

Modeling language evolution and feature dynamics in a realistic geographic environment

Rhea Kapur

The Pingry School
rhea.kapur1@gmail.com

Phillip Rogers

Department of Linguistics
The University of California, Santa Barbara
phillip@ucsb.edu

Abstract

Recent, innovative efforts to understand the uneven distribution of languages and linguistic feature values in time and space attest to both the challenge these issues pose and the value in solving them. In this paper, we introduce a model for simulating languages and their features over time in a realistic geographic environment. At its core is a model of language phylogeny and migration whose parameters are chosen to reproduce known language family sizes and geographic dispersions. This foundation in turn is used to explore the dynamics of linguistic features. Languages are assigned feature values that can change randomly or under the influence of nearby languages according to predetermined probabilities. We assess the effects of these settings on resulting geographic and genealogical patterns using homogeneity measures defined in the literature. The resulting model is both flexible and realistic, and it can be employed to answer a wide range of related questions.¹

1 Introduction

The distribution of linguistic feature values across time and space is uneven. The overall frequencies of particular feature values can be skewed with some highly frequent and others vanishingly rare, and many feature values have been observed to cluster geographically or within particular language families. Such patterns are most commonly explained in terms of the competing influences of inheritance, diffusion, universal tendencies of language drift, and chance. Understanding these dynamics has long been a goal of linguists, and it has important implications for historical linguistics, areal linguistics, and linguistic typology. Despite growing attention to these issues and advances in comparative and quantitative methods, many questions remain unanswered.

The well-known problem of spatial autocorrelation limits the usefulness of a direct statistical approach to disentangling geography from genealogy, and as a result various other approaches have been innovated. There are a number of studies exploring the stability or biases of typological features within language families (Dediu and Cysouw, 2013; Dediu and Levinson, 2012; Maslova, 2000; Wichmann and Holman, 2009). Of these approaches, those that account for geography are limited to comparisons within families or between families. Other successful approaches which account for both genealogy and geography are limited to a particular region (Cysouw, 2013), use typological distance measures that do not allow for the comparison of features separately (e.g. Comrie and Cysouw, 2012), or apply a ratio of genealogical and geographic measures that reduces independently varying factors to their quotient (Parkvall, 2008).

Computer simulations have also emerged as a promising method by which to explore these and related issues (e.g., Holman et al., 2007; Oliveira et al., 2008). However, nearly all of these models employ abstract representations of space, such as a lattice. A recent exception is Wichmann (2017), which uses ‘populated places’ from the GeoNames database to add geographic realism to its simulations of language

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

¹Source codes and data are freely available at https://github.com/rkapur102/language_phylogeny_feature_simulations.

family expansions. However, Wichmann’s simulations are limited to individual language families, so they do not model interactions between language families.

In this paper, we introduce a model for simulating languages and their features over time in a realistic geographic environment. The foundation of this model—and a prerequisite for exploring feature dynamics—is a reliable model of language speciation, migration, and death. We will collectively refer to these aspects of the model as phylogeny, and we compare the output of our phylogeny model to real language family data in order to demonstrate its validity. This model is then equipped with a flexible mechanism for representing linguistic features. Feature values are allowed to change randomly or under the influence of nearby languages according to pre-set probabilities, and we use various homogeneity measures from the literature to explore how these settings interact to produce feature value distributions. This model can be employed to answer a wide range of questions related to the distributions of languages and features.

2 Foundations of the Model

Implemented in R 3.6.0 (R Core Team, 2019), our simulations operate over a fixed number of discrete timesteps, where each timestep is intended to represent 20 years. Unless otherwise specified, the simulations presented in this paper run for 500 timesteps (10,000 years), and event probabilities apply to each language at each timestep.

We start with a list of 1.8 million populated places in Eurasia from the GeoNames database². Using data on populated places has been shown to be useful when simulating language phylogeny and migration, as it introduces elements of geographical realism (Wichmann, 2017). For purposes of computational efficiency, we randomly sample 50,000 of the 1.8 million coordinates for use in our simulations. To begin a simulation, unrelated languages are randomly placed at 300 of the populated places. Each of these 300 languages are assigned values of a feature of any number of values. These values can be assigned randomly (such as the simulations presented in this paper) or according to a pre-determined frequency distribution. At each timestep in the simulation, languages undergo phylogenetic simulation—they can either speciate, die, migrate, or remain unchanged—and feature simulation—they can obtain a new feature value via diffusion or internal change.

3 Modeling Languages

The utility of a model for investigating language feature dynamics is limited by its realism with regard to language phylogeny and migration. Put another way, our simulations must accurately model languages before they can be expected to accurately model the features of those languages. In this section, we describe the simulation of language speciation, extinction, and migration. Then, we test the validity of our phylogenetic model against characteristics of real language families and their geographic dispersions.

In order to accurately model language evolution in time and space, it is essential for languages to undergo real-world changes to their phylogeny: namely, speciation, death, and migration. These phylogenetic aspects of the simulation are critical to consider to establish a realistic basis before examining feature dynamics. We first discuss their details and validations, and then delve into features.

3.1 Speciation and Extinction

Language speciation is modeled via splits. The result of a split is two languages; one will remain in the original location, and the other will be given a new location. The location of the second language is determined by the same process for language migration described in Section 3.2. In the case of language extinction, the affected language will simply be removed from the simulation.

