Check for updates

RESEARCH ARTICLE

REVISED Gaussian process models for geographic controls in phylogenetic trees

[version 2; peer review: 1 approved, 3 approved with reservations, 1 not approved]

Frederik Hartmann[1], Gerhard Jäger[2]

[1]Department of Linguistics, University of North Texas, Denton, Texas, 7620, USA
[2]Seminar für Sprachwissenschaft, University of Tübingen, Tübingen, Baden-Württemberg, 72074, Germany

v2 **First published:** 17 Apr 2023, **3**:57
https://doi.org/10.12688/openreseurope.15490.1
Latest published: 22 Jan 2024, **3**:57
https://doi.org/10.12688/openreseurope.15490.2

Abstract

Geographical confounding in phylogenetic inference models has long been an issue. Often models have great difficulty detecting whether congruences or similarities between languages in phylogenetic datasets stem from common genetic descent or geographical proximity effects such as language contact. In this study, we introduce a distance-based Gaussian process approach with latent phylogenetic distances that can detect potential geographic contact zones and subsequently account for geospatial biases in the resulting tree topologies. We find that this approach is able to determine potential high-contact areas, making it possible to calculate the strength of this influence on both the tree-level (clade support) and the language-level (pairwise distances).

Keywords

Phylogenetics, Gaussian processes, geospatial confounding, control for language contact

This article is included in the European Research Council (ERC) gateway.

This article is included in the Horizon 2020 gateway.

Open Peer Review

Approval Status ? ✓ ? ? ✗

	1	2	3	4	5
version 2 (revision) 22 Jan 2024		✓ view	? view	? view	✗ view
version 1 17 Apr 2023	? view	? view			

1. **Tiago Tresoldi**, Uppsala University, Uppsala, Sweden
2. **Thomas Brochhagen**, Universitat Pompeu Fabra, Barcelona, Spain
3. **Frederic Blum**, Max Planck Institute for Evolutionary Anthropology, Leipzig,, Germany
   University of Passau, Passau, Germany
4. **Grégoire Clarté**, The University of Edinburgh, Edinburgh, UK
5. **Natalia Chousou-Polydouri**, University of Zurich,, Zurich, Switzerland
   Nico Neureiter, University of Zurich (Ringgold ID: 27217), Zurich, Switzerland

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Languages and Literature gateway.



This article is included in the Linguistic Diversity collection.

Corresponding author: Frederik Hartmann (frederikhartmann1@gmx.de)

Author roles: Hartmann F: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Jäger G**: Conceptualization, Data Curation, Funding Acquisition, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [834050]) and from the DFG Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Hartmann F and Jäger G. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hartmann F and Jäger G. **Gaussian process models for geographic controls in phylogenetic trees [version 2; peer review: 1 approved, 3 approved with reservations, 1 not approved]** Open Research Europe 2024, **3**:57 https://doi.org/10.12688/openreseurope.15490.2

First published: 17 Apr 2023, **3**:57 https://doi.org/10.12688/openreseurope.15490.1

REVISED **Amendments from Version 1**

The new version contains an extended investigation of the relationship between Stan chain convergence and clade credibility. It also contains a link to all model and data preprocessing files used for the analyses. Further, minor changes to wording and clarity have been applied.

Any further responses from the reviewers can be found at the end of the article

1 Introduction

Phylogenetic linguistics started with the aim to apply phylogenetic inference algorithms from computational biology to data based on collections of cross-linguistic lexical and grammatical data. The goal of this enterprise is to infer phylogenies, i.e., tree diagrams representing the diversification history of language families. These family trees are interesting for historical linguistics in their own right, but they also provide stepping stones for investigations into deep population history, statistical control for typological studies, and several other applications (see, *inter alia*, Bouckaert *et al.*, 2012; Bowern & Atkinson, 2012; Hruschka *et al.*, 2015; Jäger & Wahle, 2021).

A well-known potential problem of phylogenetic inference algorithms in this domain is that they model language change exclusively as vertical transmission with random mutations. However, historical linguists have been well aware since the 19th century (Schmidt, 1872) that horizontal transmission via language contact plays an important role in language change. Modeling the effects of language contact as independent random mutations potentially introduces a bias. Also, it ignores an important source of information about language history and concomittant population history processes.

This problem has not gone unnoticed. There are studies such as (Greenhill *et al.*, 2009) assessing the robustness of phylogenetic inference. A novel approach by Neureiter *et al.* (2022) enriches phylogenies with contact edges, i.e., connections between simulatenous points on different branches of a phylogeny enabling information transfer. The phylogenetic skeleton and the contact edges are inferred simultaneously.

This approach, however, ignores an important source of information, namely the fact that language contact almost always happens under spatial proximity. It seems therefore natural to model the effect of language contact as a spatial stochastic process. Spatial statistics have recently gained popularity in cross-linguistic studies, e.g., to identify linguistic areas or as a statistical control for typological studies (Guzmán Naranjo & Becker, 2021; Ranacher *et al.*, 2021). Phylogenetic information is either not used at all or only in a rudimentary fashion in this context though.

The present paper presents a hybrid approach combining spatial with phylogenetic modeling. Briefly put, the probability of a certain language L having a certain feature f is assumed to depend on both the presence or absence of f in L's spatial neighbors and in L's genetic relatives. Phylogenies are obtained via statistical inference while controlling for language contact. By comparing the results with vanilla phylogenetic inference, regions of intense language contact can be identified.

First, we present the model architecture and mechanics of the model before testing the model on datasets of known language families with known contact effects, namely *ASJP* (Wichmann *et al.*, 2016) in the modified form with cognate inferences produced by Jäger (2018) and *Northeuralex* (Dellert *et al.*, 2020).

2 Model architecture

The goal of this model is to infer phylogenetic similarity between languages based on data while geographical confounds are accounted for. This can be achieved by building a model that treats phylogenetic similarity as a latent variable that is inferred as the residual similarity of two languages when linguistic similarity based on geographic proximity is accounted for. Such a model structure can be graphically visualized as the graph in Figure 1.

In this model, the statistical model **M** takes in the geographical similarity **G** and the phylogenetic similarity P and models those on the linguistic data **D**. Crucially, P is an inferred (latent) node which is the estimand in this analysis. More concretely, the model is built with **G** and **D** as inputs, whereas P remains in variable form. As a result, information in the model flows from the linguistic data and the geographic similarity to the model which makes it possible to infer the phylogenetic similarity. The following model architecture was used for this analysis:

$$C_{s,l} \sim Binomial(1, p_{s,l})$$

$$logit(p_{s,l}) = \rho_{s,l} + \gamma_{s,l}$$

$$\begin{bmatrix} \gamma_{s,l_1} \\ \gamma_{s,l_2} \\ \dots \\ \gamma_{s,l_n} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}, G \right)$$

$$\begin{bmatrix} \rho_{s,l_1} \\ \rho_{s,l_2} \\ \dots \\ \rho_{s,l_n} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}, P \right)$$

$$G_{tj} = \alpha_G e^{-\frac{\delta^2_{G,tj}}{\lambda_G}}$$

$$P_{rk} = \alpha_P e^{-\frac{\delta_{P,rk}}{3}}$$

$$\delta_{P,rk} \sim Exponential(1)$$

$$\alpha_G, \alpha_P, \lambda_G \sim Exponential(2)$$

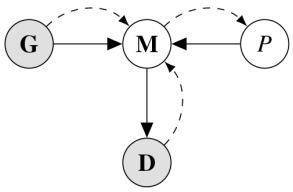

Figure 1. Graphical representation of the inference model. Grey nodes represent observed nodes (i.e. data) while white nodes are inferred. Black arrows indicate model-internal connections while dashed arrows show the structure of information flow in the model.

Here, the outcome C is a binary linguistic character of a certain language l and site s.[1] C is modelled as a binomial outcome with a probability of p for this specific site and language. ρ is the phylogenetic similarity effect to be inferred from the latent variable and γ is the geographic similarity effect inferred from the geographical data. Both variables are modelled as Gaussian processes, drawn from a multivariate normal distribution with a mean vector of 0 and the covariance matrices **G** and **P**. Their logit sum (of the variables ρ and γ) then enters the binomial process as p, effectively making $\rho_{s,l}$ and $\gamma_{s,l}$ the log-odds values of site s of language l being inherited or borrowed. They are further derived from the kernel functions that are, in case of the geographic similarity (G), based on a quadratic kernel (cf. McElreath, 2018, ch. 14.5) and the kernel of P is based on a Ornstein–Uhlenbeck kernel. The kernels take in the pairwise distances (phylogenetic or geographic) δ_P/δ_G between two languages and output the covariance matrix of all languages. The pairwise distances for the phylogenetic kernel are drawn as a latent variable from an exponential distribution with rate 1, while the parameters for the kernels are sampled from an exponential distribution with rate 2. This results in less prior probability being allocated to high maximum covariances and high covariance values for each language pair. In effect this means that the model assumes there to be little covariance between languages a priori.

