Journal of Linguistic Geography (2024), **12**, 24–31
doi:10.1017/jlg.2024.8

CAMBRIDGE UNIVERSITY PRESS

Article

Detecting linguistic variation with geographic sampling

Ezequiel Koile[1] and George Moroz[2]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany and [2]Linguistic Convergence Laboratory, HSE University, Moscow, Russia

Abstract

Geolectal variation is often present in settings where one language is spoken across a vast geographic area. This can be found in phonological, morphosyntactic, and lexical features. For practical reasons, it is not always possible to conduct fieldwork in every single location of interest in order to obtain the full pattern of variation, and a sample of them must be chosen. We propose and test a method for sampling these locations, with the goal of obtaining a distribution of typological features representative of the whole area. We apply *k*-means and hierarchical clustering algorithms for defining this sample, based on their geographic distribution. We test our methods against simulated data with several spatial configurations, and also against real data from Circassian dialects (Northwest Caucasian). Our results show an efficiency significantly higher than random sampling for detecting this variation, which makes our method profitable to fieldworkers when designing their research.

Keywords: Linguistic variation; geographic sampling; Circassian languages; computational methods; clustering

1. Introduction

Languages spoken across a vast geographic area tend to present dissimilarities in their local varieties, to a higher or a lower degree. The term *geolectal variation* stands for the different forms that a language takes across its geographically separated varieties (Labov, 1963). This variation can be present in different forms, such as phonological, morphosyntactic, and lexical features. This type of linguistic variation has been addressed by dialectologists since the mid-nineteenth century and was considered generally by linguists only in the second half of the twentieth century (Chambers and Trudgill, 2004). Traditionally, this phenomenon has been overlooked, with homogeneous distributions being assumed in most studies (Dorian, 2010).

In certain regions, it is possible to find a geographic dialect continuum formed by a group of settlements whose dialects differ slightly with distance. This generates a situation where every group of contiguous settlements speak very similar, intelligible varieties, but more distant settlements of the same continuum speak varieties that are non-intelligible (Chambers and Trudgill, 2004). This is the case of several groups of languages in Europe, such as the West Romance, East Slavic, or Scandinavian dialect continua, and it is particularly relevant in the Caucasus.

In order to account for the variation in a geographic region, it is necessary to collect data from a significant number of different locations. Although methods for collecting a large amount of linguisitc data in a short period of time exist (and they thrived during the COVID-19 pandemic [Leemann et al., 2020; Staff, 2019; Sprouse, 2011; Piller, Zhang, & Li, 2020]), in this kind of research, linguists need take into account geographical imbalance, quality, and representativity of the obtained data (e.g. in terms of recording conditions, possibility of speech disorders, and others). Therefore, this collection must be done in a systematic way, such as field expeditions where questionnaires are elicited and completed by experienced researchers with high quality recording devices. For practical reasons, it is not always possible to collect fieldwork data from every single location in order to obtain this full pattern of variation. Instead, it is more efficient to collect data from a selected group of these settlements, which we will call a sample. The question arises of which locations should be surveyed, either in order to resemble the real distribution of linguistic features, or at least to detect the degree of different values that exist for each of them.

1.1. Spatial sampling

Spatial sampling is a well developed field (see Stehman and Overton, 1995; Delmelle, 2009 and references therein), but no systematic study specific to linguistics has been carried out to date. According to Delmelle (2009), spatial sampling can be used in order to solve three categories of problems: estimating non-spatial characteristics of a spatial population, summarizing spatial variation of a variable in the form of a map, and obtaining observations independent from each other from a spatially structured dataset. All of them are relevant to us.

We focus on a two-dimensional discrete spatial sampling, where our goal is to sample a set n out of a population of N spatially structured datapoints. Delmelle (2009) divides such sampling methods into i) uniform random sampling, ii) systematic sampling (which can be further divided into regular, random, and unaligned), and iii) stratified sampling.

In the uniform random sampling, the dots are randomly selected, and the selection of a unit does not influence the selection

Corresponding author: Emails: ezequiel_koile@eva.mpg.de and agricolamz@gmail.com

Both authors contributed equally to the manuscript.

Cite this article: Koile E and Moroz G. (2024) Detecting linguistic variation with geographic sampling. *Journal of Linguistic Geography* **12**: 24–31, https://doi.org/10.1017/jlg.2024.8

of any other. The advantages of this method are its operational simplicity and its capacity to generate a wide variety of distances among pairs of points in the sample. Its main disadvantage is that the distribution of selected points may not be representative of the underlying population, resulting in some areas being oversampled and others undersampled.

In the systematic sampling, the space of interest is divided into a uniform grid of subregions or cells, and one sample is taken from each of these. It is called regular systematic if the sample is always taken from the same location inside the cell (e.g. its center), systematic unaligned if there is a differentiate algorithm for choosing the sample inside each cell, or systematic random if the sample inside each cell is taken at random. The benefits of this approach are a good spreading of observations across the whole region of interest, resulting in a representative sampling coverage, and avoiding sampling clustering and redundancy. Its inconveniences are that the distribution of distances between points in the sample is biased (many points are separated by the same distance, i.e. multiples of the cell size), and there is the possibility of missing spatially periodic behaviors.

Finally, in stratified sampling, the region of interest is partitioned into non-overlapping strata, making the systematic sampling a special case of stratified sampling. For each stratum, one or more samples are collected. The main challenge here is deciding the shape and size of each stratum. Also, a criterion is necessary for deciding how many dots will be sampled in each stratum (e.g. an amount proportional to the size of the stratum or only one for each stratum) and how this will be selected (e.g. always in the center of the stratum, or at random).

