a logistic regression in order to prove those observations by quantifying the relation between:
  - One **binary variable** (outcome):
    - All variation discovered vs. Not all variation discovered
  - Three parameters:
    - Proportion of villages sampled $p$ (numeric: $0.1, \ldots, 0.90$)
    - Type of clustering (hirarchical, *k*-means, random)
    - Type of geographic distribution (central-periphery, equidistant, uniform)

outcome ~ (spatial configuration + cluster type) * proportion_of_villages

# Regression results

| term | estimate | std.error | statistic | p.val |
|---|---|---|---|---|
| (Intercept) | -2.045 | 0.048 | -42.960 | *** |
| equadistant | 0.295 | 0.051 | 5.736 | *** |
| central-periphery | 0.362 | 0.051 | 7.036 | *** |
| k-means | 0.208 | 0.052 | 4.035 | *** |
| h. clustering | 0.384 | 0.051 | 7.510 | *** |
| proportion_of_village | 6.057 | 0.112 | 54.256 | *** |
| equadistant:proportion_of_village | -0.068 | 0.125 | -0.546 | 0.59 |
| central-periphery:proportion_of_village | -0.050 | 0.126 | -0.396 | 0.69 |
| k-means:proportion_of_village | 0.159 | 0.126 | 1.258 | 0.21 |
| h. clustering:proportion_of_village | -0.121 | 0.125 | -0.969 | 0.33 |