Note that the Ornstein–Uhlenbeck kernel has a fixed parameter $\lambda = 3$. This follows a modelling decision that takes into account the latent property of the phylogenetic distances: were λ (the magnitude of the covariance between languages) inferred, there would be more than one solution for every covariance-distance pair. In practice, the distance function $e^{-\frac{\delta}{\lambda}}$ at $\delta = 1, \lambda = 1$ has the same solution (≈ 0.368) as at $\delta = 2, \lambda = 2$. If both parameters are inferred, for the model, both solutions are equivalent. In other words, if the model infers a covariance between two languages of 0.368, it is mathematically compatible with either $\delta = 1, \lambda = 1$ or $\delta = 2, \lambda = 2$.[2] This means that if both δ and λ are inferred parameters, there cannot be a single-valued solution for any value, leading to extreme collinearity between both parameters. As a result, we fixed λ such that this issue is avoided.

[1]The structure of the data is explained in Section 3.

[2]In fact, this is true for any real number where $\delta = \lambda$.

3 Data and contact area selection

As datasets on which to test the model, we used the databases *ASJP* (Wichmann *et al.*, 2016) with cognate inferences produced by Jäger (2018) and *Northeuralex* (Dellert *et al.*, 2020). Both datasets are large multi-language lexical cognate databases intended for use primarily for phylogenetic inference. *Northeuralex* was specifically selected since with Dellert (2019), there exists a detailed prior study of geographical contact effects on loanwords particularly in the Baltic Sea regions. Therefore, the results of the study at hand, and thus the performance of the model, can directly be compared to a prior study.

Fromboth datasets, a subset of languages (25 from *ASJP* and 24 from *Northeuralex*) were selected in such a way that known high-contact areas are reflected in the dataset on which the model's performance can be tested. Those areas are: contact between Celtic and English in the British Isles; the Balkan Sprachbund that includes South Slavic languages alongside Albanian, Romanian, and Greek; the Baltic Sea area with contact between Germanic, Slavic, Uralic, and Baltic; the contact zone between Indo-Iranian languages. As a condition for the Gaussian process approach to succeed, we expect the model to recognize these zones by increasing the geographical covariance between languages in these areas, thus increasing the likelihood for individual cognates to be borrowed between constituent languages.

The datasets were adapted to a format suited for the model architecture presented in Section 2: after conversion, the data consist of binary vectors where each vector represents a shared cognate between two or more languages. Table 1 illustrates this data shape.

Here, hypothetical languages *A*, *B*, and *C* share hypothetical cognate 1 whereas only languages *B* and *D* share cognate 2. Due to this structure, the model can calculate a probability for each language to have that cognate given the geographic and phylogenetic distances to all other languages.[3]

4 Results

We present the results of the analysis on three levels: firstly, we investigate the strength of the geographical confound between languages on a general level, before analyzing the effects the geographical control has on the tree topologies of the datasets. Lastly, we look at the results on a character level, comparing the model's inference for the likelihood of

Table 1.

	A	B	C	D
Cognate 1	1	1	1	0
Cognate 2	0	1	0	1

[3]The code to recreate the datasets as well as the code for the models can be found at 10.5281/zenodo.10247032.

individual borrowings in the *Northeuralex* dataset with the borrowings identified in Dellert (2019).

The model was run separately on each dataset so as not to add noise to the analysis due to potentially different coding schemes. Likewise, a 'confounded' model was run where the geographical variable was omitted. This allows for a direct comparison between the models with and without the geographical confound.

For modelling, we used the Bayesian programming language Stan (Stan Development Team, 2022). All results were based on posterior samples. Concretely, we extracted the posterior samples of the inferred pairwise phylogenetic distances between languages and constructed a UPGMA phylogram for each posterior sample. This yields a large number of trees from which clade support values were extracted and the consensus and maximum clade credibility (MCC) tree were constructed.

It was observed that the pairwise distances in the four-chain setup of the analysis did not achieve convergence. However, when analysing the individual chains, they yield the same tree topology. We found that this convergence issue is in part due to the fact that the same tree topology can be expressed with different distance matrices. For example, the following hypothetical upper-triangular distance matrices, M_1 and M_2, yield equivalent tree topologies between three hypothetical languages (not considering branch length).

$$M_1 = \begin{bmatrix} 1 & 2 \\ & 1 \end{bmatrix} M_2 = \begin{bmatrix} 2 & 4 \\ & 2 \end{bmatrix}$$

Here, matrix M_2 is a factor of M_1, meaning that the relative pairwise distances are equal. Since these distances are inferred as a latent variable in the model, each chain of the Bayesian model can converge at either M_1 or M_2, yielding the same tree topology but, by virtue of having converged on different values, inter-chain convergence is not reached. As a result, one needs to check model accuracy based on the tree topologies obtained from posterior samples directly. As described above, the diffent sampling chains of the models yielded tree topologies with minor differences in topology and clade support (see discussion in Section 6.3 in the appendix).

4.1 Evaluating the model

Before we can investigate the model results concerning the accuracy in identifying contact effects, we need to ascertain whether the model succeeds at identifying the major phylogenetic clades. Recall that the goal of this analysis is *not* to infer the correct tree topology of the given languages. Rather, we analyse the tree topologies here solely to check if the model is able to capture the clades well established in earlier research. The consensus trees constructed from the posterior samples of the model show that the model was able to capture the general topology of relationships between the languages in question (see Figure 2 and Figure 3). As reference trees for this comparison, we use the trees in Jäger (2018) and Dellert (2019) for the *ASJP* and *Notheuralex* databases respectively.

The models correctly identify several clades with high support. In the *ASJP* dataset, these include Slavic, Indo-Iranian, Romance, Germanic, and Celtic. In the *Northeuralex* dataset, they include Germanic (subdivided into North and West Germanic), Balto-Slavic, Slavic (split into East and West Slavic), Baltic, and Uralic (further divided into Finnic and Sámi). Further, in the *ASJP* dataset, Albanian is identified as an Indo-European outgroup, as is common in such analyses (cf. Ringe *et al.*, 2002, p. 90). However, the model was not able to retrieve more opaque relationships that are sometimes proposed such as Graeco-Armenian (Clackson, 1994; Jørgensen, 2022) or Italo-Celtic (see discussion in Eska, 2009; Weiss, 2022). Or, even if the status as a subgroup is unclear, frequently emerge as subgroups in phylogenetic analyses, e.g. in Ringe, Warnow, and Taylor (2002), Chang *et al.* (2015), and Heggarty *et al.* (2023).

The level of support that well-established clades exhibit in this model indicates that this inference technique was successful. Despite the drawbacks of distance-based methods (see Section 5), the method can be deemed accurate enough to continue analyzing the results.

4.2 Confound strength

Figure 4 and Figure 5 show the geographical confound strengths between each language pairing.[4] Those strengths are the normalized covariance values of the geographical covariance matrices. The strength line segments in the figures indicate the strength of the geographical confound. From the modelling perspective, the confound strength has to be interpreted as the geographically conditioned similarity of the linguistic data. In other words, if two languages exhibit a strong geographical signal, it means they are in part similar because of the geographical variable. Interpreted from the perspective of phylogenetic similarity, strong geographical covariance between two languages means that they are less similar phylogenetically than they seem, as some of the similarity can be attributed to geographical proximity.

Both maps show that strong geographical covariances are inferred for the Balkan region, northwest Europe, the northeastern Baltic Sea area and northern Scandinavia. This coincides with the previous assumptions about the linguistic contact zones. Further, strong pairwise covariances can be detected for Italian and Corsican, Czech and Polish, and Western and Eastern Armenian. The findings in Figure 5 are compatible with previous computational analyses of contact lines such as Dellert (2019, pp. 262–265) which show strong contact effects especially in the are of the eastern Baltic coastline and northern Scandinavia.