One issue to take into account is that of spatial auto-correlation. In the case of spatially structured data, we expect stronger similarities among closely spaced datapoints, and it is therefore redundant to oversample in those areas. Spatial auto-correlation, defined as how similar the values of the variable of interest are at different locations as a function of their separating distance, generally decreases as the distance among sample points increases, and the functional form of this decrease makes one or other sampling methods more efficient, that is, to have a lower variance.

It has been argued that linear distances between point locations are not the best way of accounting for the possibility of linguistic contact, but an inverse-square law might be more useful (Nikolaev, 2019). Although we think this is a relevant observation for larger distances (such as the macro-area of Eurasia, studied in the mentioned reference), linear distances are a good enough measurement for our region of interest, the Caucasus.

It also should be mentioned that, in our paper, we used fake latitude and longitude values for simulations and real latitude and longitude values for Circassian data. Using those values as a simple two-dimensional space is equivalent to projecting our observation in the Mercator projection, known for causing a huge distortion in sizes. This is not a big problem for the small area that we analyze, but it could be so for other regions of interest. Therefore, we propose applying our methods onto projected coordinate systems with a projection suitable for the area under analysis.

1.2. The aim of this study

In the present paper, we propose and test a method for sampling different locations, with the goal of obtaining the distribution of linguistic features most representative of the whole area. For this goal, we use different clustering algorithms, such as *k*-means and hierarchical clustering of the locations. Using spatial clustering is not particularly new, since it can be found in the list of methods from Delmelle (2009). The novelty in our approach is that we test this algorithm using four spatial configurations for the locations involved, which we call *circular equidistant*, *center–periphery*, *dialect chain*, and *uniform*. We test our methods against both simulated data with a different number of categories, with all possible distributions of our linguistic feature of interest, and on various spatial configurations, and also against real data from Circassian dialects (Northwest Caucasian). This assures us a geographically representative sample, less affected by spatial auto-correlation than a random selection. Our results show an efficiency higher than random sampling, both for detecting variation and for estimating its magnitude, which makes our method profitable to fieldworkers when designing their research.

In section 2 we describe the data used, both in our simulation and in our example of real data on Circassian languages. In section 3 we describe the clustering methods used. In section 4 we describe and discuss the results obtained. In section 5 we summarize our conclusions. We used the packages lingtypology (Moroz, 2017), partitions (Hankin, 2006), stats (R Core Team, 2021), and tidyverse (Wickham et al. 2019) in the programming language R. All code and data are available on GitHub.[1]

2. Data

In this paper we focus on the scenario where a researcher is interested in one or more linguistic features with discrete values. This can represent the realization of a phonological feature, a morphosyntactic feature, the attestation of a lexical item, or an overall feature of the variety, such as the dialect spoken in a region. However, our method can be extended to numeric variables such as number of cases or vowel length. We give one single value to the feature of interest in each settlement. This is a simplification of the real situation, where inter-speaker and intra-speaker variation can be present in the same settlement (see Dorian, 2010; Moroz and Verhees, 2019).

2.1. Generation of simulated data

We generated data that resembles different distributions found in realistic linguistic settings. In all cases, we generated a number of settlements N_s, ranging from 30 to 90 in decades, and a number N_c of different categories for the feature of interest, ranging from 3 to 9. For each combination of N_s and N_c, each category can be differently populated. For example, if we have $N_s = 50$ and $N_c = 5$, an even distribution will group exactly 10 settlements in each category (configuration 10–10–10–10–10), while an extremely skewed distribution will have one overly populated category with 46 settlements, and the remaining four categories with only one settlement each (configuration 46–1–1–1–1). We call this distribution the *count configuration* Q, and its associated entropy is $H(Q)$ (Shannon, 1948).

We distinguish four spatial configurations, that we call *circular equidistant*, *center–periphery*, *dialect chain*, and *uniform*, described below. In all configurations, the locations of settlements labeled with each category are generated as a bimodal normal or log-normal distribution around a center, as explained below. The different configurations refer to how these centers are located in space, as well as the parameters of the normal distribution. More complex configurations that have been defined in the bibliography (see e.g. Goossens, 1977:78), such as enclave, funnel, tubular, or ring, are not considered *a priori*, although some enclave-like configurations can emerge from the overlapping regions of our