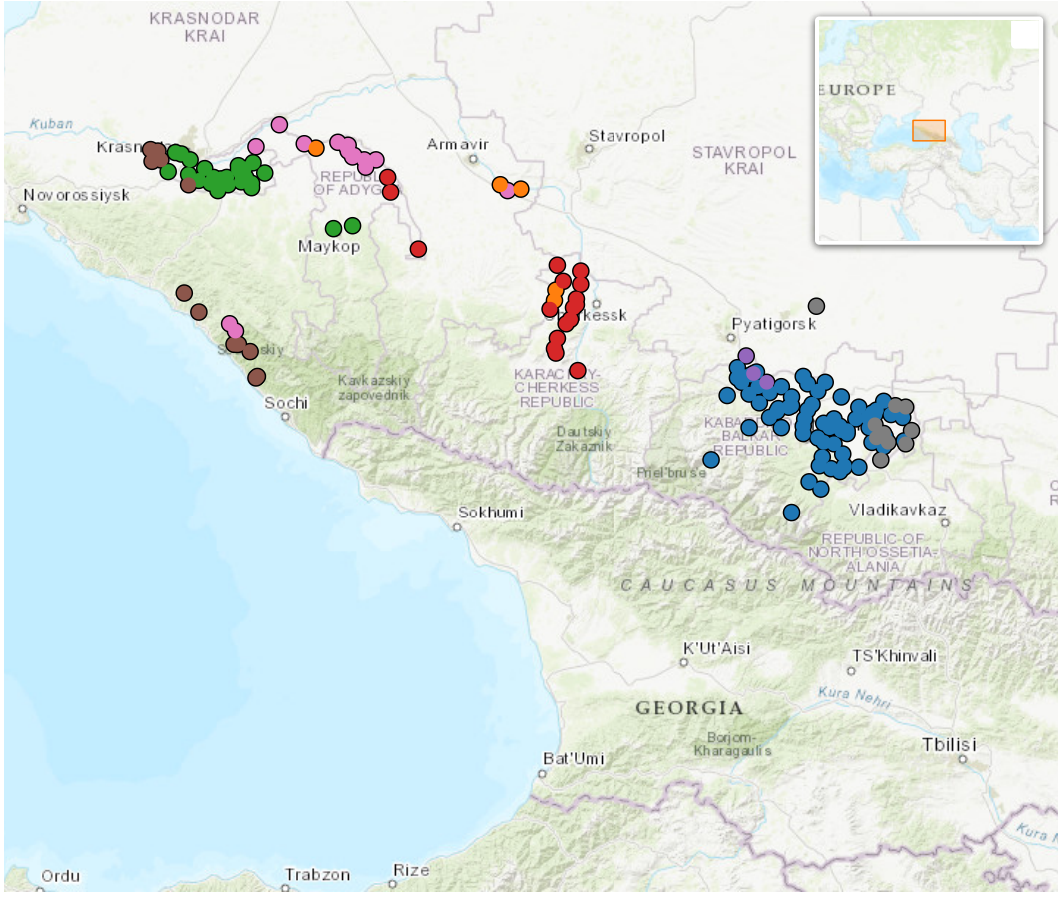
# Regression results



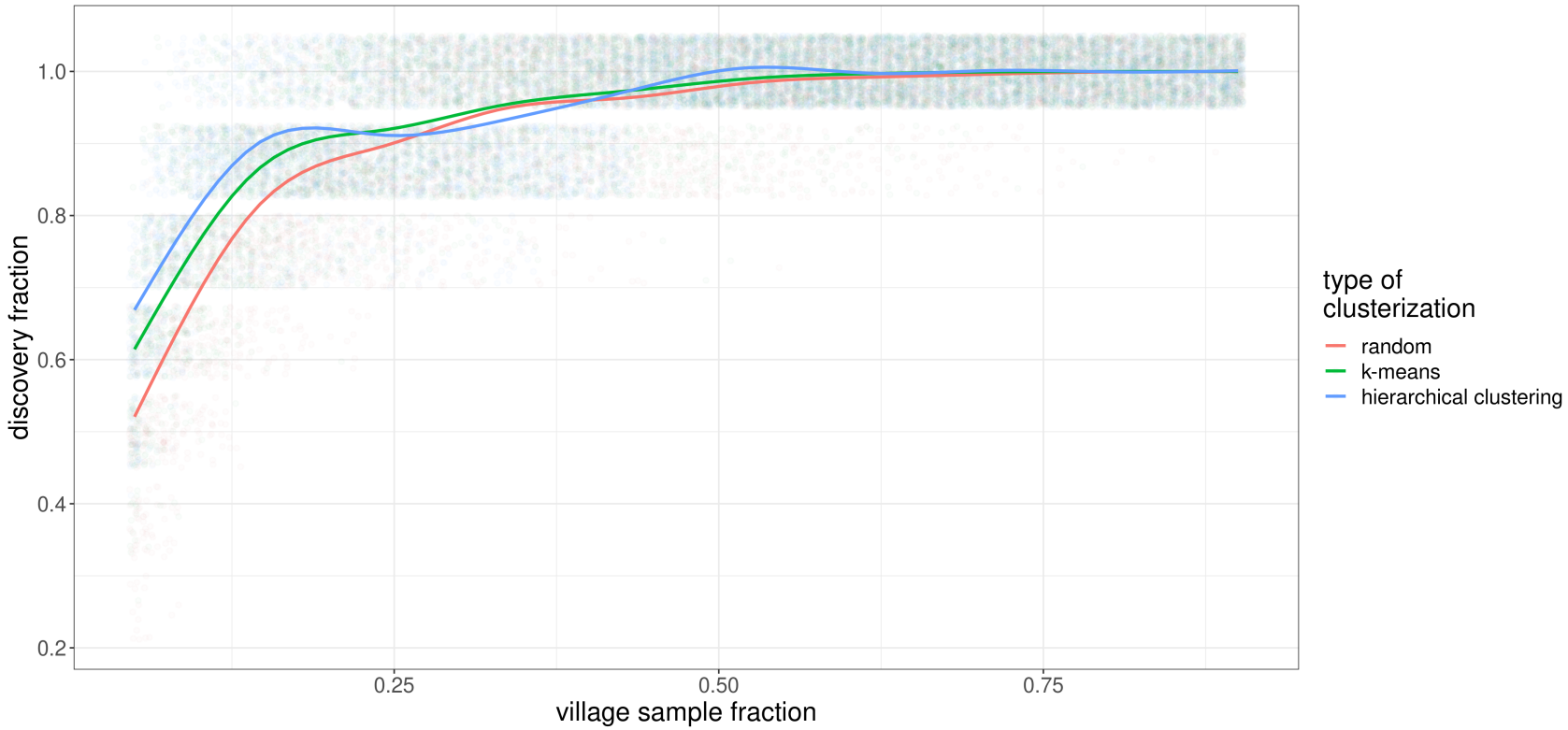
# Outline of the talk

# Algorithm evaluation using Circassian data [Moroz 2017]

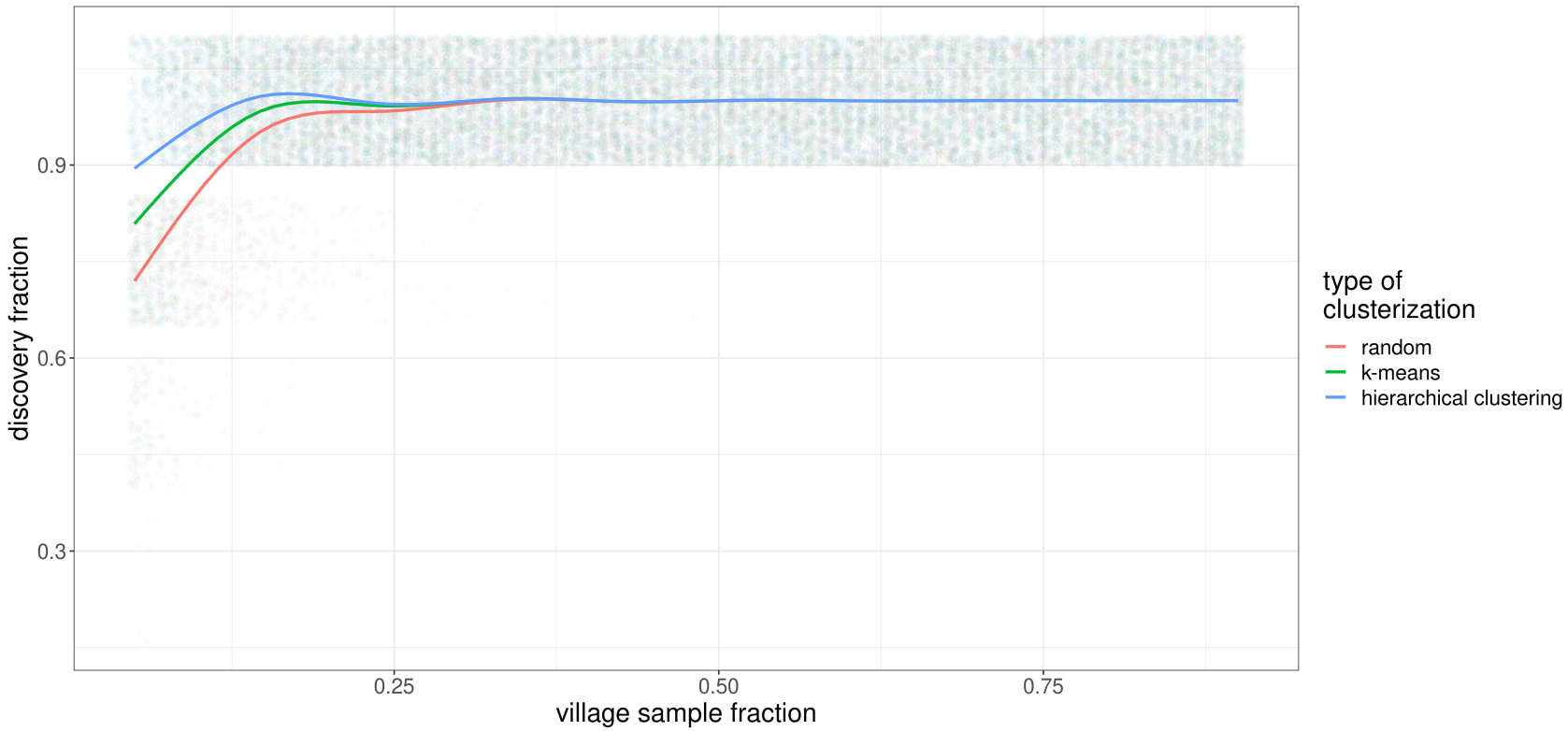
- 158 villages
- proportion of villages sampled: 0.05, 0.06, …, 0.89, 0.9
- true count configuration: 68-27-17-15-13-10-5-3
- 100 runs of each method on the same dataset



# Algorithm evaluation using Circassian dialects data [Moroz 2017]



# Algorithm evaluation using Circassian $q^h$ data



# Outline of the talk

Introduction

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

# Information entropy

In order to measure how the count configuration $c$ affects our sampling method, we use the information entropy, introduced in [Shannon 1948]:

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

# Information entropy

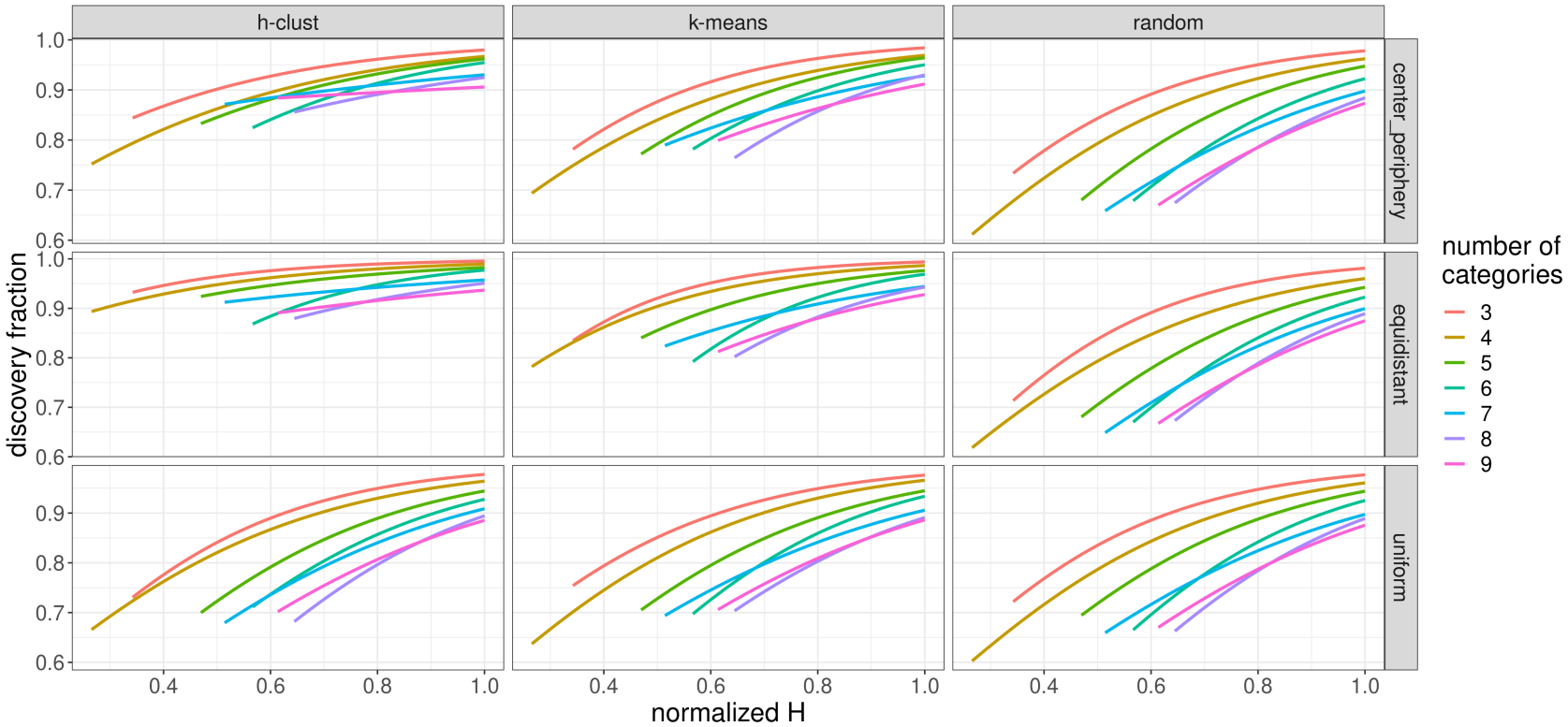
In order to measure how the count configuration $c$ affects our sampling method, we use the information entropy, introduced in [Shannon 1948]:

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

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

Our approach

Simulated data

Results and modelling

Circassian data example

Entropy

Conclusions

# Conclusions

- We introduced a geographic sampling algorithm for detecting variation in doculects, based on different clustering methods
- We tested our algorithm against random sampling in data simulated with different geographic distributions and numbers of observations
- We have found that our algorithm outperforms random sampling on simulated data in the cases where an underlying geographical structure is present, and performs as well as random sampling when variation is uniformly distributed across space
- We applied our algorithm to real data from Circassian languages
- We have found that our algorithm outperforms random sampling on real data for small sample proportions, but hirarchical clustering becomes worse than random sampling on larger sample proportions on those specific data
- We found that our algorithm has optimal results when entropy is lower

# Problems

- dialect fluctuation
- do we need statistics?
- is discovered fraction a good measure? We already showed that entropy matters...
- should we be interested in estimating discovered proportions?

# 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.

Moroz, G. (2017). *lingtypology: easy mapping for Linguistic Typology*.

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.