4.3 Clade-level analysis

For the tree-level analysis, we compare the posterior consensus trees of the model with and without the geographical variable. Going forward, the pure phylogenetic model is referred to as

[4]The same figures can be found in the appendix without the map data and with language labels.

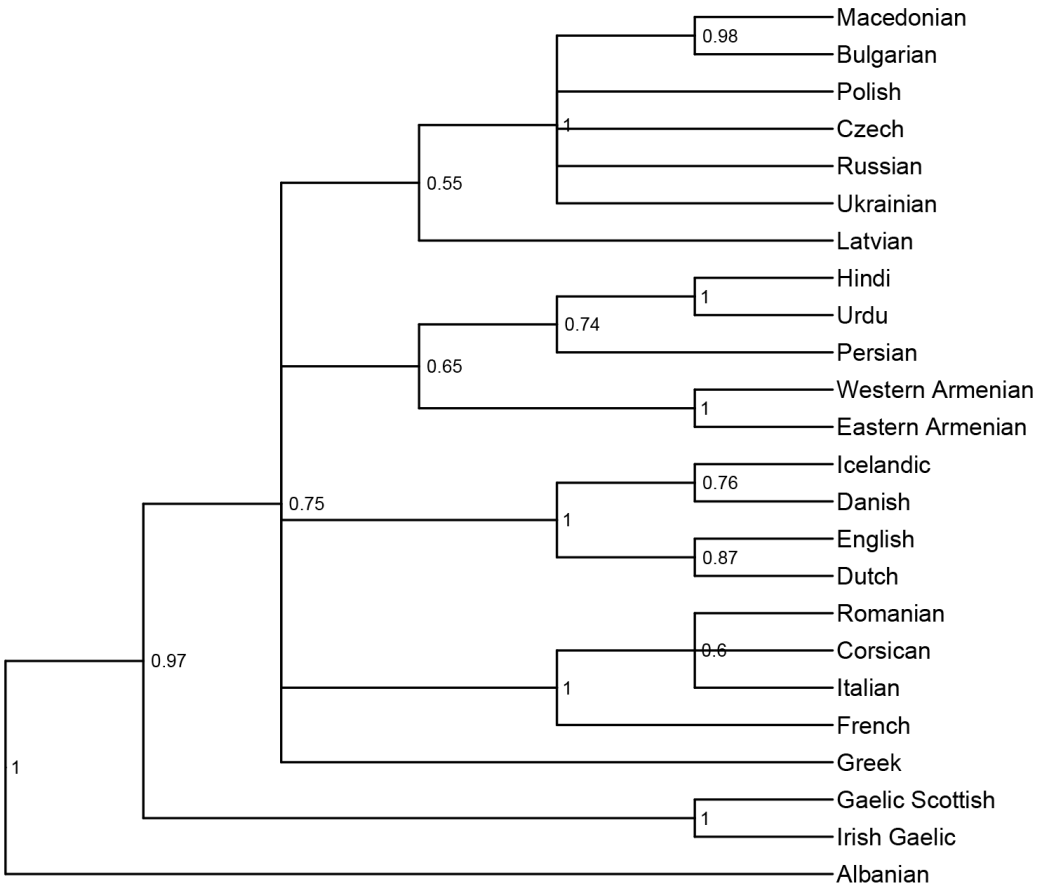

Figure 2. Consensus tree (cladogram) of the *ASJP* language similarity inference. The values at each internal node indicate the posterior clade support of that node. Clade support threshold for inclusion in the consensus tree was 0.5.

the *confounded* model while the *deconfounded* model refers to the model with geographical control.

Firstly, we investigate changes in the consensus trees between the two models. Figure 6 shows a direct comparison between the two consensus trees of the *ASJP* dataset.

In this tree comparison we can observe that Greek was moved from being clustered with Romance to a not clearly detemi-nable outgroup position within Indo-European, a reclassifica-tion that is more in line with the consensus view of Greek not being a Romance language. Further, Corsican is then clustered more closely with Italian instead of French.

However, not in line with the consensus view is that in the geographically controlled model, Polish and Czech are no longer clustered together, but are moved to an indeterminate position within Slavic. This suggests that the model might be over-correcting for geospatial effects in this instance (see discussion in Section 5).

The tree topologies of the *Northeuralex* dataset do not show differences between the confounded and deconfounded consensus trees. This does not mean that the geographic control did not affect the results, merely that the differences in clade support do not change a clade support from > 0.5 to < 0.5 or vice versa since the consensus trees show a clade whenever support is > 0.5.

As a next step, we compare the differences in tree support for individual clades between the confounded and deconfounded models on both *ASJP* and *Northeuralex* databases. Table 2 and Table 3 show the ten clades with the strongest differences (both positive and negative). A reduction in clade support when transitioning from the confounded to the deconfounded model indicates that adding the geographical control variable leads the model to interpret some of the similarities among the con-stituent languages as contact-induced rather than phylogenetic. Conversely, an increase in support indicates that the clade in question is phylogenetically more plausible based on the geographical analysis.

It has to be noted, however, that these changes in clade sup-port are not independent. In essence, an increase in clade support for one clade could be caused by a reduction in the competitor clade. For example, the increase in clade support for Italian, Corsican, and Romanian we observe in Table 2 might be caused by the fact that the model downweighs the clade consisting of Italian, French, and Romanian based on inferred contact strength. It is therefore not necessarily the

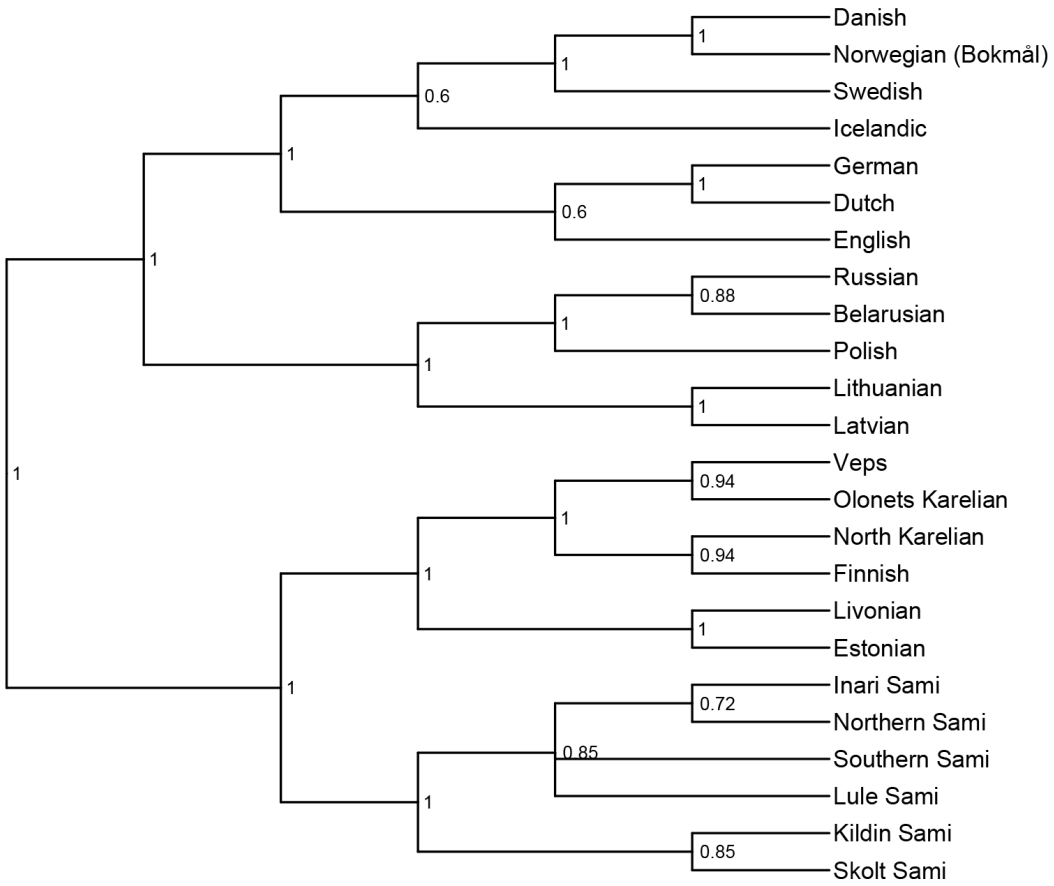

Figure 3. Consensus tree (cladogram) of the *Northeuralex* language similarity inference. The included language families are Indo-European and Uralic. The values at each internal node indicate the posterior clade support of that node. Clade support threshold for inclusion in the consensus tree was 0.5.

case that these three languages are more likely to constitute a clade based on geography, rather that the alternative hypothesis – Italian, French, and Romanian – has been deemed less likely.