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press

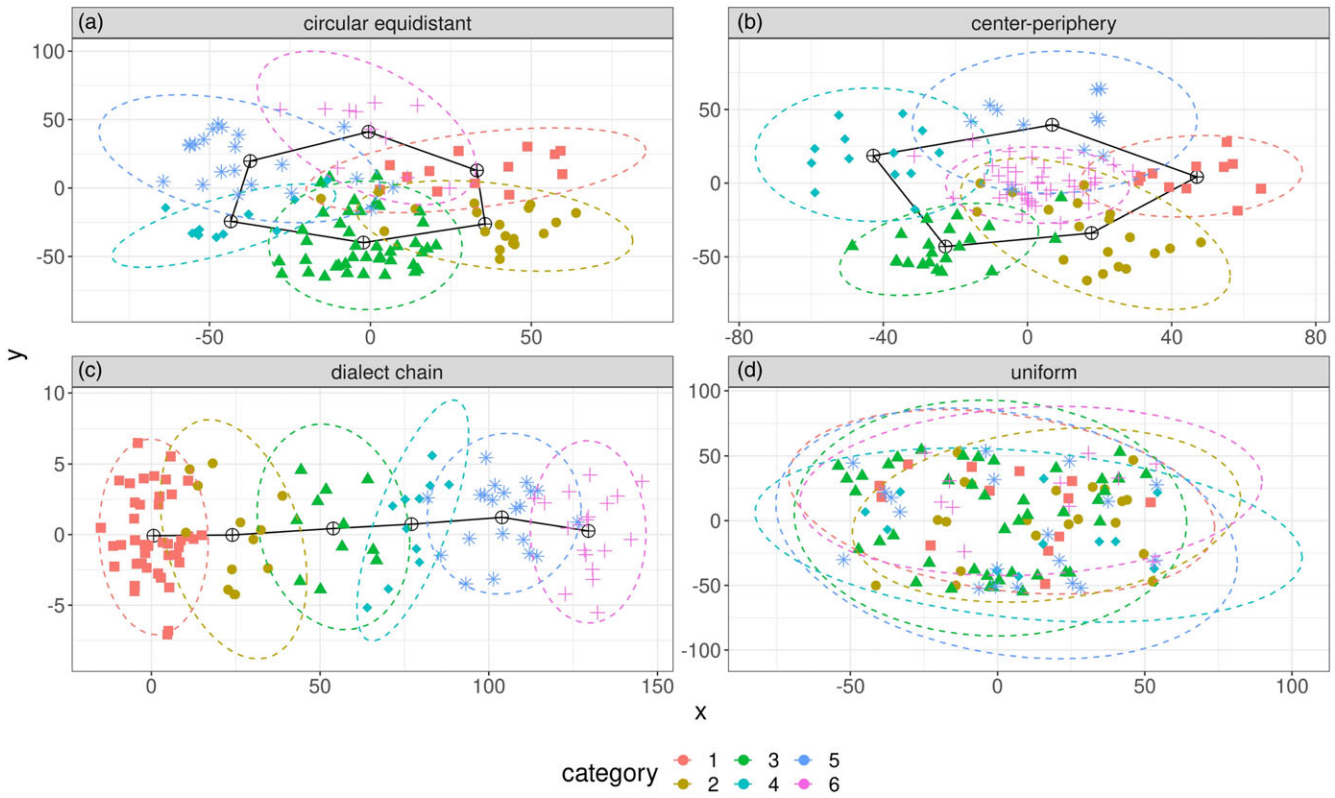


Figure 1. (a) Circular equidistant, (b) center–periphery, (c) dialect chain, and (d) uniform distributions for $N_s = 117$ settlements distributed in $N_c = 6$ categories, with a count configuration $Q = (42,21,19,13,11,11)$ and $r = 26$. In (a), the distributions' centers form a regular hexagon. In (b), the most populated category lies in the origin, while the other centers form a regular pentagon around it. In (c), they form a straight line. In (d), all centers coincide at the origin. The dots within each category are surrounded with their normal ellipses (Fox and Weisberg, 2018). The centers around which the data are generated in (a) and (b) form regular polygons, while the polygons showed here link the centroids of the generated data *a posteriori* and are therefore not regular. Similarly, the dialect chain in (c) does not form a horizontal segment, although the centers for data generation did.

simulated data, which might be associated with the mixed zones (*Mischgebiet*) in the mentioned reference.[2]

Circular equidistant. In this configuration, the centers of each category correspond to the vertices of a regular polygon of N_c sides, circumscribing a circumference of radius r, centered in the origin. Therefore, all centers are at distance r from the origin, and each category has two closest neighbors to its sides at a distance equal to the polygon's side, $d_t = 2r \cdot \sin\left(\frac{\pi}{N_c}\right)$. In the tangential direction, settlements belonging to a given category follow a normal distribution centered in the corresponding vertex of the polygon and with a standard deviation equal to half of the distance between closest neighbors $\sigma_t = d_t/2$, allowing for some overlap. In the radial direction, we allow settlements to approach the origin, but not to expand too much outwards. For this reason, we choose a log-normal distribution bound by $2r$, and decaying towards the origin, with standard deviation $\sigma_r = r$. Figure 1(a) shows an example of this distribution for $N_s = 117$, $N_c = 6$, and $Q = (42,21,19,13,11,11)$.

Center–Periphery. This configuration is built in a fashion similar to the previous one, with one main difference: The most populated settlement is located at the origin, and all others form an $(N_c - 1)$-side regular polygon around it. All of these peripheral settlements lie at the same distance r from the central one and have two closest neighbors in addition to this one, at a distance $d_r = 2r \cdot \sin\left(\frac{\pi}{N_c - 1}\right)$. Again, in the tangential direction, settlements belonging to a given category follow a normal distribution centered in the corresponding vertex of the polygon, with a standard deviation equal to half of the distance between closest neighbors $\sigma_t = d_t/2$; and in the radial direction they form a log-normal distribution bound by $2r$, decaying towards the origin, with standard deviation $\sigma_r = r$. Figure 1(b) shows an example of this distribution for $N_s = 117$, $N_c = 6$, and $Q = (42,21,19,13,11,11)$.