Wichmann (2017) provides a geographically realistic model and some justification for speciation and extinction probabilities. However, Wichmann’s simulations begin with a single language growing into a family in a vacuum—that is, without competing languages or language families. The result is that Wichmann’s simulations nearly always end up with far more languages than they start with. To rectify this, we set an extinction rate that is proportional to the speciation rate. In doing so, we model a state of relative equilibrium in terms of the number of languages. We discuss in further detail the refinement

²Available at <http://www.geonames.org/>.

of these birth and death rates to match real language family distributions in Section 3.3. Note, however, that our model can be easily adapted to simulate scenarios of disequilibrium as well.

An important additional mechanism of our model of language phylogeny is a time limit on language family relations. After 300 timesteps (representing 6000 years), a genetic link between two languages (and their descendants) is lost. In other words, two branches of a language family are no longer considered related to each other after 300 timesteps, and they become two separate families. This design is meant to reflect the well-documented limits on determining relationships among languages and language families at deep time depths (Nichols, 1992; Ringe, 1995). By treating these deep genetic relationships as unknown, the output of our simulations can be compared to data on real language families.

3.2 Migration

To account for language migration (including the placement of a new language during speciation), we utilize the quadrilateral-based method described in Wichmann (2017). A brief description of this method is offered here. First, the simulation defines a quadrilateral of area 1 square degree around the current language location by adding and subtracting a *start_size* of 0.5 degrees from the latitude and longitude. The number of populated places falling inside the quadrilateral (from the 50,000 selected at the start of the simulation) is then counted. If this number exceeds a set quantity *ch*, which is the minimum number of acceptable populated places within the quadrilateral, the algorithm will select a migration location randomly from those places. If not, the algorithm will repeatedly add a set number of degrees to each side of the quadrilateral until the number of populated places within its boundaries exceeds *ch*, at which point the migration location will be chosen.

Simulations revealed that an unintended consequence of this migration method is the clustering of simulated languages in areas of densely-aggregated populated places. In other words, languages would gradually migrate out of sparsely populated areas and into densely populated areas. To combat this effect, we added a mechanism of preferred migration into areas underpopulated by other simulated languages.

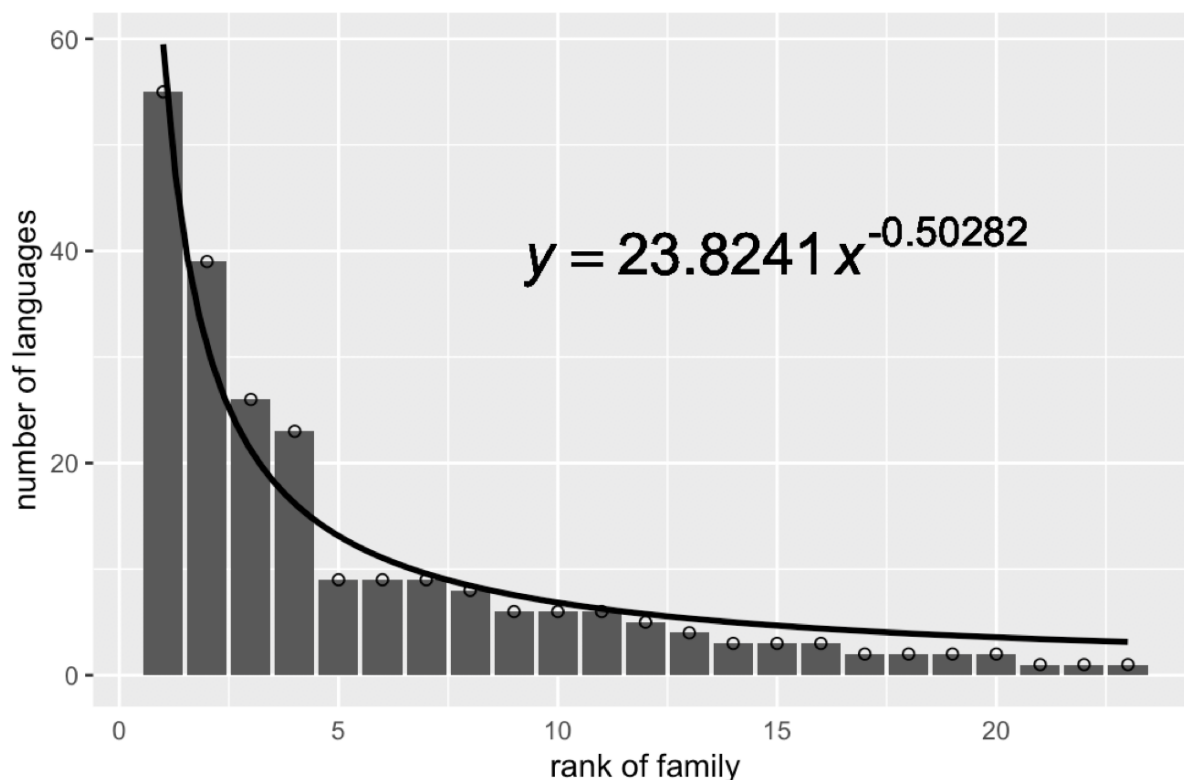


Figure 1: Power-law distribution of simulated language family sizes with speciation and extinction rates set to 2%. The trendline is a nonlinear least-squares regression with formula $y=ax^{-b}$.