In Table 2, we see that the largest reduction in clade support is in a clade consisting of the Romance and Germanic languages plus Greek which is strongly decreased due to geographical effects. As observed above, Czech and Polish, as well as a clade grouping Greek with the Romance languages, was reduced. Moreover, both the support for Indo-Iranian as well as a clade of Italian, French, and Romanian was decreased. This indicates that some of the similarities between those languages are partly geographically conditioned.

The results from the *Northeuralex* database showmany changes pertaining to Sámi, particularly identifying contact networks in Lule Sami, Southern Sami, and Northern Sami and Inari Sami, Skolt Sami, and Kildin Sami. Further, the North Germanic (Icelandic, Norwegian (Bokmål), Danish, and Swedish) andWest Germanic (English, Dutch, and German) clades receive less support in the deconfounded model.

4.4 Language-level analysis

In a second step, we can retrieve the pairwise phylogenetic distances of both models and compare the changes. Doing this enables the direct comparison of language pairs which decreased/increased their phylogenetic distance due to adding the geographical variable. Table 4 and Table 5 display the top 10 distances for each dataset, demonstrating the most significant changes.[5]

In Table 4, we find that, predominantly, distances involving Latvian and Albanian changed the most between the two models: the phylogenetic distance between Albanian and other languages in the Balkan Sprachbund were reduced while distances between other languages were increased. However, the increase in similarity between these languages calls for comment: according to the consensus tree (Figure 2), Albanian is still an outgroup even under the deconfounded model. The changes in the pairwise distances therefore mean that Albanian has been, relatively speaking, moved away in phylogenetic similarity from the other Balkan Sprachbund languages, which resulted in an inevitable decrease in distance from other languages. As can be seen in the consensus tree,

[5]Note that in order to make the distance matrices comparable, the pairwise distances were normalized such that the largest inferred distance corresponds to 1 in each model before calculating the differences between the models. A change in -0.34 means that from the confounded to the deconfounded model, the pairwise distance between the two languages in question shrunk by 34 percent of the largest phylogenetic distance.

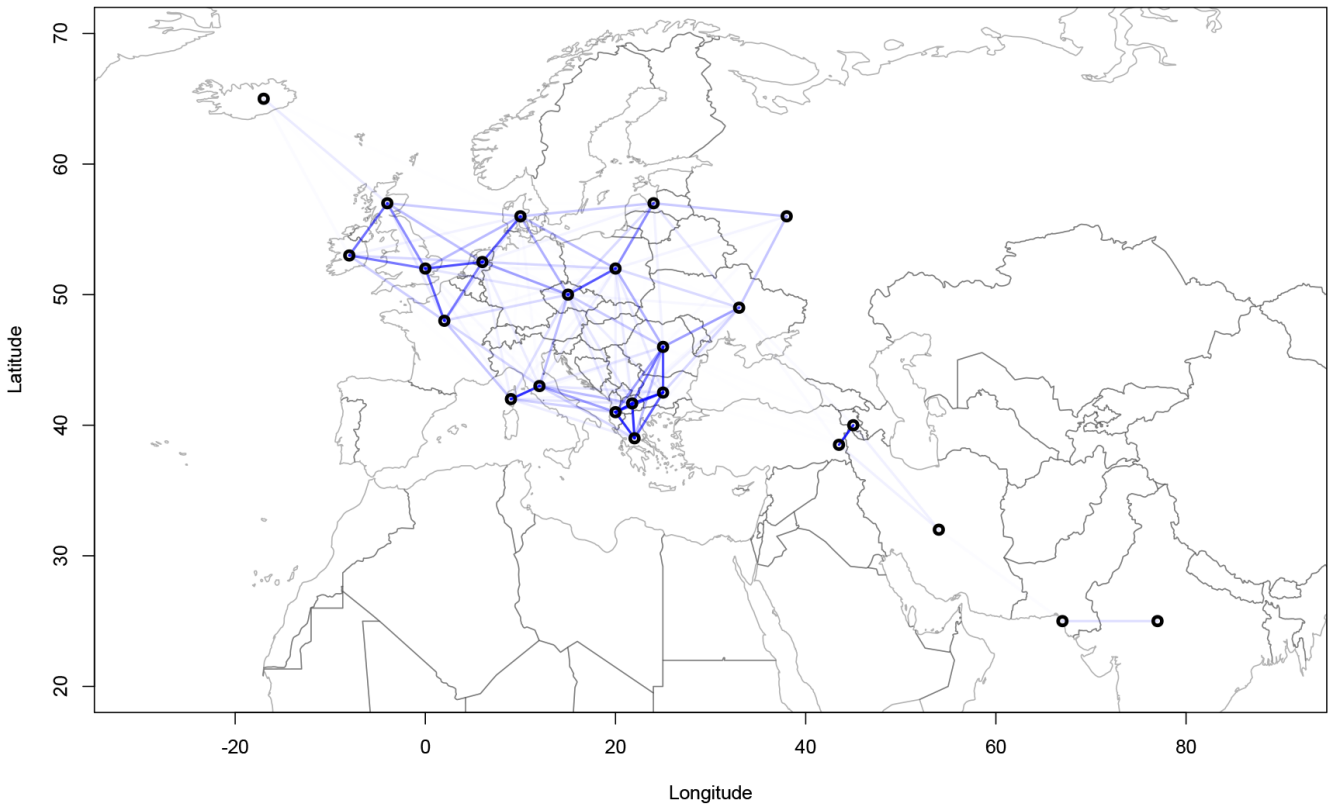

Figure 4. Strength of the geographical confounds between selected languages in the *ASJP* dataset.

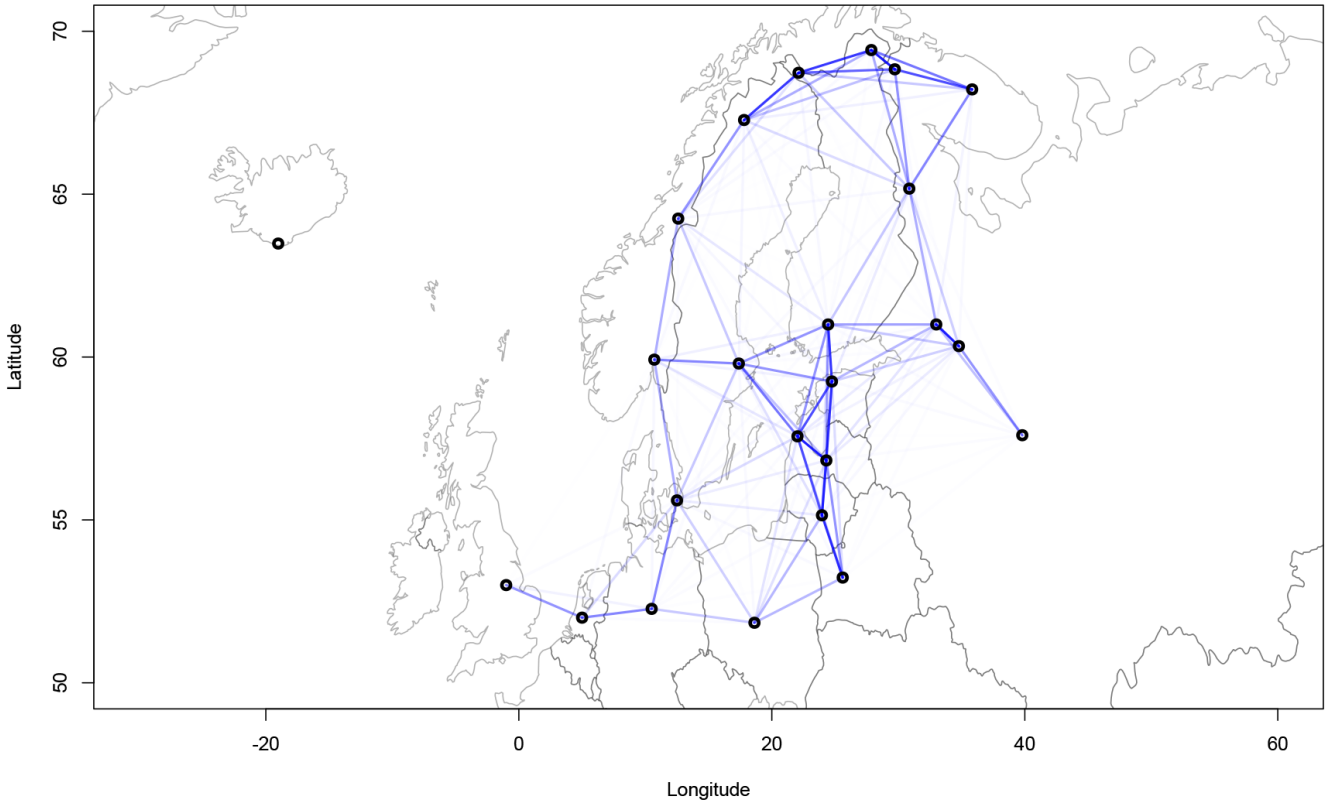

Figure 5. Strength of the geographical confounds between selected languages in the *Northeuralex* dataset.