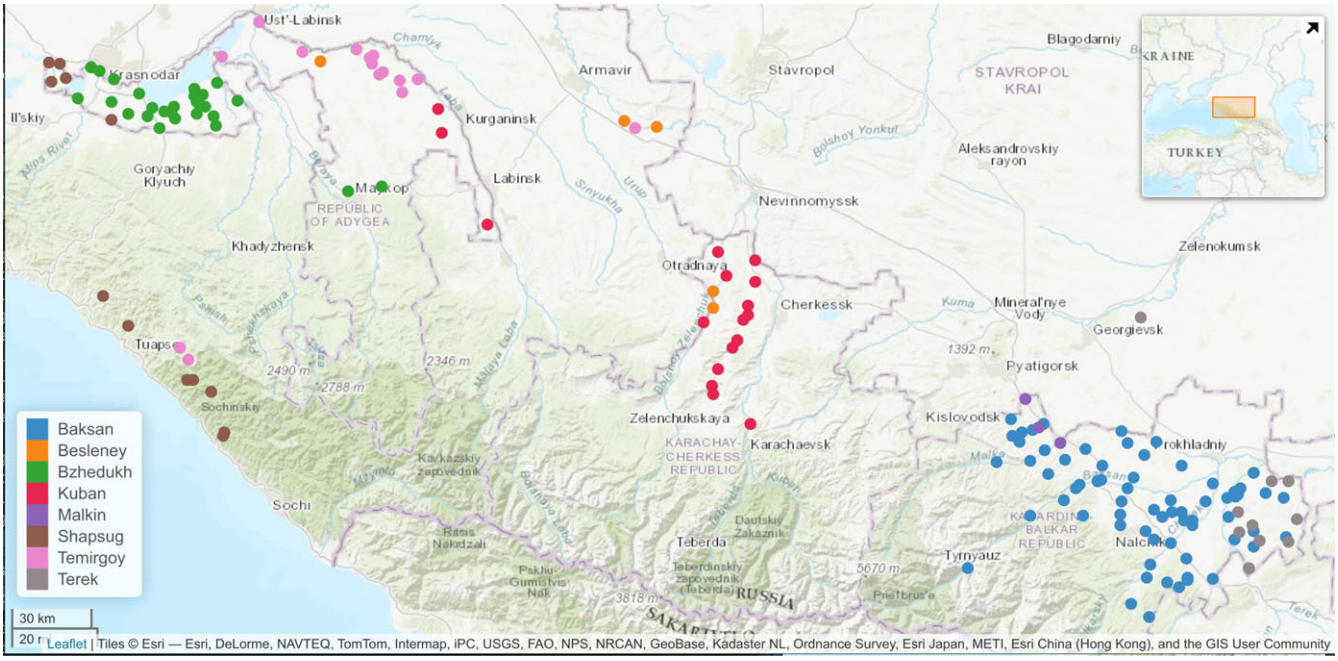
Dialect chain. This configuration is similar to the circular equidistant one, but along a straight line. Here, the N_c centers of each category lie on a straight line, each separated by distance r from its closest neighbors. Therefore, the distance between two contiguous centers is $d = r$, while distance between the two most extreme centers is $d_M = (N_c - 1) \cdot r$. The settlements belonging to a given category follow a normal distribution centered in the corresponding point in the straight line, with standard deviation $\sigma = r$ in both the direction along the line and the perpendicular one, resulting in circular distributions with some overlap between neighbors. Figure 1(c) shows an example of this distribution for $N_s = 117$, $N_c = 6$, and $Q = (42,21,19,13,11,11)$.

Uniform. In this configuration, the centers of all N_c categories coincide at the origin, and settlements around them form a uniform squared distribution, with standard deviation r. Figure 1(d) shows an example of this distribution for $N_s = 117$, $N_c = 6$, and $Q = (42,21,19,13,11,11)$.

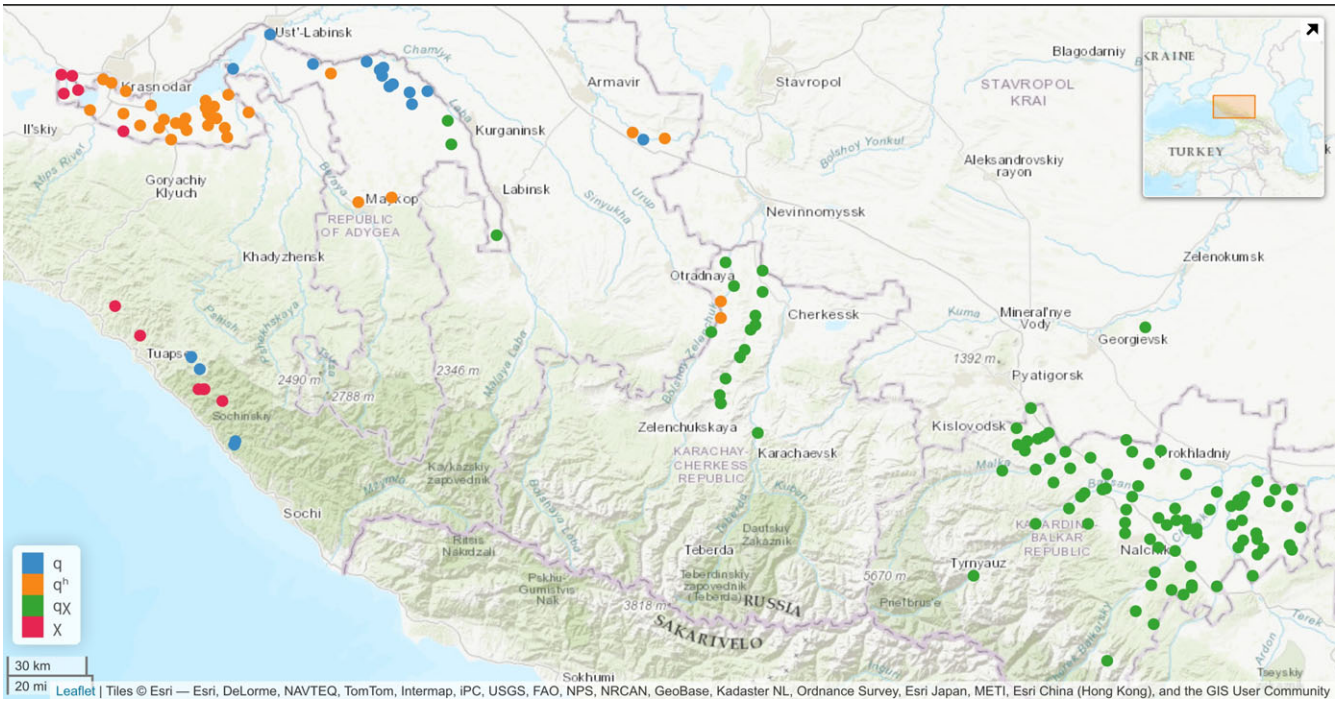
2.2. Real data: Circassian languages

As a real example application of this procedure, we use data for Circassian languages (Northwest Caucasian), from Moroz (2017). This dataset contains different linguistic data for 158 Circassian settlements in the Russian Caucasus. We consider two different

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press



Map 1. Distribution of dialects in the Circassian settlements.



