# Testing methods of homeland detection using synthetic data

Søren Wichmann[1,2] and Taraka Rama[3]
[1]Leiden University Centre for Linguistics, Leiden University, Postbus 9515, Leiden, 2300 RA, The Netherlands
[2]Laboratory for Quantitative Linguistics, Kazan Federal University, Kremlevskaya Street 18, Kazan, 420000, Russia
[3]Department of Linguistics, University of North Texas, Discovery Park Room B201, 3940 N Elm St., Suite B201, Denton, TX, 76207, USA.

ID SW, 0000-0002-3257-3087; TR, 0000-0002-4531-6733

**Keywords:**
historical linguistics, homelands, migration, phylogenetics, Bayesian phylogeography, ASJP
**Author for correspondence:**
Søren Wichmann
email: wichmannsoeren@gmail.com

## Abstract
There are two families of quantitative methods for inferring geographical homelands of language families: Bayesian phylogeography and the 'diversity method'. Bayesian methods model how populations may have moved using the backbone of a phylogenetic tree, while the diversity method, which does not need a tree as input, is based on the idea that the geographical area where linguistic diversity is highest likely corresponds to the homeland. No systematic tests of the performances of the different methods in a linguistic context are available, however. Here we carry out performance testing by simulating language families, including branching structures and word lists, along with speaker populations moving in areas drawn from real-world geography. We test five different methods: two versions of BayesTraits; the random walk model of BEAST; our own RevBayes implementations of a fixed rates and a variable rates random walk model; and the diversity method. Each method is tested on the same synthetic family of 20 languages, evolving in 1000 different random geographical locations. The results indicate superiority in the performance of BayesTraits and different levels of performance for the other methods, but overall no radical differences in performance.

## 1. Introduction
Information on the location of proto-languages is important in many studies of linguistic prehistory. Such origins may be of interest in and of itself, such as in the case of Indo-European, whose origins have received much attention and also been subject to controversy [1]. Language group origins may also, for instance, form the backbone for a study relating linguistic typology.

When combined with information on dates for language groups [2], inferences about homelands become particularly powerful as contributions to world prehistory.

In the past, researchers had to rely mainly on *reconstructed* lexical items for clues to the origin of a language group. For instance, if a word for a particular biological species or material item which is diagnostic of a certain geographical or archeologically defined area, can be reconstructed for a proto-language, then the speakers of the proto-language can hypothetically be assigned to the area in question. In addition to this approach, known as linguistic paleontology, other types of evidence that can sometimes be drawn upon are old place names with known linguistic affiliations or early loanwords showing different proto-languages to have been in contact. These various approaches only promise to apply when a language group is already very well researched, and even then there are strong limitations since securely reconstructed and geographically diagnostic proto-words are hard to come by, and information on early loanwords and place names is most often not available.

Instead of, or in addition to, applying linguistic paleontology, researchers have often taken a bird's eye view of a language family, inferring origins and directions of dispersion through an application of what can be called the center of gravity or diversity method. Apparently first introduced by Edward Sapir [3], the idea here is that the area of origin will usually see more diversity building up than areas in which members of a given language group were latecomers. For instance, Sapir himself argued [3] that the large Algonquian family of North America is more likely to have originated in the west than in the east because the most divergent languages are found in the west. The same type of argumentation has been applied to several families in South America [4], to Austronesian [5], to Sino-Tibetan (apparently) [6], and other families. A qualitative (impressionistic) application of this approach is admissible in obvious cases, such as that of Austronesian, all but one of whose subgroups are confined to one and the same area (Taiwan), but in less obvious cases it needs a quantitative implementation of some kind.

Fortunately, homeland detection methods of a more quantitative nature have appeared which only rely on a type of evidence which is available for any language group, namely *basic lexical items as attested in the extant languages*. These approaches are either inspired by or directly stem from approaches in biology. The early part of the 2010's saw both the application of biological (specifically Bayesian) phylogeographical approaches [7-8] and a quantitative implementation of the diversity approach [14]. These methods, however, have come with no warranty in terms of how well they perform. We rarely know the origin of a language group, so it is not obvious how to test the methods on empirical data. Even if we did collect information on cases of known origins and ran different methods to check their results we would not have enough data points to get good statistics on variability in performance or for getting insights about possible systematic causes for challenges to the methods.

The lack of performance tests of linguistic homeland detection methods motivates this paper. We solve the problem of lack of empirical testing data by producing synthetic data on one randomly chosen language phylogeny and its geographical dispersal under 1000 different origin

scenarios and then look at the performance of a range of methods in terms of the distance between the true and inferred homelands. The next section describes this simulated data.