When ch is exceeded, rather than choosing a random populated place, the modified algorithm selects ten possible candidates. A new quadrilateral is drawn around each of these ten candidates, and the number of simulated languages (not populated places) within each is counted. The populated place surrounded by the least amount of simulated languages is chosen. This mechanism encourages the migration of languages into less crowded areas.

3.3 Language Family Sizes

It is well-known that language families approximate a power-law distribution (Wichmann, 2005). When language families are plotted in descending order according to the number of members they have, the distribution closely models a curve by the form $y=ax^{-b}$. In our simulations, the distribution of language family sizes is modulated by speciation and extinction parameters. As mentioned above, we chose to keep speciation and extinction rates equal in order to simulate an approximate equilibrium in the number of languages in the model. We ran simulations co-varying this parameter from 1% to 10%, and we found that 2% yields a distribution closest to that of real language family sizes. As such, this is the parameter value that we use for the simulations presented in this paper. An example of the power-law distribution resulting from one such simulation is given in Figure 1³.

3.4 Language Family Dispersions

Just as speciation and extinction rates can be validated by their effect on the distribution of language family sizes, migration parameters can be validated by their effect on the dispersions, or geographic spread, of language families. We define the dispersion D of a language family to be the area of the smallest con-

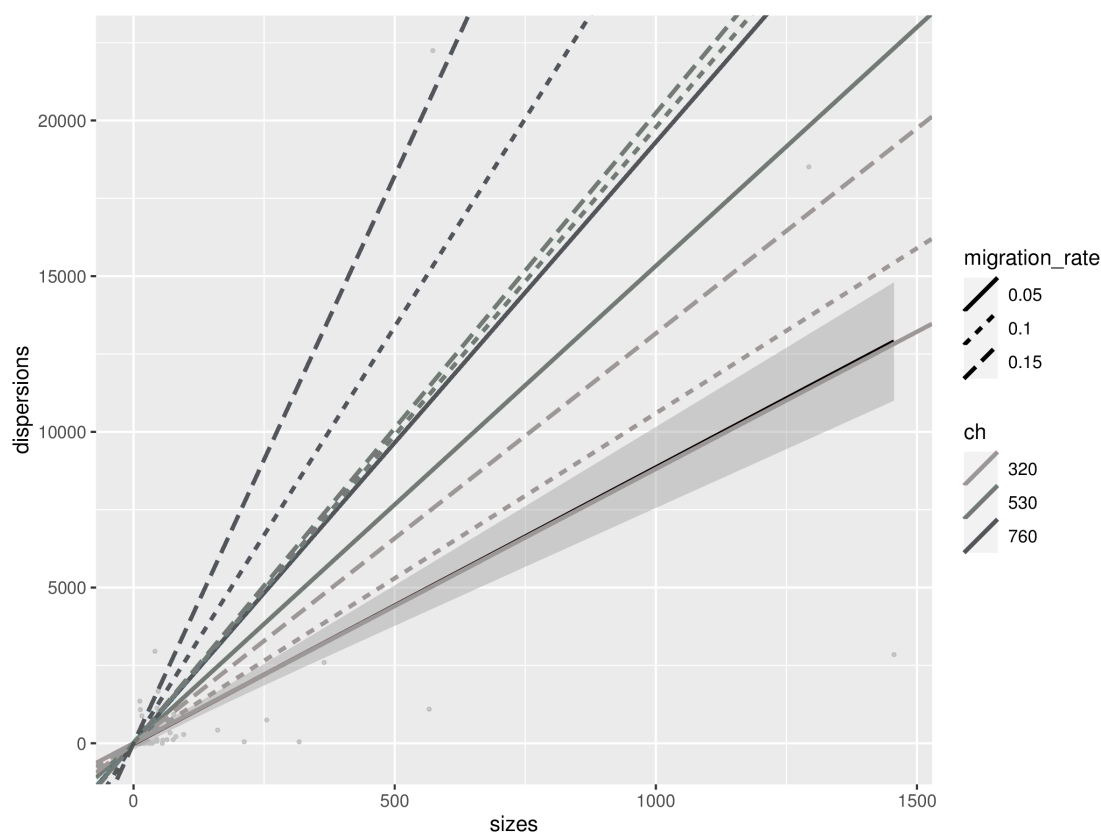


Figure 2: Average simulated dispersion vs. family size for different migration rates and ch values. The black line and accompanying grey confidence band represent the dispersion of real language families, generated from the grey points in the plot.

³All plots in this paper were generated using the ggplot2 R package available at <https://ggplot2.tidyverse.org>.