Map 2. The distribution of Circassian reflex of *q^h.

features for these data. First, we study the dialect spoken in each settlement. There are eight dialects, each spoken in the following number of settlements: Baksan (68), Bzhedukh (27), Kuban (17), Temirgoy (15), Shapsug (13), Terek (10), Besleney (5), and Malkin (3). These dialects show a distribution with a clear geographical pattern, as can be seen in Map 1. The second feature we study is the reflex of *q^h (Moroz, 2021), which has different values in different settlements: q^h in Neshukay (Bzhedukh) and Besleney (Besleney), q in Bolshoy Kichmay (Shapsug) and Pshicho (Temirgoy), χ in Pseytuk (Shapsug) and Khadzhiko (Shapsug), and $q\chi$ in Zhako (Kuban) and Khodz' (Kuban). This feature has four categories, and its distribution does not show an obvious geographical pattern, as can be seen in Map 2.

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press

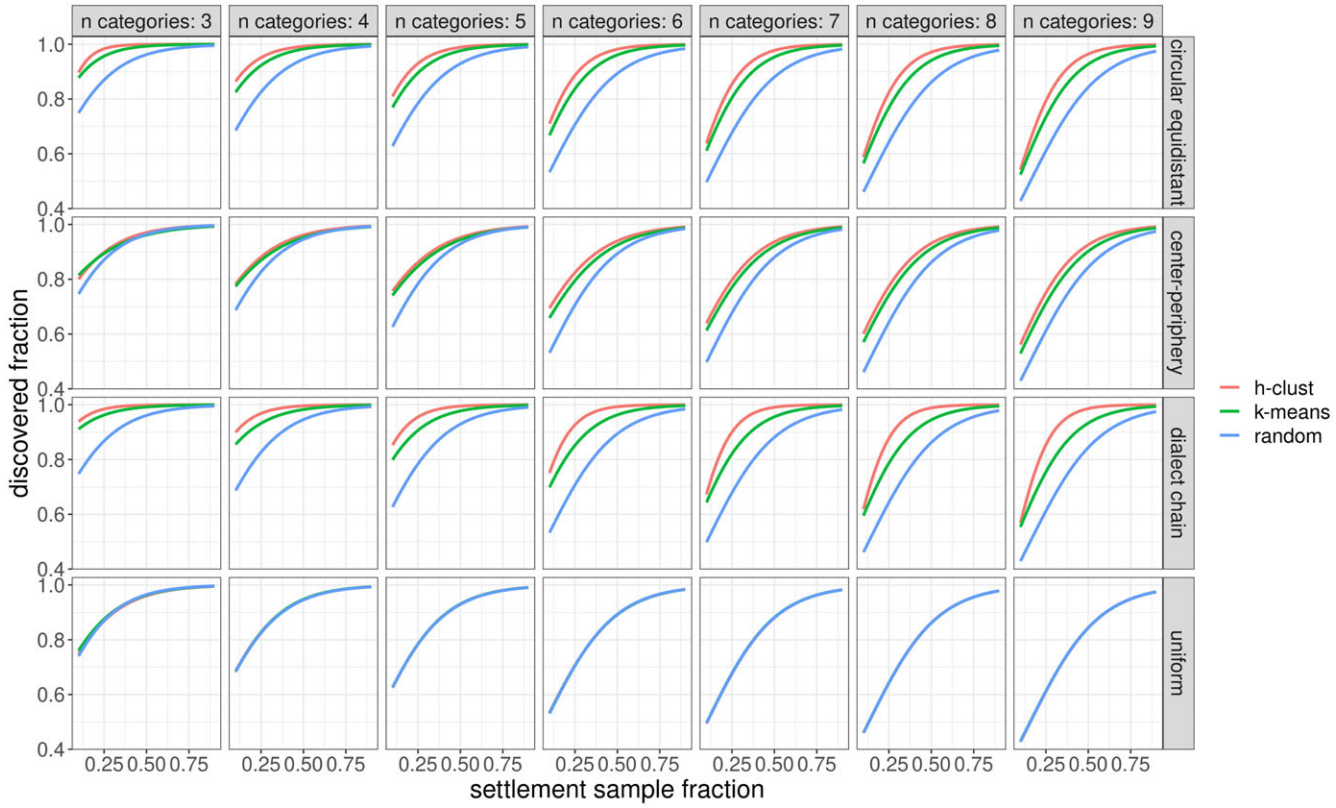


Figure 2. Results for the discovery fraction as a function of the settlement sample fraction for different values of the parameters. We can see a clear improvement in the discovery fraction when using clustering algorithms for the data with spatial structure (rows 1–3) and no improvement or depreciation in performance for the cases with no spatial structure (last row).