## 2. Data

We simulate a single language family having 20 languages, and each of the languages is associated with a 100-item word list containing words that have evolved over 120 time steps. The simulations are described in Online Appendix 3 to [9].[1] Briefly, the ancestral language is constructed from an inventory of phonological symbols identical to the symbols used in the ASJP database [10]. Proto-words with semi-realistic shapes are constructed, and at each time step these words undergo phonological and lexical changes according to preset probabilities tuned to reality. Probabilities for speciation and extinction, also preset, allow lineages to grow into families of sizes that are not predictable but still probabilistically related to the parameters of the birth-death process and the number of time steps. The program was modified for the present paper to accommodate an output of 100-item rather than 40-item word lists and to include information about cognacy among the simulated words. It is found in the electronic supplementary material (SI-01) along with accompanying files. The program outputs five files whose names are identical except for the suffix. The names are composed of the letter 'F', an underscore, a number representing the number of terminal taxa, another underscore, and then (for distinctiveness of the name) the first word in the first word list in the data for the terminal taxa. The file called '*.par' specifies the parameter settings. The file called '*.top' gives the topology, the file called '*.dat', contains word lists for the terminal taxa in a tab-delimited format, and the file called '*.txt' contains the same word lists in the style of ASJP input files (see http://asjp.clld.org/help). Finally, the file called '*.cog' is similar to '*.dat' except that the words are shown in the shape they would assume if they did not undergo phonological changes. This is a handy way of encoding cognacy relations. For instance, in the F_20_wanej.cog file the first lexical item assumes one of the three different shapes *wim3y*, *kahCehd*, and *bmiu* across the twenty languages, corresponding to three different cognate classes. In the F_20_wanej.dat file the phonological changes are included, and it is seen, for instance, that the word with the original shape *wim3y* has changed into *wanej*, *vo5aj*, *vozej*, etc. during the course of its evolution. Given both the information on cognate relations and actual forms having undergone phonological changes, the simulated data could be used, for instance, to test techniques for automated cognate recognition, currently a rapidly developing field [11]. Here, however, the two alternative ways of presenting the lexical data serve to satisfy the requirements of the different methods that we are testing: the Bayesian methods rely on information about cognacy and are here supplied with the full, accurate information about the relations among 100 words; the diversity method relies on lexical distances and are here supplied with lists of 40 words in their actual shapes as input to the string distance computations. This somewhat less generous input is similar to real data that the method would use.

[1] Posted at https://datadryad.org/stash/dataset/doi:10.5061/dryad.4gg07

For the present study we selected one particular simulated family that seemed adequate in terms of its size—it is large enough to contain plenty of data for the methods to work with but not too large to be unwieldy. Its topology is shown in Figure 1. The number of time steps was set to 120, but the initial lineage happens to not split until step 77. Thus the root of the lineage is anchored at step 76, which, for all practical purposes, can be considered step 0. The justification for pruning the initial branch is that the methods for detection homelands, even if successful, will only recover the most immediate location of the ancestral language. As is apparent from Figure 1, the two initial lineages split at the same time, 4 steps after the root. At a total of 44 steps (44 + 76 = 120) we reach the tips of the terminal branches.