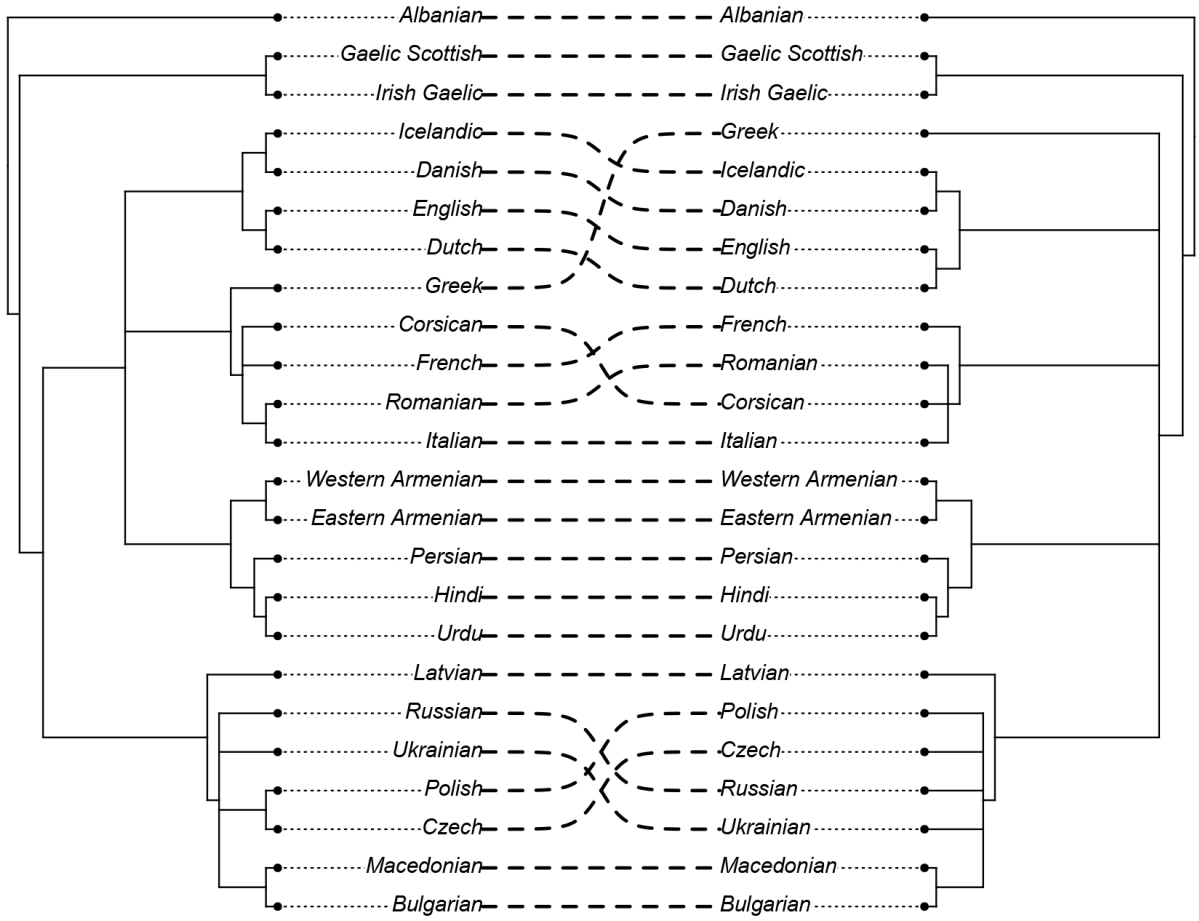

Figure 6. Comparison of the consensus trees of the *ASJP* dataset (created using the R-package *phytools* (Revell, 2012). *Left*: confounded tree. *Right*: deconfounded tree.

Table 2. Differences in clade support in the *ASJP* database between the confounded and the deconfounded models. Column *change* indicates the change in tree support when switching on the geographical control.

change	difference	clade constituents
0.43 → 0.16	-0.27	Greek, Italian, French, Corsican, Danish, Dutch, English, Icelandic, Romanian
0.60 → 0.39	-0.21	Czech, Polish
0.60 → 0.40	-0.20	Greek, Italian, French, Corsican, Romanian
0.34 → 0.15	-0.20	Italian, French, Romanian
0.92 → 0.74	-0.17	Urdu, Hindi, Persian
0.00 → 0.25	0.25	Latvian, Danish, Dutch, English, Icelandic
0.06 → 0.26	0.20	Urdu, Hindi, Persian, Italian, French, Corsican, Romanian
0.05 → 0.25	0.20	Czech, Polish, Bulgarian, Macedonian
0.44 → 0.60	0.17	Italian, Corsican, Romanian
0.07 → 0.23	0.16	French, Romanian

Table 3. Differences in clade support in the *Northeuralex* database between the confounded and the deconfounded models. Column *change* indicates the change in tree support when switching on the geographical control.

change	difference	clade constituents
1.00 → 0.60	-0.40	English, Dutch, German
1.00 → 0.60	-0.40	Icelandic, Norwegian (Bokmål), Danish, Swedish
0.26 → 0.15	-0.11	Inari Sami, Skolt Sami
0.26 → 0.15	-0.11	Inari Sami, Skolt Sami, Kildin Sami
0.26 → 0.15	-0.11	Lule Sami, Southern Sami, Northern Sami
0.72 → 0.85	0.13	Lule Sami, Southern Sami, Northern Sami, Inari Sami
0.74 → 0.85	0.11	Skolt Sami, Kildin Sami
0.33 → 0.41	0.08	Lule Sami, Southern Sami
0.66 → 0.72	0.06	Northern Sami, Inari Sami
0.83 → 0.88	0.05	Belarusian, Russian

Table 4. Top ten changes in pairwise phylogenetic distance in the *ASJP* dataset between the confounded and deconfounded models.

difference (norm.)	Language pair
-0.34	Macedonian, Albanian
-0.28	Macedonian, Latvian
-0.26	Bulgarian, Albanian
-0.20	Russian, Latvian
-0.19	Czech, Latvian
0.56	Albanian, Greek
0.47	Albanian, Persian
0.46	Albanian, Eastern Armenian
0.36	Corsican, Albanian
0.35	Albanian, Western Armenian

Table 5. Top ten changes in pairwise phylogenetic distance in the *Northeuralex* dataset between the confounded and deconfounded models.

difference (norm.)	Language pair
-0.10	North Karelian, Swedish
-0.09	Olonets Karelian, Swedish
-0.07	North Karelian, Norwegian (Bokmål)
-0.07	Finnish, Swedish
-0.04	Veps, Swedish
0.23	Kildin Sami, Swedish
0.21	Polish, Finnish
0.21	Belarusian, Finnish
0.19	Russian, Finnish
0.19	Northern Sami, Danish

this does not mean that Albanian is therefore particularly close to e.g. Persian, just that a movement away from geographically close languages results in an overall decrease in distance from other languages.

Further, Latvian shows decreased similarity with both western and eastern Slavic languages.