3. Clustering methods

For each case study, we examined how much of the real variation in the data we would be able to discover by using three different sampling strategies. Remember that we had N_s settlements with a given geographical configuration, and our feature of interest was distributed across N_c categories, according to a count configuration Q. We wanted to characterize the amount n_c of categories discovered after sampling n_s settlements with different criteria. The discovery fraction, $df = \frac{n_c}{N_c}$, was used as a score for comparing sampling methods. All methods described below (except for random sampling) were our clustering methods based on geographical distance.

Random sampling. The most trivial form of sampling is random: If we assume that there is no association between the geographic location of a settlement and the value for the linguistic feature of interest (or if we have no information about the geography whatsoever), the most simple method is to sample at random. We randomly drew n_s out of the N_s settlements and counted the number n_c of categories discovered in the sampling.

***k*-means clustering.** This method (Lloyd, 1982) uses a pre-specified number of clusters. It starts with k randomly chosen clusters and iterates searching for the mutually exclusive set of clusters of spherical shape based on a similarity measure. In our case, we used geographic distance as the parameter to build the clusters of settlements, and we built $k = n_s$ clusters. After this, we randomly chose one settlement inside each cluster, forming in this way our sample of n_s settlements. The different categories in this sample were our result n_c.

Hierarchical clustering. This method (Ward, 1963) ranks the elements of a set according to their similarity (we used Euclidean distance), and builds a dendrogram (tree-like plot). Then, it clusters the different subgroups according to their grouping in the dendrogram. The total number of clusters can be either selected *a priori* or data driven. We used the former option here. We again used geographic distances to cluster settlements, and set the number of clusters to n_s. As in the previous case, we randomly chose one settlement as a representative of each cluster, forming our sample of n_s settlements. The different categories in this sample were our result n_c.

4. Results and discussion

4.1. Simulated data

Figure 2 shows the results of applying the three different clustering algorithms to our simulated data. The plots show the discovery fraction $df = \frac{n_c}{N_c}$ as a function of the proportion of settlements sampled $\frac{n_s}{N_s}$ for different values of total number of categories and spatial configuration. Each column of plots shows the results for a different value of the total number of categories N_c from 3 to 9, while each row of plots shows a different spatial configuration: circular equidistant, center–periphery, dialect chain, and uniform. Each group was approximated with a logistic regression line.

From the rows 1–3 in Figure 2 we can see that in the cases with spatial structure (configurations circular equidistant, center–periphery, and dialect chain) there was a clear improvement of the clustering methods over random sampling, the hierarchical clustering consistently performing slightly better than the *k*-means

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press

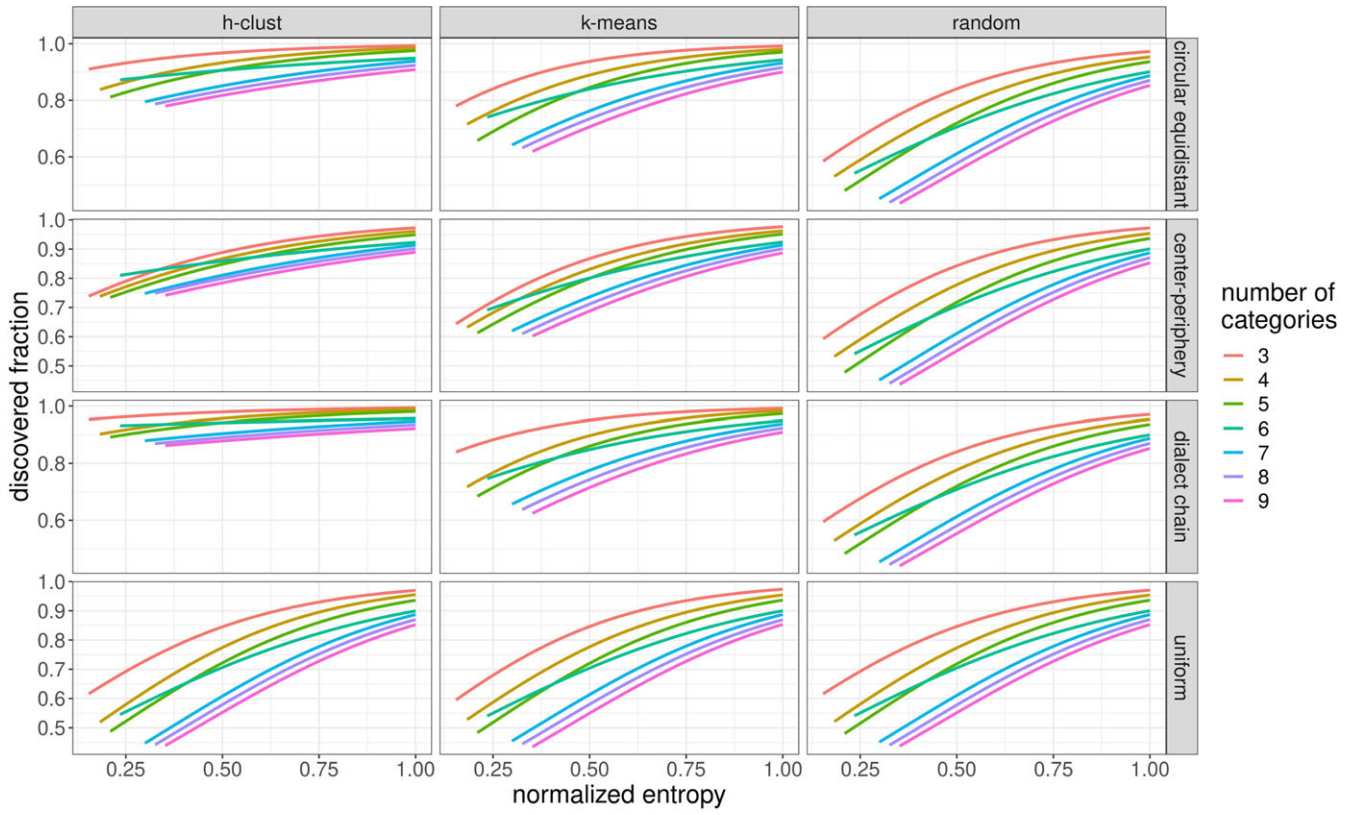


Figure 3. Results for the discovery fraction as a function of entropy. We can see a clear improvement in the discovery fraction when there is a higher value of entropy, although the obtained results depend on the number of categories.