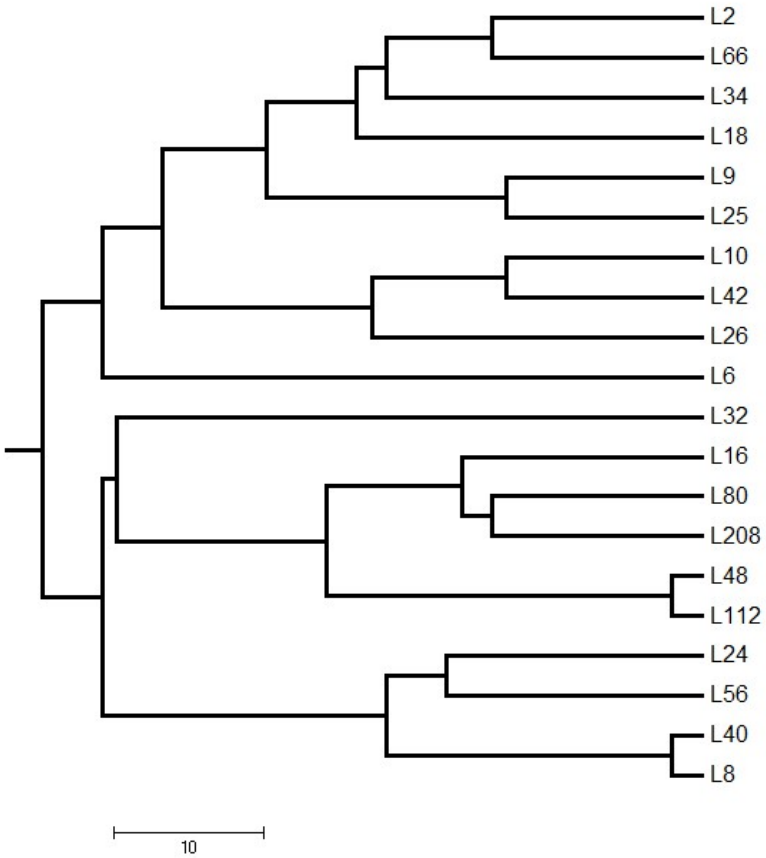


Fig. 1. Simulated tree used in this study showing topology, branch lengths, and taxon names at the tip of each branch.

For simulating the geographical diffusion of languages we again use published software with a few modifications and make the updated software available here (electronic supplementary material SI-02, output in SI-03). Thus, for a full description and justification of the simulation procedure a paper [12] is already available. The simulation process can be summarized as follows. Movements are constrained to any populated place on Earth, i.e. a place included in the geonames.org database. A starting point is found by randomly choosing from this

set of populated places. At each time step there is a preset probability of moving to a new place within a square containing at least *ch* populated places. The *ch* parameter is here set to 800 since it was found to produce results that are realistic in terms of language densities [12]. Apart from the coordinates for the world's populated places, the only input required is a tree structure such as the one displayed in Fig. 1, encoded as in the output for the language family simulation program. Using one and the same phylogeny, we produce 1000 diffusion scenarios. Since these take place in real-life geography they will be sensitive to natural boundaries such as mountains, deserts, and bodies of water. If the diffusion starts on a small island containing less than 800 populated places there is a high chance that a lineage will make a jump away from the island; if it starts in a densely populated area the family may possibly never extend beyond this area. Places that are currently populated work as a proxy for the carrying capacity of different areas of the world. The kind of movement we simulate here may be called a semi-random walk: it would obey the properties of a random walk if it was not constrained to populated places. Since this constraint introduces the possibility of jumps, it is envisaged that a method for reconstructing a linguistic homeland will encounter both cases where languages have a relatively even distribution and cases where languages end up clustering in separate regions. Maps of all 1000 cases, showing the homeland, intermediate stations, locations of current languages, and inferred homelands similarly to Figure 2 below, as well as the script that produced the maps are provided in the electronic supplementary material (SI-12).

The reason why we choose to work with just a single simulated phylogeny unfolding in many different geographical locations is that we hypothesize that the factor that will affect the performance of a homeland detection method the most is the different geographical distributions of extant languages. Other parameters, such as the size of families, the language change rate or the extinction rate, could be varied, but each additional parameter would imply a multiplication of computational efforts that are already large. Thus, we focus exclusively on the role of the most interesting and complex parameter, namely geography, in the performance of homeland detection methods. The methods thus tested are described in the next section.

## 3. Methods

For each of the methods tested here and described in subsections (a)-(g), we provide all files needed for replication in the transparently named folders contained in the electronic supplementary material (one folder per method).

### (a) Baselines (rand, centr, md)

As baselines we apply three approaches. The first approach, abbreviated 'rand', picks a random language in the family and assigns the homeland to the location of that language. Since this is merely a baseline and not a method meriting deeper investigation we just did a single run across the 1000 cases, even if the results would vary somewhat as different languages are selected. We contend that across 1000 cases a good, real method should always have a better average performance than this baseline, even if the latter might often 'strike it lucky'. The second

approach, abbreviated 'centr' for 'centroid' is built on the simple-minded assumption that the homeland is located in the center of the polygon represented by the extension of current languages. The centroid location is computed using the function 'centroid' of the R package geosphere [13]. In the third approach, abbreviated 'md' for 'minimal distance', we compute the average distance (as the crow flies) from each language to all the other languages. The location of the language with the smallest average distance to the others is equated with the homeland.

## (b) Fixed rates model of BayesTraits (BTF)

The fixed rates model implemented in BayesTraits [16] is a Brownian motion model where the latitudes and longitudes are mapped to the three dimensional Cartesian coordinate system. The three dimensional coordinates are treated independently in this model. In the BTF model, there is a single parameter, the variance of the normal distribution, which is sampled using a Monte Carlo Markov Chain (MCMC) procedure along with the three dimensional coordinates. The fixed rates model takes a single tree with branch lengths and the geographical coordinates of the 20 languages as input and then reconstructs the internal nodes' geographical locations using the MCMC procedure.