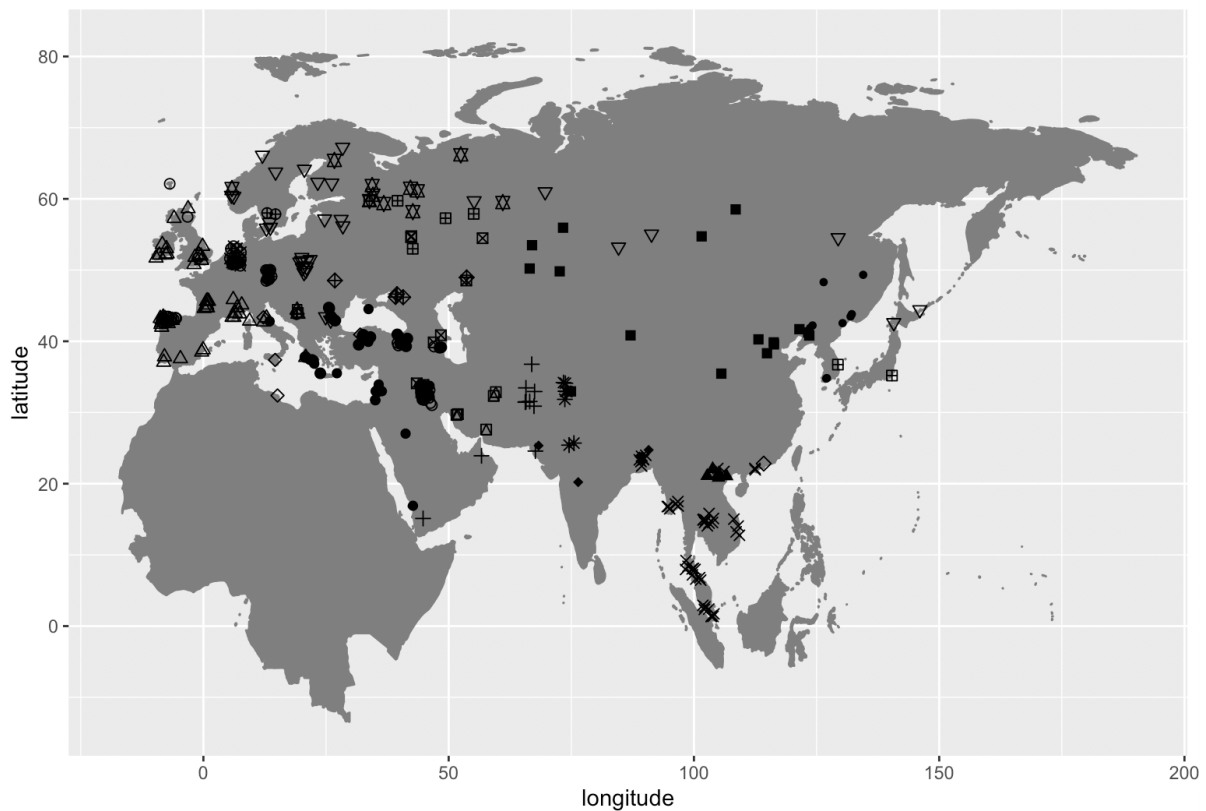


Figure 3: Language families resulting from a simulation of the model. Each shape represents a different simulated family.

vex polygon, or convex hull, that covers all of the coordinates of its languages. The algorithmic problem that details finding the convex hull of a finite set of points in a Euclidean space is a well-known problem in computational geometry. Along with the back-and-forth conversion from geographical coordinates to Euclidean ones, algorithms concerning the convex hull problem become useful here.

To determine the dispersions of real language families, we used language coordinates and genetic classification of 235 language families from Glottolog (Hammarström et al., 2019). (We exclude sign languages as they are not classified genetically in Glottolog.) For each family, we calculated the dispersion in square kilometers using the R package *GeoRange* (v0.1.0; James Boyle Developer, 2017). The backend of this dispersion calculation relies on a convex hull algorithm proposed in Eddy (1977). Finally, we plotted the dispersions as a function of language family size. The resulting trend for the real language data is the darkest black line (and grey confidence interval) in Figure 2. The slope of this line is 8.87 square kilometers per language. (Note that for families containing 2 or fewer languages, the dispersion is 0 square kilometers, as a convex polygon cannot be drawn for lack of vertices.)

In our model, parameters that affect family dispersion are the migration rate and aforementioned *ch* value. We ran a series of simulations varying these parameters to determine the migration rate and *ch* value that would produce a dispersion slope similar to that of real language families. In all, nine different simulation settings were used representing every possible combination of three values of migration rate and three values of *ch*. (The values of *ch* are recommended values from Wichmann [2017].) For each setting, ten simulations were run and the resulting dispersion values were averaged. Figure 2 shows the averaged results of these simulations. The settings which produced the best results were a migration rate of 5% and a *ch* of 320. In fact, the slope for these settings is almost exactly the same as the slope of that darkest black line representing the real language relationship (the two lines are nearly on top of each other). These migration settings are used in the subsequent simulations exploring feature dynamics.

Figure 3 is an illustration of the phylogenetic aspects of the model. It shows plausible language families realistically dispersed across the continent at the end of a simulation.

	stability	diffusion	internal change
L-D/L-IC	70%	15%	15%
H-D/L-IC	50%	40%	10%
L-D/H-IC	50%	10%	40%
H-D/H-IC	20%	40%	40%

Table 1: Key to feature simulations. Acronyms are defined as follows: L-D/L-IC: low diffusion, low internal change; H-D/L-IC: high diffusion, low internal change; L-D/H-IC: low diffusion, high internal change; H-D/H-IC: high diffusion, high internal change.