In the *Northeuralex* dataset, we find weaker changes overall; most reduction in phylogenetic similarity is found in pairwise distances between Germanic and Sámi languages on the Scandinavian Peninsula. Conversely, the model reports increases in phylogenetic similarity for Finnish, and some Indo-European languages, which goes against the common consensus. The reason for this result may be that the model overcompensates for Finnish showing geographical effects in the Uralic context by moving it closer to other languages instead.

4.5 Character-level analysis

Recall that themodel infers the the geographical and phylogenetic distances between languages on the character-level, which means that for every character, the model calculates a corresponding log-likelihood value for the character to be geographically conditioned (γ in the model formula in Section 2). High values of this parameter for a certain character indicate that it has a strong geographical conditioning; i.e., the character is likely to be borrowed. If the model succeeds at finding geographical contact patterns, we would expect to find that the actual loanwords we see in the languages in question correspond to high values of γ in the model results.

To evaluate the character-level accuracy of the inferences, we relied on the goldstandard loanword dataset provided in the supplements of Dellert (2019), which is itself a modification of the World Loanword Database (WOLD) (Haspelmath & Tadmor, 2009). This enables the direct comparison between

the *Northeuralex* dataset model output and the loanword database. For this analysis, we selected all loanwords where the source and target languages are present in the model, which yields a total of 82 borrowed characters. Figure 7 shows the distributions of log-likelihood values for the geographical variable for each character by whether or not they are indicated as loans in Dellert (2019).

The figure shows that the borrowed characters are attributed a larger log-likelihood in the model. To further corroborate the difference between the groups, we ran a univariate Bayesian regression model with standard normal priors using the Rpackage *brms* (Bürkner, 2017). In this model, the log-likelihood was the outcome variable and the binary variable of the borrowings was the predictor. The results show a detectable difference between borrowings = 'yes' (mean 0.63, 95% CI [0.42, 0.84]) and borrowings = 'no' (mean -0.40, 95% CI [-0.41, -0.40]). Therefore we can conclude that the model succeeds in assigning borrowed words a higher loglikelihood. However, the difference is only present in population means. Concretely, although the model assigns borrowings a higher geographical conditioning on the population-level, it fails at clearly identifying individual characters as borrowings. An accurate classification of characters could therefore not be done. This is crucial for evaluating the usefulness of the model insofar as the model can not be used for identifying loanwords for this reason.

5 Advantages, biases and limitations of the approach

After evaluating the results on several levels of analysis, we now turn to a more in-depth scrutiny of the implications and usefulness of the model for phylogenetic research.

5.1 Summary of the analysis

The analysis has shown that the Gaussian process model approach yields mostly accurate inferences of contact zones and their effects on phylogenetic inference. The model succeeds in many respects at deconfounding phylogenetic distances between languages. However, the model seems to overestimate certain confounds especially with languages that are both highly phylogenetically related and geographically close (see discussion in Section 5.2).

Overall, however, the method provides insights at different levels of analysis, which is useful to investigate micro-level effects such as pairwise distances as well as macro-level effects on a tree topology or individual clades. Despite this, the granularity is not infinitely scalable as the model cannot be used for loanword classifications (see discussion in section 4.5).

Moreover, it has been noticed that due to the strong correlation between reductions and increases in clade support or pairwise language distances, it is not always discernible whether an increase in tree support or language distance for a clade or pair of languages is due to a reduction in another part of the tree.

5.2 Biases and limitations that reduce the efficacy of the model

Several issues limit the accuracy of the approach; some of which are methodological limitations due to how the model is set up, and some of which are biasing factors that interfere with the model's inferences. This section is intended to sketch out these issues along with some considerations about how to address them.

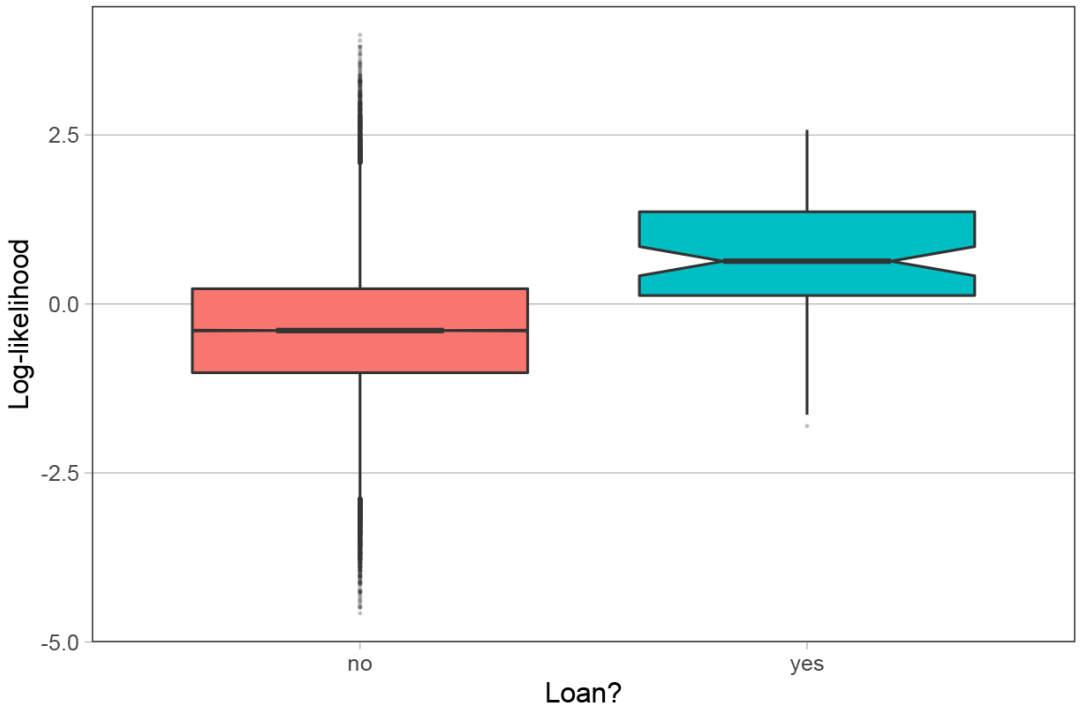

Figure 7. Distribution of per-character geographical log-likelihood by whether or not they are loanwords.

First of all, the most obvious limitation of this approach is that it relies on a distance-based architecture rather than inferring character evolution as most state-of-the-art phylogenetic models do. This is an issue that cannot be overcome without entirely re-designing the model under an evolutionary model paradigm: the Gaussian processes used in this study are inherently distance-based as they take in information about geographical and phylogenetic relationships in the form of pairwise distances. Related to this point, the model cannot capture historical contact lines as accurately as the geospatial distance matrices used in these analyses are stationary across time can. In effect, this approach does not model changes in contact strength over time as would be possible in a character-evolution model.

A related limitation that encompasses several sub-issues is the inability of singular Gaussian process kernels to handle non-positive-definite matrices. This means that each one of the categories (languages in this case) needs to be proportionally distant from any other category. This means that if there are three languages *A*, *B*, and *C*, *A* cannot be at the same time close to *B* but different from *C* without *B* also being different from *C*. Consider the two distance matrices D_1 and D_2.

$$D_1 = \begin{bmatrix} & 1 & 4 \\ 1 & & 3 \\ 4 & 3 & \end{bmatrix} \quad D_2 = \begin{bmatrix} & 1 & 4 \\ 1 & & 1 \\ 4 & 1 & \end{bmatrix}$$

Here, D_1 yields a positive definite matrix outcome of the GP kernel while D_2 does not. The reason for this is that D_2 is not representable in 2D Euclidean space since the distances between the points are not proportional.[6]

Why is this issue important in linguistic datasets? Geospatial (or, for that matter, all contact-induced) relationships between languages cannot be accurately be represented as points in Euclidean space. That is, some languages have several contact points with other languages or nonlinear contact relationships. For example, French is in close contact with both Flemish and Spanish, hence the pairwise distances between French and Flemish and French and Spanish would both be very small while the distances between Flemish and Spanish would be much greater. Further, there can be long-distance or nonlinear contacts such as contacts between English and Spanish in the Americas or French and Arabic in Northern Africa. Additionally, there are contact relationships that are non-geographic, such as Latin influence on several central and western European languages in the middle ages. One could argue, however, that a point-like representation is a sufficiently informative abstraction, yet in such models, we cannot rule out potential mismatches between the Euclidean representations and the in reality observable contacts.