In this paper, we infer the phylogenetic tree of the 20 languages using a Metropolis-coupled MCMC ($MC^3$) procedure. In the MCMC procedure, the cognate data of the 20 languages is supplied as an input to the MrBayes phylogenetic software [26]. We use the uniform model [20] that infers rooted trees using a binary continuous time Markov chain model of lexical evolution with sites weighted by a four-category discrete Gamma distribution. We perform two independent runs starting from two different randomly initiated starting points. Each independent run consists of three hot chains and one cold chain to navigate the multiple peaked tree landscape efficiently without getting stuck in local optima. An MCMC chain is run for 10 million iterations where every $1000^{th}$ sample is written to a file. The first 25% of the iterations are discarded as burn-in. We constructed a majority consensus tree from the set of 7500 trees and used the majority rule consensus tree as input to both the BTF and the BTV BayesTraits models (subsection c), as well as to the RevBayes [17] model (described further below).

## (c) Variable rates model of BayesTraits (BTV)

The variable rates model is a relaxation of the single parameter Brownian motion [21], which assumes that the rate of change is fixed across all the branches. In this model, the branch lengths are allowed to shrink or expand reflecting large movements in space. In contrast to the MCMC models, where the number of parameters is fixed through the sampling process, the BTV model has two parameter changes: whether to scale a branch and to sample the scaling parameter of a particular branch. The decision to scale a branch increases the number of parameters by 1 and requires the use of a Reverse Jump MCMC (RJMCMC) procedure. The RJMCMC procedure does not easily converge, requiring long running times depending on the number of languages. In this paper, we run the model for 1 million iterations (sampled at every 1000th iteration) preceded

by a burnin of 100,000 iterations. In addition, we discard the first 500 samples and only use the sample with the best likelihood for evaluation purpose.

### (d) The random walk model of BEAST (BRW)

The random walk model as implemented in BEAST 2.6.3 [22] features a joint phylogenetic inference and phylogeographic model. In the phylogenetic inference model, lexical evolution follows a covarion model where some cognate sets are allowed to change faster than others. The tree model of evolution is based on a birth-death model where the height of the tree is drawn from a Gamma distribution with mean and standard deviation set to 1. The phylogenetic inference model also has a relaxed branch rates model where the branch rates are drawn from a uncorrelated lognormal distribution [24].[2] The phylogeographic model is a relaxation of the Brownian motion model where the likelihood of latitude and longitude drawn from independent uniform distributions are estimated using a bivariate normal distribution. The parameters of the variance matrix of the bivariate normal distribution are scaled by a lognormally distributed scaler for each branch leading to a separate variance matrix for each branch. The model was run for 50 million iterations with every 1000th iteration written to a file. The first 50% of the samples were discarded as part of burnin. Out of the remaining 25,000 samples, the geographical coordinates from the sample with the best posterior probability is used to evaluate the model.

### (e) RevBayes implementation of a fixed rates random walk model (RBF)

We implemented a simple model in RevBayes where the geographical coordinates are both drawn from uniform distributions of real numbers in the [-90, 90] range for latitudes and the [-180,180] range for longitudes. The phylogenetic tree is the majority consensus tree inferred from MrBayes software described in section 3(b). The standard deviation parameter of the normal distribution is assumed to be drawn from a uniform distribution on a log scale. The standard deviation parameter along with latitudes and longitudes are sampled using an MCMC chain. As part of burnin, the chain was run for 5000 iterations followed by a run for 50,000 iterations sampled at every 100th iteration to reduce autocorrelation.

### (f) RevBayes implementation of a variable rates random walk model (RBV)

The RBV model [23] allows for variable rates of evolution among branches using a relaxed model where each branch's rate is allowed to shift or not to shift. If there is no shift, the rate for the branch is the same as the rate for its parent branch. If there is a rate shift, the parent branch's rate is scaled by a scalar that is sampled through an MCMC procedure. The rate parameter at the root node is sampled through an MCMC move. The rate shift probability is drawn from a Gamma distribution with mean 1 and variance set to 0.33. The probability that there is no rate shift is given as the expected number of rate shifts (5) divided by the number of branches (2*20-

[2] The XML file was crafted based on the tutorial provided here: https://taming-the-beast.org/tutorials/LanguagePhylogenies/

2 = 38). The rate shift multiplier being 1 (no rate shift) or not 1 is sampled using a Reverse Jump MCMC move since a value not equal to 1 would increase the number of parameters in the model by 1. We ran the model for 100,000 iterations with every 200th iteration written to file. The run was preceded by a burnin of 5000 iterations. RBV uses the same phylogenetic tree and nexus files as the RBF (see Section 3b) model as input.