4 Modeling Features

Equipped with a reliable model of language phylogeny, we turn to the simulation of linguistic features. Our model is intended to explore one feature at a time, but it is flexible with regard to the characteristics of that feature. Features can have any number of values, and they can be structured or unstructured. (One example of a structured feature is an ordinal feature, in which a particular value might only be allowed to change to an adjacent value in the ordinal structure.) At each timestep, the particular value of a feature assigned to a language is subject to internal change (random), change via diffusion (influence from nearby languages), or no change at all. These outcomes are determined by probabilities specified at the outset of the simulation.

4.1 Diffusion

Diffusion is modeled using another variant of the quadrilateral method described above. When a language is selected for diffusion, a quadrilateral is drawn around it and expanded (according to the process described in Section 3.2) until it contains the locations of at least ten other simulated languages. The relative proportions of each feature value represented among these languages become the diffusion result probabilities. In other words, the language undergoing diffusion will take on the feature value of one of these languages. If most of the surrounding languages have a particular feature value, that value is most likely to be adopted by the language undergoing diffusion. Note that this also means that languages can retain their original feature value via the influence of diffusion. For example, if a language has no tones (vs. level tones vs. contour tones) and is surrounded by many languages that also lack tone, there is a high probability that it will remain without tone when selected for change by diffusion. Note also that the effects of diffusion can reach over long distances as a result of overlapping subsequent diffusion events.

5 Assessing Model Output

To explore feature dynamics in the model, we ran a series of simulations varying the probabilities for diffusion and internal change. Four different settings were used, representing the combinations of relatively low and high values of each of these parameters. The exact probabilities for each setting are reflected in Table 1. Each setting was featured in ten separate simulations, and we report the averages for each setting. These simulations were run over 750 timesteps to ensure that the results are not biased by the initial distribution of unrelated languages and random feature values. To evaluate the simulation results, we borrow measures of genealogical and areal (geographic) homogeneity found in the literature. These metrics and their results are detailed in the following sections. They offer a systematic way of assessing the impact of diffusion and internal change on resulting geographic and genealogical patterns.

5.1 Areal Homogeneity

Parkvall (2008) describes measures of both areal and genealogical homogeneity. He combines these measures into a ratio of feature stability, but we use them independently. The areal homogeneity measure is detailed first, and we return to the genealogical homogeneity measure in the next section.

Parkvall's genealogical homogeneity begins with Greenberg's diversity index, a version of the Herfindahl-Hirschman index described in Greenberg (1956; see also Hirschman, 1945; Hirschman,

1964). This diversity measure is defined as

$$D_g = 1 - \sum P_n \quad (1)$$

D_g represents the diversity within an area g of languages, and P_n is the fraction of the entire area that has the feature value n . To get H_g , the measure of homogeneity within an area, the reciprocal of the diversity measure is taken. Next, H_{FAM} , the average homogeneity across all areas, is calculated by taking the averages of all H_g . H_{FAM} is then divided by H_{WORLD} , the average homogeneity regardless of areal distinctions, to get C_{FAM} , the areal homogeneity measure.

We apply this measure to our simulations through yet another adaptation of Wichmann’s (2017) quadrilateral method. In this variation, quadrilaterals are drawn around ten randomly selected languages. These quadrilaterals define the areas on which the homogeneity measure is based. This process can be repeated with quadrilaterals of different sizes to assess the areal homogeneity at different granularities. We compute this measure for quadrilaterals with lengths of 2° , 5° , 10° , 25° , and 50° .

Figure 4 illustrates changes in areal homogeneity at these granularities over the course of simulations for each of the four diffusion/internal change settings. Recall that the leftmost values in each pane reflect the random distribution of values at the start of the simulations. As such, there is little differentiation among the different settings. As the simulations progress, the differential parameters for diffusion and internal change take effect. As expected, the highest areal homogeneity values generally belong to the high diffusion/low internal change simulations. This is because the effect of diffusion is to assimilate languages in proximity to each other, while the effect of internal change is to disrupt this homogeneity with randomness. On the other hand, the lowest areal homogeneity values generally belong to the low diffusion/high internal change simulations (with the exception of an apparently anomalous spike in the final timesteps).

One of the most important insights from this exercise is the evidence for spatial autocorrelation. Even when diffusion is held constant, the simulations show consistent effects of internal change rates on areal