This modelling issue could be addressed, however, by giving each language a covariance matrix of its own that is inferred separately at each step based on a custom distance matrix outfitted with bespoke distances, based on topology and actual contact lines. However, this would increase computation times by a factor equal to the number of languages (i.e. for the approach at hand this would mean a computation time increase of a factor of 25). Even with a moderate number of languages in the dataset (10–30), this would result in computation times of several weeks, even with fully optimized code.

[6]This, however, is only a cursory explanation as the mathematical reason why the posititve definiteness is not fulfilled is grounded in the definition of the Gaussian process kernel.

6 Conclusion

We have shown in this study that the proposed Gaussian process model can identify areas, clades and language pairs that have been subject to geospatial effects, making them seem genetically closer than they are. This approach succeeds at partially removing those geospatial confounds, thereby mitigating the risk of mistaking contact-induced similarity between languages as support for closer genetic relationships.

While this model is effective overall, we identified some weaknesses that are due to how Gaussian processes in general handle distance matrices and the known issue that languages are represented as point-like units in such models. Further, we found that the character-based loanword identification failed in the model, meaning that the model in its current form cannot be used to identify borrowings in the dataset, as the model accuracy is too low on such a fine-grained level. We identified problems with chain convergence in the model stemming from the Gaussian process setup and the large number of latent parameters. However, these issues can be mitigated by running more chains and averaging over the posterior samples.

Acknowledgements

Many thanks to Johannes Dellert and Johanna Nichols for their input and comments on the project. Many thanks also to Miri Mertner and Molly Rolf for commenting on the final draft. This research was supported by the DFG Centre for Advanced Studies in the Humanities *Words, Bones, Genes, Tools* (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050).

Appendix

6.1 List of languages included

Languages selected from the *ASJP* database:

> Hindi, Urdu, Persian,Western Armenian, Eastern Armenian, Romanian, Italian, Corsican, French, Greek, English, Dutch, Icelandic, Danish, Polish, Czech, Macedonian, Bulgarian, Russian, Ukrainian, Latvian, Scottish Gaelic, Irish Gaelic, Albanian

Languages selected from the *Northeuralex* database:

> Danish, Norwegian, Swedish, Icelandic, German, Dutch, English, Russian, Belarusian, Polish, Lithuanian, Latvian, Veps, Olonets Karelian, North Karelian, Finnish, Livonian, Estonian, Inari Sami, Northern Sami, Southern Sami, Lule Sami, Kildin Sami, Skolt Sami

6.2 Topologies of confounded trees

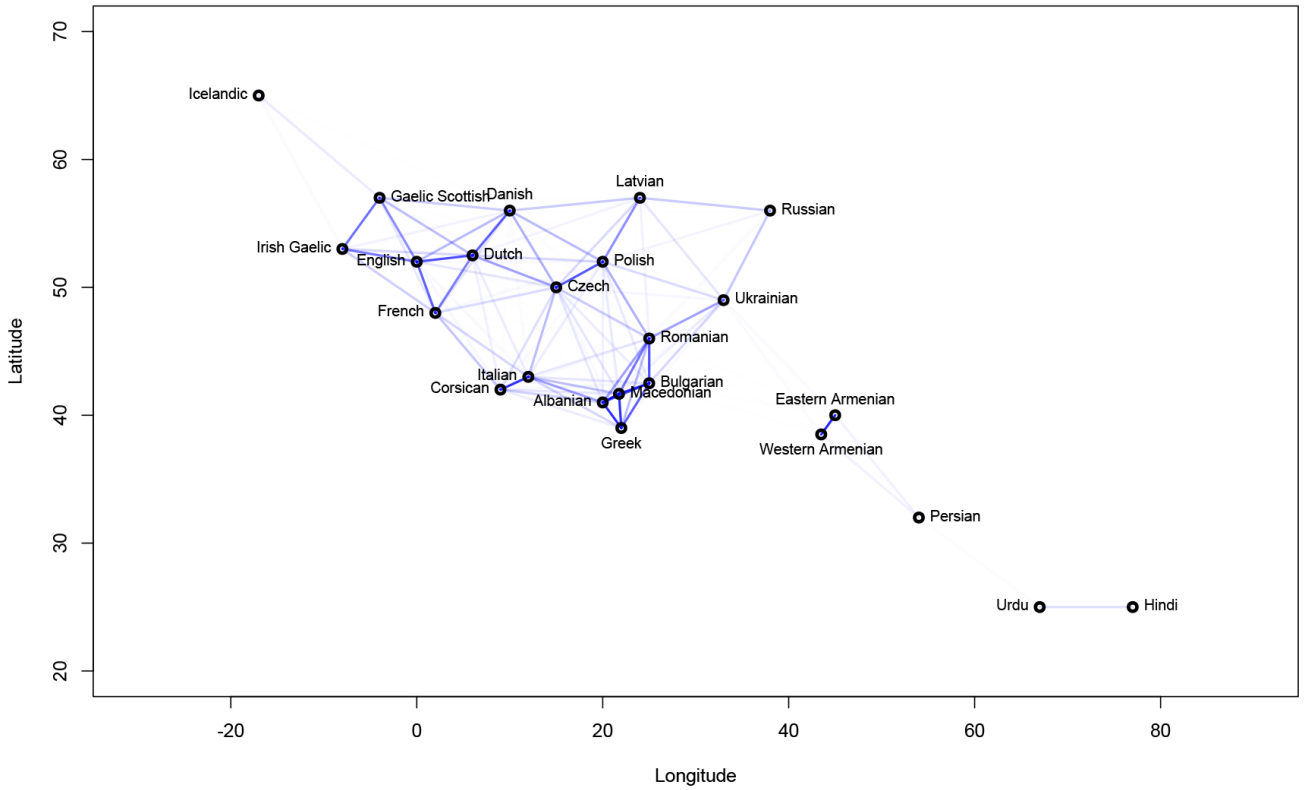

Figure 8. Strength of the geographical confounds between selected languages in the *ASJP* dataset, map data removed.

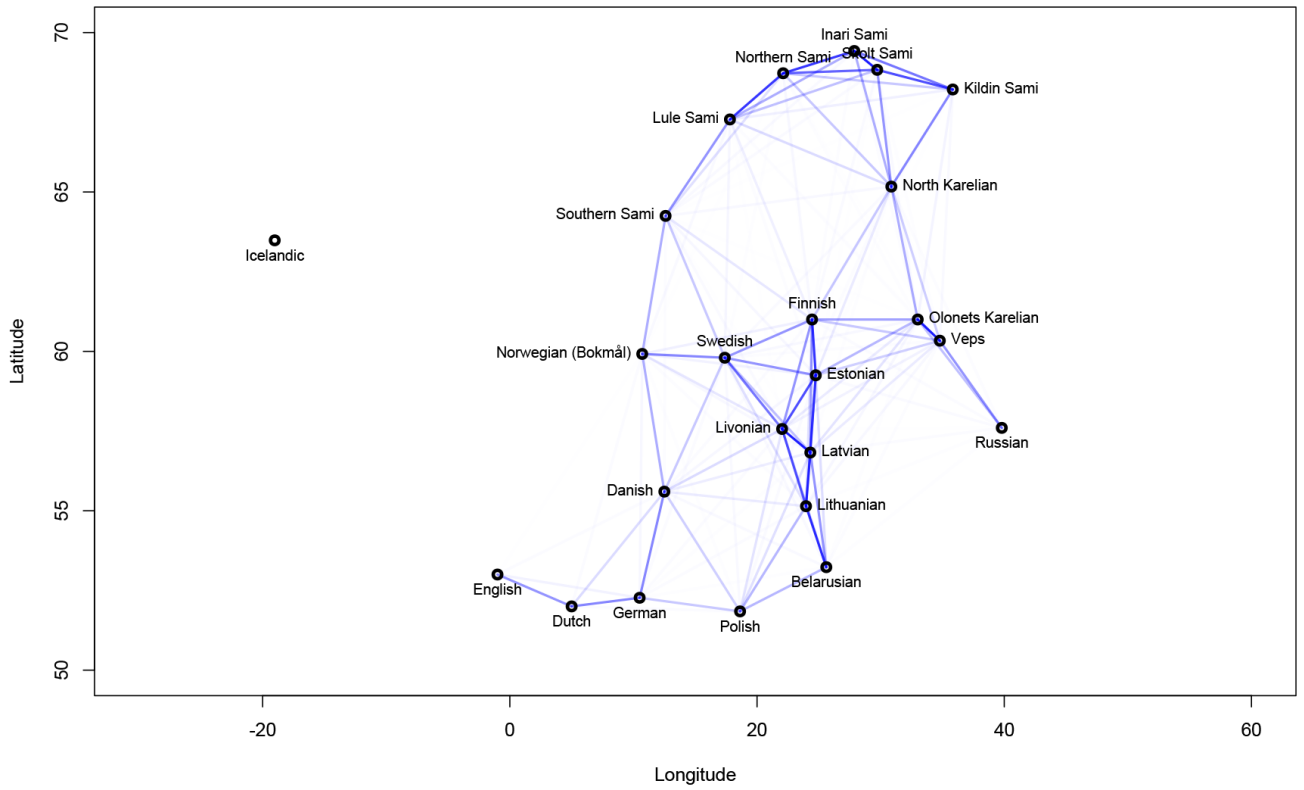

Figure 9. Strength of the geographical confounds between selected languages in the *Northeuralex* dataset, map data removed.