### (g) The diversity method (Div)

This method represents a quantitative implementation of the old idea in linguistics and biology [3,18] that the area of greatest diversity of a family/species is most likely the homeland. We follow [14], where the method is presented and results for empirical data from across the world's language families are discussed. An actual program for doing the calculations has not been published previously, but is provided in the electronic supplementary material (SI-10). Since the method is already amply described [14], we only provide a summary here. A basic tenet of the method is that it is a reasonable approximation to assume the location of one of the current languages can be identified with that of the homeland. Given this assumption, the question becomes how to assign diversity values to different languages. This question is answered by assuming that larger linguistic distances coupled with greater proximity to the linguistically distant relatives means more diversity. This solution is then implemented quantitatively by computing all pairwise linguistic distances as well as all pairwise geographic distance (as the crow flies) among the languages. The linguistic distances are average string distances across word lists, the particular string distance used being defined as as LND or Levenshtein Distance Normalized—the Levenshtein distance divided by the length of the longest of the two words compared—divided by the mean LDN across word pairs not referring to the same concept. This is called the LDND (Levenshtein Distance Normalized Divided) [19]. Now a diversity index $D_L$ for each language $L$ is calculated as the mean of the proportion of linguistic to geographical distance between a given language and each of the other languages. The location of the language with the highest $D_L$ value is assumed to be the homeland.

## 4. Results

Results for the different methods are given as distances in km (as the crow flies) from the true to the inferred homelands. For the Bayesian methods we ignore the fact that they provide a set of plausible homelands, i.e., inferred areas rather than single locations. This is of course an advantage of these methods that should carry weight in a qualitative, comparative assessment. But for the purposes of a quantitative performance comparison it is necessary to operate with a single location whose validity can be assessed as a number. For BayesTraits (BTF, BTV) we select the location associated with the highest likelihood (the software only outputs likelihoods, not posterior probabilities) and for BEAST (BRW) and RevBayes (RBF, RBV) we select the location with the highest posterior probability.

Table 1 provides mean and median absolute errors (distances). It is based on a larger table for individual cases provided as a file summary_errors.txt in the electronic supplementary

material (SI-11). Across the table, we observe large differences in means and medians, suggesting that errors are not normally distributed, outliers having a large impact.

Table 1. Mean and median absolute errors (in km, rounded)

| Error | rand | centr | md | BTF | BTV | BRW | RBF | RBV | Div |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 214 | 353 | 153 | 148 | 172 | 179 | 167 | 172 | 183 |
| Median | 156 | 160 | 106 | 103 | 114 | 123 | 118 | 125 | 131 |

While the absolute errors in Table 1 are suggestive of some performance differences they should be taken with a grain of salt since the absolute errors for baselines and methods are generally highly correlated with the mobility of a family. The latter can be measured as the total distance traversed at all time steps by the twenty lineages. For the centr (centroid) baseline the correlation is small although still significant ($r = .173$), but this is also the 'odd man out' in terms of performance. For the other baselines and methods the correlations, which are all significant at the $p < .00001$ level, are in the range $.657 \leq r \leq .721$. Given correlations between absolute error and mobility we prefer to assess performance using relative error, calculated as the absolute error divided by total distance traversed. These numbers are displayed in Table 2, multiplied by 100,000 and rounded to integers for easier viewing (results for individual cases are in the file summary_relative_errors.txt in the online electronic supplementary material, SI-11).

Table 2. Mean and median relative errors

| Error | rand | centr | md | BTF | TV | BRW | RBF | RBV | Div |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 1137 | 2051 | 794 | 781 | 884 | 931 | 878 | 915 | 955 |
| Median | 1008 | 997 | 682 | 679 | 749 | 805 | 773 | 791 | 833 |

The results in Table 2 allow us to establish a hierarchy of methods and baselines. According to the results for means the ranking is as follows:

$$\text{BTF, md} < \text{RBF, BTV, RBV, BRW, Div} < \text{rand, centr}$$

To construct this hierarchy, we first ordered the methods and baselines, separating them by the less-than symbol. Subsequently we applied a pairwise Wilcoxon rank sum significance test to assess the differences between *neighbors* in the hierarchy and replaced the smaller-than sign by a comma when neighbors were not significantly different at the .05 level. This caused