clustering. This improvement in the discovery rate was more evident for the cases with the higher quantity of categories N_c, and the effect was different for different spatial configurations: it was the most prominent for the dialect chain configuration, followed by equidistant, and finally center–periphery.

From the last row, we see that in the cases of data with no spatial structure (uniform), the performance of all three sampling methods was practically the same: there was no improvement (and also no depreciation) in the performance when choosing a spatial clustering algorithm over a random sampling.

We now want to study how the count configuration Q affected our results. Shannon entropy (Shannon, 1948), defined as

$$H(Q) = -\sum_{i=1}^{n} P(q_i) \cdot log_2 P(q_i) \quad (1)$$

gives us an idea of how evenly distributed counts in the different categories are, for a count configuration of the form $Q = (q_1, \ldots, q_n)$, where $P(q_i)$ is the probability of finding the i-th value of Q in each settlement. If the values are evenly distributed across all categories, the value of entropy will be higher, while it will be lower if the distribution is skewed. Going back to a previous example of $N_s = 50$ settlements and $N_c = 5$ categories, the most even count configuration, with 10 settlements in each category, $Q_{\text{even}} = (10,10,10,10,10)$, has an entropy $H(Q_{\text{even}}) = 3.32$, while the most uneven count configuration with all categories populated, $Q_{\text{uneven}} = (46,1,1,1,1)$, has an entropy $H(Q_{\text{uneven}}) = 0.56$.[3]

In Figure 3, we plotted the discovery fraction as a function of normalized Shannon entropy $h = H/H_{max}$ of the count configuration Q, for different values of the spatial configuration, clustering method, and total number of categories N_c. Plots were aggregated in the settlement fraction sampled. In this figure, we can see several behaviors. First, for each spatial configuration and clustering method, the discovery fraction increased with entropy. This is what we expected, since the more evenly the categories were distributed in settlements, the easier it was to discover more categories. We can also see how, when we increased the number of categories N_c, the overall curve decreased and moved to the right. This means, on the one hand, that it is harder to discover a large fraction of categories if there are more of them, and that the higher N_c, the larger configurations space is, and therefore the higher the possible values of entropy.

4.2. Real data: Circassian languages

In Figures 4 and 5, we can see the results for the discovery rate as a function of the settlement sample fraction, for the Circassian dialect recognition, and reflex of Circassian *q^{h}. Shadowed dots are the jittered values for each individual observation. We can see how hierarchical clustering had the highest performance, followed by the k-means clustering, while the random sampling performed the worst for both cases. This is consistent with the behavior observed for the simulated data discussed above.

4.3. Overall results

As we can see from Figures 2–5, our algorithm worked better than random sampling in all cases. It is worth mentioning the obvious fact visible through our analysis that the number of categories influenced the discovery rate. If a researcher sampled around

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press

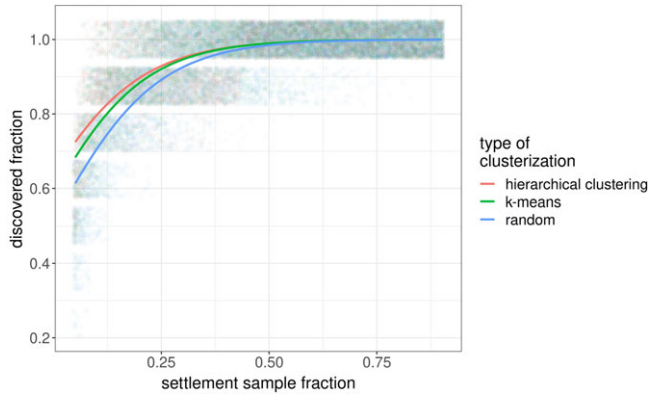


Figure 4. Discovery rate as a function of settlement sample fraction for Circassian dialect data (eight categories).

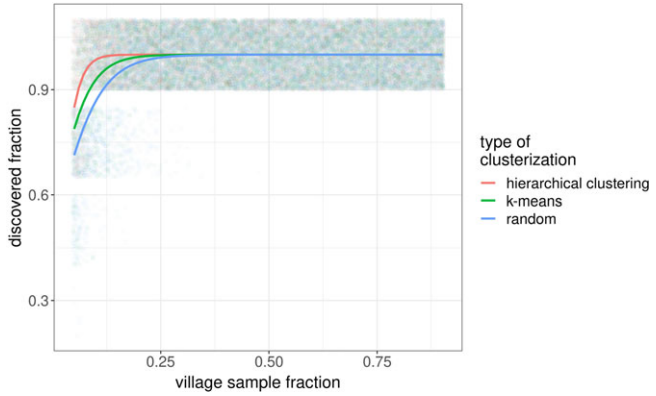


Figure 5. Discovery rate as a function of settlement sample fraction for Circassian reflex of *q^h (four categories).