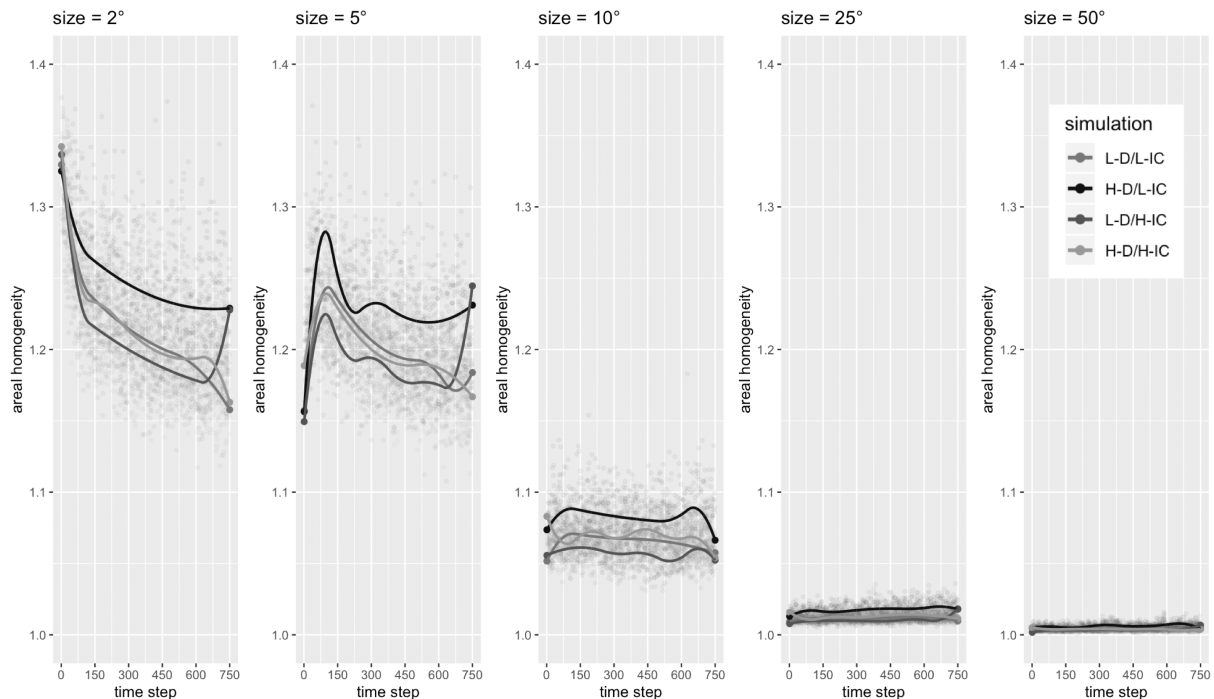


Figure 4: Parkvall’s (2008) areal homogeneity metric at different granularities (area sizes), calculated at each timestep and averaged across simulations of the same settings. Trendlines are constrained B-spline curves to preserve endpoints while also maintaining overall regression fit⁴.

⁴Spline curves were generated using the cobs R package available at <https://CRAN.R-project.org/package=cobs>.

homogeneity. The inverse of internal change is genealogical stability, and this stability cannot be disentangled post-hoc from geographic measures. Not only do these simulations provide evidence for this spatial autocorrelation, but they offer a means by which to investigate the scope and dynamics of this pervasive phenomenon in future work.

5.2 Genealogical Homogeneity

For genealogical homogeneity, we use the measures described by Parkvall (2008) and Wichmann and Holman (2009). Parkvall’s (2008) measure of genealogical homogeneity is, unsurprisingly, quite similar to his measure of areal homogeneity. First, the degree of linguistic diversity is calculated through the aforementioned Greenberg index. The diversity measure mirrors (1), with the differences being that D_g now represents the diversity within a language family rather than the diversity within an area, and P_n is the proportion of languages in the family in question with each feature value. The calculation proceeds exactly as outlined in Section 5.1 to arrive to C_{FAM} , in this context the genealogical homogeneity measure.

‘Metric C’ of Wichmann and Holman (2009) works as follows: for each family and for all pairs of languages, the proportion of pairs that have the same feature value is recorded. This proportion is then averaged across all families, with each family weighted by the square root of the total number of language pairs in that family. This weighted averaged proportion is denoted by R . Another proportion, U , is calculated in the same way for just one group consisting of all unrelated language pairs. The final formula for the stability, S , is

$$S = \frac{R - U}{1 - U} \quad (2)$$

Just as for the areal homogeneity measures, these two genealogical homogeneity measures were calculated at each timestep of our feature simulations. Both measures revealed the same trends, and the results