BTF/md, RBV/BRW/Div, and rand/centr to merge. We then iterated this procedure, also replacing the less-than sign with a comma when there were non-significant differences between members of neighboring *groups*. The relevant, non-significant pairs here were BTV/BRW, RBF/BRW, RBV/DIV, and RBF/RBV. The resulting picture is the one where there are just three groups. BayesTraits (specifically BTF) works best, followed by RevBayes (RBF, RBV), BRW, and Div, which are rather indistinguishable. All the methods beat the baselines rand and centr, which provides a good sanity check on the results. A surprising result, however, is that the md (minimal distance) baseline shares the winning position with BayesTraits. This baseline does not draw on any linguistic information, only geographical coordinates. It is 'one half' of the diversity method (Div) in the sense that Div uses average geographical distances in its denominator in the calculation of diversity indices used to select the best homeland language. Apparently the numerator does little to improve on the utility of this measure for determining which language location is the best proxy for the homeland. Nevertheless, we still contend that md is best regarded as a baseline, not a real method.

## 5. Discussion

As a first step towards comparing the methods we did a simple pairwise correlation of errors across methods, also including the baselines (the full matrix of correlations is in the electronic supplementary material folder SI-12). It turns out that these errors are highly correlated, except for the baselines rand and centr. For all pairs of methods (also including md) correlations lie in the range $.809 < r < .935$. Given the overall similar performances it does not seem productive to look for factors that systematically incur *differences*. Instead, we will look for factors that affect the performance in *similar ways* across methods. The large differences in mean and median errors (Tables 1 & 2) indicate that the variability of performance is mainly due to large outliers. Inspection of pictures showing inferred and true homelands for each method (included in the electronic online supplementary material folder SI-12) suggest that outliers are mainly due to a situation where all current languages are found in areas other than that of the homeland, presumably due to early geographical movements pertaining to the two main lineages. Figure 2 shows such a situation where languages have ended up in two different areas, apparently as a result of early jumps out of the homeland of both major lineages. All the six methods infer wrong homelands, but differ a little in the magnitude of the errors.