90 percent of the data, they will discover nearly all of the present variation. However, if the sample is smaller, than the value where both approaches (clustering and random sampling) converge and all variation is found depends on the number of categories N_c. For example, from Figure 2 we can see that if our variable had three categories, then both approaches converged near 50 percent, but if our variable had nine categories, then the approaches would not converge until almost all settlements were sampled. We can also see this behavior for the case of the Circassian languages: The stabilization of the results (100 percent discovery rate for all methods) was reached at a lower sampling factor when the reflex of *q^h was studied (Figure 5, four categories) than when the dialect was investigated (Figure 4, eight categories). Another fact easily visible through our analysis is that it was easier to discover variables with evenly distributed frequencies. From Figure 3, we can see that the higher discovery rate corresponded with higher values of entropy. We can also see that the increase in the number of categories made it harder to discover all possible values; however it was easier to discover nine evenly distributed categories than three hardly skewed ones. The consequence of this could also be turned around: if the researcher discovered a given number of categories in a large amount of settlements, there is always a possibility that the variable under investigation is skewed and that there are really more rare values of this variable to discover. This uncertainty cannot be avoided by any sampling method, but it can be estimated by generating simulated data, as in this study.

5. Conclusion

This paper proposes a solution to the problem of sampling settlements in the limited conditions where it is not possible to collect fieldwork questionnaires from every single location. We performed two studies: First, we simulated data with different spatial and count configurations, as well as different number of categories, and compared the results from our algorithm with the random sampling. For all spatial configurations, our algorithm discovers more variation than random sampling. Second, we tested our algorithm with two different linguistic features from Circassian languages. For both the results are the same: hierarchical clustering performed the best, *k*-means slightly worse, and random sampling the worst. In a case with no spatial structure (uniform configuration), our algorithm performs exactly the same as random sampling. We also showed that not all variables are equal: It is harder to discover the variation of those with lower entropy, as expected. Our recommendation to the field linguist is to follow this algorithm: Estimate the number of settlements n_s that they can visit, then perform clustering of all settlements, and finally, randomly sample one location from the n_s clusters obtained from the clustering process. During the work on this paper, we wrote an R code for creating different spatial relationships between datapoints (circular equidistant, center–periphery, dialect chain, and random) that could be useful for other projects where spatial relations are important.

Acknowledgments. Both authors contributed equally. We would like to thank the members of the Linguistic Convergence Laboratory (HSE University) for useful comments in the early stage of this paper. Ezequiel Koile's work was funded by the Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology. George Moroz has received funding from the Basic Research Program at the HSE University.

Competing interests. The authors declare none.

Notes

1 See https://lingconlab.github.io/Detecting_linguistic_variation_with_geographic_sampling/.

2 Strictly speaking, our mixture zones are regions where *villages* with two or more values of the feature of interest coexist, but the value is unique for each village, while the mixed zone in the citation refers to a region where a mixed variety is spoken.

3 The extreme configuration $Q_{\text{extreme}} = (50,0,0,0,0)$ has $H(Q_{\text{extreme}}) = 0$.

References

Chambers, J. K. & P. Trudgill. 2004. *Dialectology*, 2nd edn. Cambridge: Cambridge University Press.

Delmelle, E. 2009. Spatial sampling. In A. Stewart Fotheringham and Peter A. Rogerson (eds.), *The SAGE handbook of spatial analysis*, 165–186. London: SAGE Publications.

Dorian, N. C. 2010. *Investigating variation: The effects of social organization and social setting*. Oxford: Oxford University Press.

Fox, J. & S. Weisberg. 2018. *An R companion to applied regression*. London: SAGE Publications.

Goossens, J. 1977. *Deutsche Dialektologie*, vol. *2205*. Berlin: Walter de Gruyter.

Hankin, R. K. S. 2006. Additive integer partitions in R. *Journal of Statistical Software, Code Snippets* 16. doi: 10.18637/jss.v016.c01.

Labov, W. 1963. The social motivation of a sound change. *Word 19*(3). 273–309.

Leemann, A., P. Jeszenszky, C. Steiner, M. Studerus, & J. Messerli. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, *6*(s3). doi: 10.1515/lingvan-2020-0061.

Lloyd, S. P. 1982. Least square quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2). 129–137.

Moroz, G. 2017. *lingtypology: easy mapping for Linguistic Typology*. Retrieved from https://CRAN.R-project.org/package=lingtypology (March 28, 2024).

Moroz, G. 2021. *Some questions of Circassian segmental and suprasegmental phonology and phonetics*. Ph.D dissertation. Moscow: National Research University Higher School of Economics.

Moroz, G. & S. Verhees. 2019. Variability in noun classes assignment in Zilo Andi: Experimental data. *Iran and the Caucasus 23*(3). 268–282.

Nikolaev, D. 2019. Areal dependency of consonant inventories. *Language Dynamics and Change 9*(1). 104–126.

Piller, I., J. Zhang, & J. Li. 2020. Linguistic diversity in a time of crisis: Language challenges of the covid-19 pandemic. *Multilingua 39*(5). 503–515.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal 27*(3). 379–423.

Sprouse, J. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods 43*(1). 155–167.

Staff, P. O. 2019. Correction: Talk2Me: Automated linguistic data collection for personal assessment. *PLoS ONE 14*(4). e0216375. doi: 10.1371/journal.pone.0216375.

Stehman, S. V. & W. S. Overton. 1995. Spatial sampling. In S. Arlinghaus (ed.), *Practical handbook of spatial statistics*, 31–63. Boca Raton: CRC Press.

Ward, J. H. Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301). 236–244.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, & H. Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software 4*(43). 1686. doi: 10.21105/joss.01686.

https://doi.org/10.1017/jlg.2024.8 Published online by Cambridge University Press