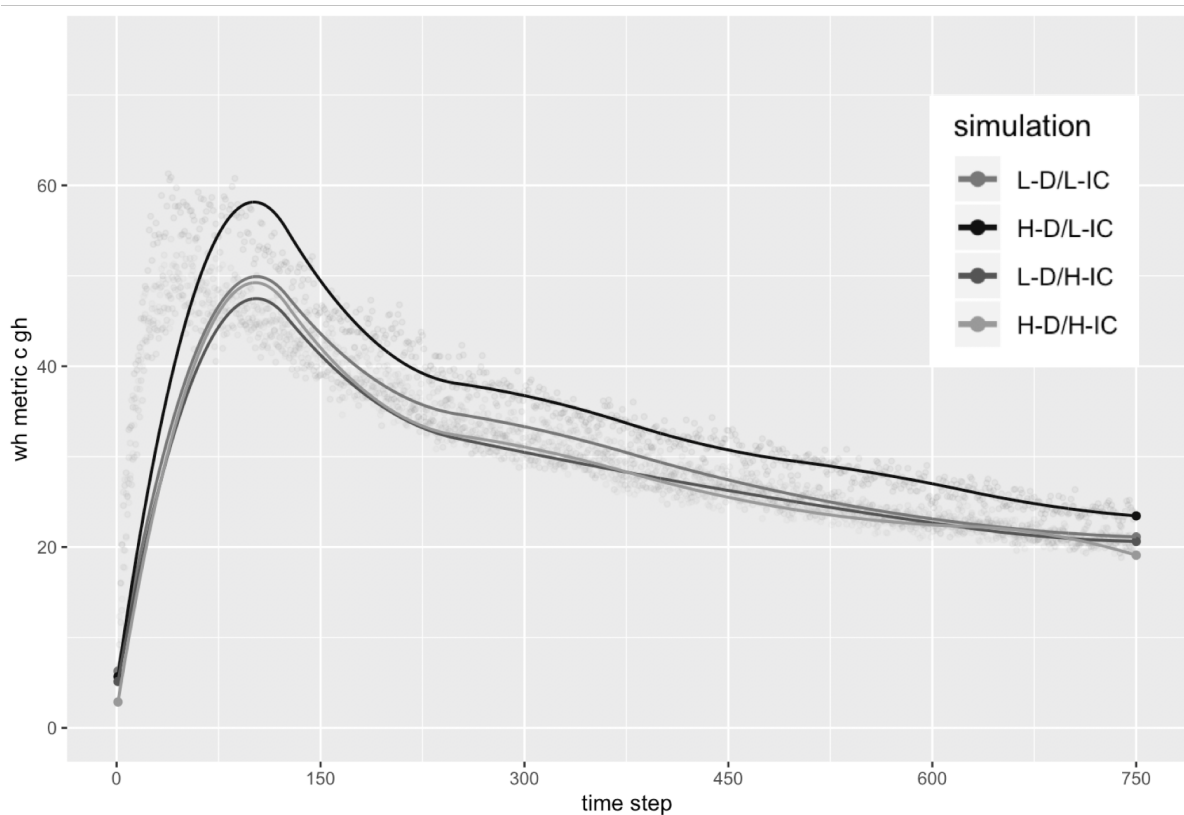


Figure 5: Wichmann and Holman’s (2009) genealogical homogeneity metric, calculated at each timestep and averaged across simulations of the same settings. Trendlines are constrained B-spline curves to preserve endpoints while also maintaining overall regression fit.

	high	medium	low
probability of change from value B to value A	80%	70%	60%
probability of change from value A to value B	20%	30%	40%

Table 2: Key to transitional probability simulations for a two-valued feature in which value A is the dominant feature value.

of the Wichmann and Holman (2009) metric can be found in Figure 5. The same settings that produce high areal homogeneity values—high diffusion and low internal change—produce high genealogical homogeneity as well.

6 Case Study: Transitional Probabilities

There is tremendous potential in the simulation of feature dynamics, and we have barely scratched the surface of it in this paper. As we mentioned earlier, the model can handle structured features. To what extent do structured features behave differently from unstructured features? Furthermore, consider the possibility that transitional probabilities for diffusion and/or internal change might be different for each value of a feature. In other words, the probability of changing from value A to value B may be different than the probability of changing from value B to value A. The model is already equipped with an adjacency matrix to represent a directed graph with connections from each feature to every other, where weights on the edges are the transitional probabilities. With structured features at play, the model also recalculates the chances of diffusion and internal change from the general transitional probabilities, dividing them up in a 3:1 ratio to allow for the necessary dependent values.

We present here a preliminary exploration into structured feature dynamics, examined in the context of a two-valued feature. We ran a series of simulations reflecting three sets of transitional probabilities. Exact values for each setting are detailed in Table 2. Similar to the simulations described in Section 5, each setting was featured in ten separate simulations, and we report the averages across each set. These

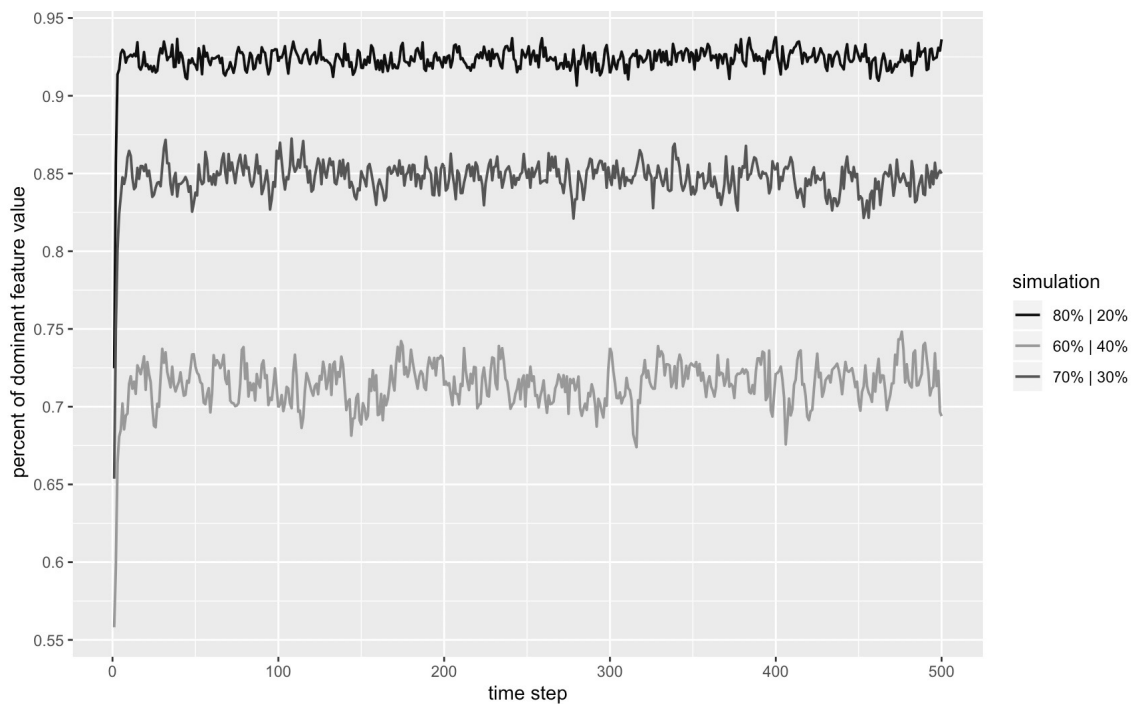


Figure 6: Percentage of dominant feature value (favored as result of change) for three experiments with varying transitional probabilities (reflected in Table 2). Calculated at each timestep and averaged across simulations of corresponding setting.