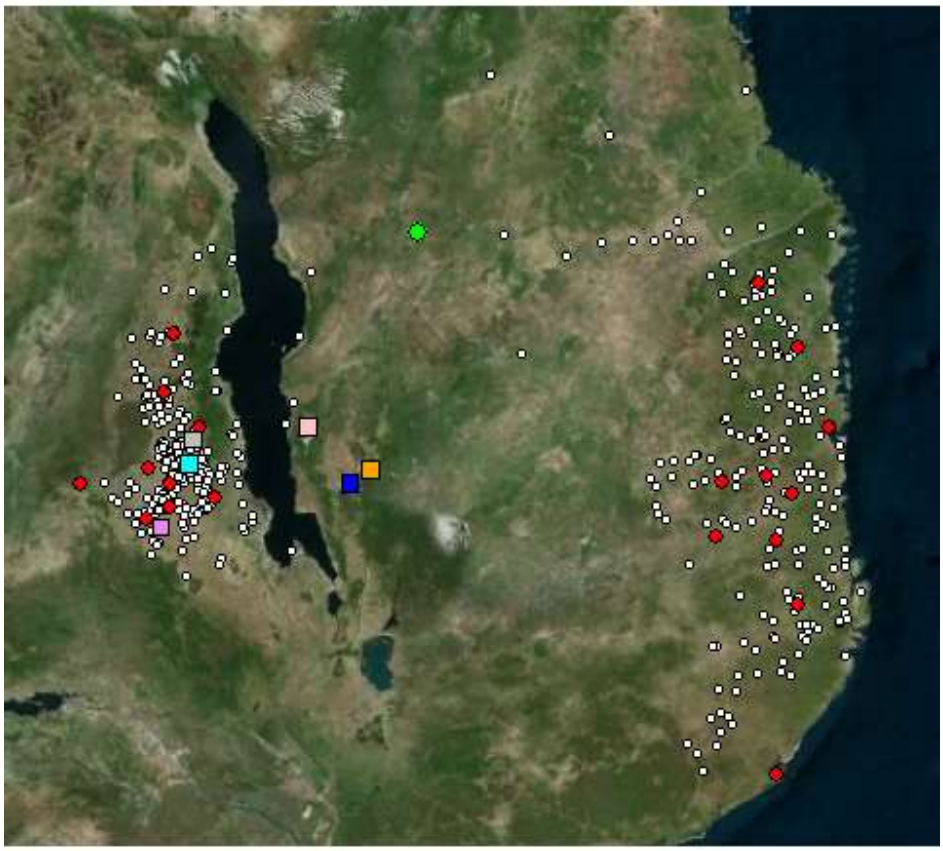

Fig. 2. A simulation scenario (file prefix F_20_wanej_-10.9_36.15) where there were early jumps out of the homeland, causing all six methods to err. Red dots: extant languages; white dots: intermediate migratory stations; large green dot: true homeland; blue square: BTF; cyan square: BTV; violet square: BRW; orange square: RBF; pink square: RBV; gray square: Div.

If it is correct that large outliers in errors are mainly due to early jumps out of the homeland, then we can gather from the tree structure in Figure 1 that these jumps would have occurred during the first 4 time steps of the simulations. In order to corroborate this hypothesis we first compute 'early jump indices' as the percentage of the distances traversed during the first 4 time steps out of the sum of all distances traversed (results are in the file percent_early.txt in the online supporting material folder SI-12). We then gauge the impact of early jumps on errors in two different ways, as follows: (a) early jump indices are correlated with relative errors across the 1000 cases; (b) following the traditional way of defining outliers in a histogram, taking a larger outlier to exceed a value of 1.5 times the interquartile range, we identify outliers in the early jump indices as well as in the relative errors and then compute an agreement index as the percentage of large error outliers that pertain to cases which are also large outliers in terms of their early jump indices. If errors are in a large part due to early jumps, the correlations produced in approach (a) should be high, but the correlations are based on all 1000 sets, so factors other

than early jumps affect the results. Approach (b) is more focused just on the early jumps, but the way of singling out outliers is somewhat arbitrary. In spite of these shortcomings, approaches (a) and (b) should give us some insights into the effects of early jumps. Results for all baselines and methods except for centr, for which the correlations (approach a) were non-significant, are displayed in Table 3.

Table 3. The relationship between the presence of early jumps and absolute errors

| Approach | rand | md | BTF | BTV | BRW | RBF | RBV | Div |
|---|---|---|---|---|---|---|---|---|
| correlation | .418 | .681 | .655 | .632 | .604 | .569 | .568 | .545 |
| agreement | 44.4 | 71.4 | 69.4 | 64.5 | 55.6 | 56.4 | 60.6 | 46.7 |

Table 3 shows that the two approaches to measuring the error rates' sensitivity to early jumps agree (Spearman's $\rho$ = .905) and that they both indicate relatively high sensitivities. They are both strongly negatively correlated with the average mean error ($\rho$ = -.857 and $\rho$ = -.905 for correlation and agreement, respectively). This indicates that while early jumps is a problem for all methods, the best methods are less affected by *other* factors causing large errors.

Besides quantitative assessment in terms of error measurements the methods also deserve to be assessed qualitatively in terms of other performance aspects. Div and md require less than a minute of running time on a laptop and BRW requires close to an hour for each case on a server, with the other methods being intermediary. The Bayesian methods require cognate information while Div only requires word lists, and md not even word lists. Div and md will avoid assigning homelands to uninhabitable areas, including water bodies, since they are restricted to populated locations. All the Bayesian methods give samples of coordinate sets associated with likelihoods and/or posteriors measuring confidence as opposed to Div and md, which just output a single location. Whether a fixed or variable rates method (BTF vs. BTV or RBF vs. RBV) is preferable is an open question. Here we are using a simulation model where the rate of lexical change is probabilistically fixed, perhaps explaining why BTF works better than BTV. But in a real-life situation BTV might be better. Such qualitative factors can feed into the choice of a particular method, sometimes even excluding some methods. In the case of BEAST, the computing time may be prohibitive in a large study involving numerous language groups. Typically, cognate information may not be accessible for all the language groups even though word lists are available. This study suggests that BTF may be the preferred method when the cognate information required to infer a phylogenetic tree is available (such information may be produced manually or through an automated procedure [25]). When it is not, Div can be used, and its performance is not going to be radically inferior in spite of the more limited information that it draws upon. The surprisingly effective md baseline may serve as a quick way of generating a hypothesis about the location of a homeland, and in cases where no linguistic information is

available, only information about the location of languages, it is meaningful to apply this as an actual method.

## 6. Conclusion

Our results of testing different methods and baselines for inferring the geographical origins of language groups can be summarized in three points.