simulations were run over 500 timesteps to ensure the results are not biased by the initial distribution of random feature values. Of particular interest in this case study is whether and to what degree diffusion will interact with the transitional probabilities to amplify or diminish the skew of the resulting feature value distribution.

Figure 6 illustrates the change in the percentage of the dominant feature value (favored as result of change) over the course of simulations for each of the three transitional probability settings. These results reveal an amplified skew in the distribution of resulting feature values relative to the preset transitional probabilities. For example, transitional probabilities of 70%—30% result in a feature value distribution in the vicinity of 85%—15%. These findings demonstrate how even relatively minor differences in the underlying stability of feature values can be amplified significantly by the dynamics of languages in space in time. In particular, diffusion plays an important role in magnifying these feature value inequalities. This is just one of many potential explorations into the complex ways in which language feature dynamics produce particular feature value distributions, all of which require a model of suitable flexibility and realism.

7 Further Applications

The primary contribution of this paper is a tool which can be used to investigate a myriad of unresolved issues concerning the distribution of languages and their features across time and space. A few further extensions and applications of the model are discussed here.

The realistic geographic environment featured in the model make explorations into spatial phenomena an obvious area for extension. For example, much is unknown about how local diffusion events and/or deep genealogical relationships can conspire to create long-distance areal effects. Similarly, the existence of linguistic areas presents a nice challenge for our model. Which settings can result in the development of such areas?

The model of phylogeny at the heart of our simulations also invites investigations into the social/cultural and geographic factors that may affect competition between languages and language families. For example, the model could easily be adapted with a mechanism of temporary speciation ‘momentum’ bestowed upon language families with a political or technological advantage. Alternatively, the model could be equipped with information about climate or terrain to constrain the movements of and influences among different language groups.

8 Conclusion

In this paper, we have presented a new model which brings a realistic geographic environment to the simulation of languages and linguistic features over time. This model is realistic, producing plausible language family sizes and dispersions. At the same time, it is also flexible, constructed with adaptability in mind and equipped to model features of various structures and properties. We have also shown how our model can be utilized to address a wide range of related issues with implications for historical linguistics, areal linguistics, and linguistic typology. It opens the door for further discoveries in the evolution of language and patterns of human migration and interaction.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions, and the COLING 2020 organizers for their efforts. Phillip Rogers is the corresponding author of this paper.

References

- James Boyle Developer. 2017. GeoRange: Calculating Geographic Range from Occurrence Data. R package version 0.1.0.
- Bernard Comrie and Michael Cysouw. 2012. New Guinea through the eyes of WALS. *Language & Linguistics in Melanesia*, 30.

- Michael Cysouw. 2013. Disentangling geography from genealogy. In Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi (eds.), *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, 21–37. Berlin: De Gruyter Mouton.
- Dan Dediu and Michael Cysouw. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS ONE*, 8(1).
- Dan Dediu and Stephen C. Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE*, 7(9).
- Matthew S. Dryer and Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- William F. Eddy. 1977. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software*, 3(4). 398–403.
- Joseph Greenberg. 1956. The measurement of linguistic diversity. *Language*, 32(1). 109–115.
- Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 4.1*. Jena: Max Planck Institute for the Science of Human History.
- Albert Hirschman. 1945. *National power and the structure of foreign trade*. Berkeley: University of California Press.
- Albert Hirschman. 1964. The paternity of an index. *American Economic Review*, 54. 761–762.
- Eric W. Holman, Christian Schulze, Dietrich Stauffer, and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology*, 11. 395–423.
- Elena Maslova. 2000. Stochastic models in typology: obstacle or prerequisite? *Linguistic Typology*, 4. 357–364.
- Johanna Nichols. 1992. *Linguistic diversity in space and time*. Chicago, IL: University of Chicago Press.
- Paulo Murilo Castro de Oliveira, Dietrich Stauffer, Søren Wichmann, and Suzana Moss de Oliveira. 2008. A computer simulation of language families. *Journal of Linguistics*, 44. 659–675.
- Mikael Parkvall. 2008. Which parts of languages are the most stable? *STUF*, 3. 234–250.
- Donald A. Ringe. 1995. ‘Nostratic’ and the factor of chance. *Diachronica*, 12(1). 55–74.
- R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Søren Wichmann. 2005. On the power-law distribution of language family sizes. *Journal of Linguistics*, 41(1). 117–131.
- Søren Wichmann. 2017. Modeling language family expansions. *Diachronica*, 34(1). 79–101.
- Søren Wichmann and Eric W. Holman. 2009. *Temporal stability for linguistic typological features*. Munich: Lincom Europa.