- When cognate information is available, BayesTraits, specifically the fixed rates version (BTF) is the preferred method.
- When cognate information is not available, the diversity method (Div), which requires neither a tree nor cognate information, may be applied, and its results will not be radically different from those of BayesTraits and largely indistinguishable in quality from those of RevBayes and BEAST.
- The minimal distance (md) baseline may have a real utility as a quick approximation to a hypothetical homeland, and is of particular utility when only language locations are known.

## Acknowledgments

SW's research was carried out under the auspices of the project "The Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage" (NWO proj. no. 335-54-102) within the European JPI Cultural Heritage and Global Change programme (http://jpi-ch.eu/). It was additionally funded by a subsidy from the Russian government to support the Programme of Competitive Development of Kazan Federal University and a major project from National Social Science Fund of China (no. 19ZDA300).

## References

1. Rama T. 2018. Three tree priors and five datasets: A study of Indo-European phylogenetics. *Lang. Dyn. Chang.* **8**, 182–218.
2. Rama T, Wichmann S. 2020. A test of Generalized Bayesian dating: A new linguistic dating method. *PLoS ONE* **15**, e0236522. (doi: 10.1371/journal.pone.0236522)
3. Sapir E. 1916. *Time Perspective in Aboriginal American Culture, a Study in Method*. Geological Survey Memoir 90.13. Anthropological Series. Ottawa: Government Printing Bureau.
4. Migliazza, EC, Campbell, L. 1988. *Panorama general de las lenguas indígenas en América*. *Historia general de América*, Vol. 10. Caracas: Instituto Panamericano de Geografía e Historia.
5. Blust, R. 1985. The Autronesian homeland: A linguistic perspective. *Asian Perspec.* **26**, 45–67.
6. Matisoff, JA. 1991. Sino-Tibetan linguistics: Present state and future prospects. *Annu. Rev. Anthropol.* **20**, 469-504.
7. Walker RS, Ribeiro LA. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proc. Royal Soc. B* **278**, 2562-2567. (doi: 10.1098/rspb.2010.2579)

8. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960. (doi:10.1126/science.1219669)
9. Holman EW, Wichmann S. 2017. New evidence from linguistic phylogenetics identifies limits to punctuational change. *Syst. Biol.* **66**, 604–610. (doi: 10.1093/sysbio/syw106)
10. Brown CH, Holman EW, Wichmann S. 2013. Sound correspondences in the world's languages. *Language* **89**, 4–29.
11. List J-M, Greenhill SJ, Gray RD. 2017. The potential of automatic word comparison for historical linguistics. *PLoS ONE* **12**, e0170046. (doi:10.1371/journal.pone.0170046)
12. Wichmann S. 2017. Modeling language family expansions. *Diachronica* **34**, 79–101. (doi: 10.1075/dia.34.1.03wic)
13. Hijmans RJ. 2019. geosphere: Spherical trigonometry. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere.
14. Wichmann S, Müller A, Velupillai V. 2010. Homelands of the world's language families: A quantitative approach. *Diachronica* **27**, 247–276. (doi: 10.1075/dia.27.2.05wic)
15. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu C-H, Xie D, Zhang C, Stadler T, Drummond AJ. 2019. BEAST2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650. (doi: 10.1371/journal.pcbi.1006650)
16. Pagel M., Meade A. 2019. BayesTraits V3.0.2. http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.2/BayesTraitsV3.0.2.html.
17. Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65, 726–36. (doi: 10.1093/sysbio/syw021)
18. Vavilov NI. 1926. Centers of origin of cultivated plants. *Trudi po Prikl. Bot. Genet. Selek.* **16**, 139–248.
19. Wichmann S, Holman EW, Bakker D, Brown CH. 2010. Evaluating linguistic distance measures. *Physica A* **389**, 3632–3639. (doi:10.1016/j.physa.2010.05.011)
20. Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn AP. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* **61**, 973–999. (doi: 10.1093/sysbio/sys058)
21. Venditti C, Meade A, Pagel M. 2011. Multiple routes to mammalian diversity. *Nature* **479**, 393–396. (doi: 10.1038/nature10516)
22. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885. (doi: 10.1093/molbev/msq067)

23. Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* **65**, 3578–3589. (doi: 10.1111/j.1558-5646.2011.01401.x)
24. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, p. E88. (doi: 10.1371/journal.pbio.0040088)
25. Rama, T, List, J-M, Wahle J, Jäger, G. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 393–400.
26. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi: 10.1093/sysbio/sys029)

## Supplementary material

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.12894218. It contains all programs and in- and output files used in this paper as well as explanatory notes.