6.3 A note on convergence

As discussed in Section 4, some convergence issues were noticed that can partially be explained by the nature of inferring distance matrices whose multiples of each other yield the same tree topology. Nevertheless we found further some convergence issues that resulted in disagreements in the topologies of the consensus trees of each chain. Those disagreements were chiefly found in the sub-structure of the Germanic clade and the clade support values of East Slavic and Sámi in the *Northeuralex* dataset, and Germanic, Romance, and Slavic in the *ASJP* dataset.

We calculated that the average standard deviation of node support of values in the consensus trees is 0.016 (*ASJP*) and 0.031 (*Northeuralex*) while the same measure for clades with support higher than ten percent is 0.077 (*ASJP*) and 0.069 (*Northeuralex*).

This means that the chains mostly disagree in low-support clades which can be interpreted that individual chains tend to rate some low-support nodes higher than others while the general tree structure is in agreement between the chains. Nevertheless, the node support standard deviation measures are high enough to be a potential issue for this method and warrants further scrutiny. To analyze further how convergence issues arise in this approach, we use the following experimental design: the investigation proceeds in a grid search simulation-based fashion that simulates phylogenies and geographic distances of languages of varying length and number of observed sites. Then, the model presented in the main approach is run on each simulated dataset. Further, we aim to investigate if covariance between phylogenetic and geographic distance is responsible for convergence issues in the model which is why we add a covariance term to the grid search (see below). Each simulated run goes through the following steps:

1. **Obtaining a phylogeny and simulated character set**: A randomtree topology is generated with exclusively extant taxa with the number of taxa specified by the associated parameter of the simulation. Taking this topology, a data generation function simulates the evolution of a binary character (‘0’ and ‘1’ with ‘0’ being the character at the root) by selecting one edge in the topology whose descendant nodes receive the innovated character 1. The edge that is selected is determined by its edge length where the probability of an innovation along this edge is proportional to its length. This simulates a character evolution process with a uniform likelihood of a change in state across all lineages. Since exactly innovation occurs in the tree over the course of the evolution process, it means that homoplastic events do not occur in this process. To simulate loss of the trait, each node and its descendants is given a probability of switching to a ‘0’ with a probability of 10%. As a result, we simulate a uniform-rate character evolution with one innovation per site and a loss rate of 10%. This process is repeated for every site specified by the number of sites parameter of the simulation.

2. **Obtaining a geographic distance**: For each taxon, we generate a random point on a quadratic two dimensional surface with bounds 0 and 3. Afterwards, we calculate the euclidean distances between each point and obtain a geographic distance matrix that can be fed to the model. To track the effects of potential covariance of phylogenetic and geographic distance on convergence, we run the models both with the aforementioned randomly generated geographic distance and a correlated geographic distance. For this, we calculate the cophenetic distance between the extant taxa in the simulated tree and multiply it by 1.2. The resulting matrix is a multiple of the phylogenetic distance in the tree. This leads the correlation between phylogenetic distance and the geographic distance to be 1 which is the worst-case scenario where geographic distance is entirely predictable from phylogenetic distance. We opted for this worst-case implementation (although this correlation strength is unrealistic in real-world cases) since, if covariance plays a role in chain convergence, we can gauge its maximal effect. Thus, the effect we see with the maximally covariant matrices is the strongest it can be.

3. **Running the model**: The model was run on the simulated dataset and geographical distance.

4. **Obtaining split frequencies and convergence metrics**: After the model run, the average *r_hat* was extracted for all parameters and only the inferred phylogenetic distance matrix separately to investigate if there are differences between overall convergence and convergence of the phylogenetic distances. Further, the posterior phylogenetic distance matrices of each sample were used to construct a tree for each of those samples, for which the split frequencies of each individual split in the trees were calculated. This was done for each chain separately to obtain the mean standard deviation of the split frequencies for each one of the chains. This was done both for the clades that had majority support and for all clades to be able to distinguish convergence in majoritysupported clades.

The grid search parameters are tabulated in Table 6.

Table 6. Parameters used in the grid search.

Parameter	Grid intervals
Number of languages	5,10,20
Number of total sites	100, 300, 500
Covariance geo-phylo	‘yes’, ‘no’

Each grid search parameter combination was run 4 times in total to capture variances between the runs with the same parameter settings. The total number of data points are 72.

The results of the grid search show three properties: firstly, none of the runs with correlated distance matrices have supported clades included in their consensus trees. This means that at extreme covariance between geographic and phylogenetic distance, the model cannot distinguish between the two effects. This is expected since in the simulation, there is no residual information from the geographic distance that can be inferred. Secondly, the correlation between the *r_hat* values of all parameters and the *r_hat* values only those of the phylogenetic distance matrix is 0.99 which means that there is no difference in convergence of just the phylogenetic distance parameters and the overall model convergence. Moreover, the mean standard deviation of the split frequencies in the majority clades is on average 0.04 (72 %) standard deviations less than that of all clades taken together. This is considerable and means that the majority clades show much better chain convergence.

To further analyze the results, we used linear models to determine the effects of each factor on the different convergence metrics. Table 7 shows the results of the results.

As Table 7 shows, it is the the number of languages in the tree that primarily causes convergence issues. This is the case both for the *r_hat* convergence metric and the split frequencies convergence metric. This effect, however, does not appear as strong in the split frequencies of only the majority clades, which indicates that even with convergence issues, the results of the posterior trees concerning majority clades are less impacted by this.

To summarize, convergence issues are related to the following factors:

- Convergence of posterior trees are related to the **number of languages** in the sample. That is, the more languages are in the dataset, the more posterior chains fail to converge. This can be explained by the fact that with increased number of languages, the pairwise distances increase to be inferred as parameters drastically increase by $\frac{n!}{2!(n-2)!}$, which results in 190 latent parameters for 20 languages and 1,225 for 50 languages. Each of these parameters could potentially be responsible for divergences in the tree topology.
- The disagreements were mostly contained to those clades where there is **considerable uncertainty in the tree structure**, i.e. where small deviations in pairwise distances can result in a change in tree topology.
- The disagreements between the chains seem to **average out** with increasing number of chains. This can be observed by the fact that disagreements in lowsupport languages lead to an increase in uncertainty in the average posterior and therefore have reduced impact on the consensus tree. In other words, with increasing number of chains, the risk of false positive inferences is reduced. In potential future applications of this method, it might therefore be advisable to run more than four chains to optimally harness the effect of this averaging.

Table 7. Model coefficients (rows) of different linear models with different outcomes (columns). Boldface figures indicate significance of the predictor of p<0.05.

outcome:predictor	total *r_hat*	split freq.	split freq. (majority)
No. languages	**0.09**	**0.01**	0.00
No. sites	0.00	-0.00	-0.00
covariance (yes)	-0.10	-0.02	
No. languages:No. sites	-0.00	0.00	-0.00
No. languages:covariance (yes)	0.00	-0.00	
No. sites:covariance (yes)	-0.00	0.00	

References

Bouckaert R, Lemey P, Dunn M, *et al.*: **Mapping the origins and expansion of the Indo-European language family.** *Science.* 2012; **337**(6097): 957–960.
PubMed Abstract | Publisher Full Text | Free Full Text

Bowern C, Atkinson Q: **Computational phylogenetics and the internal structure of Pama-Nyungan.** *Language.* 2012; **88**(4): 817–845.
Reference Source

Bürkner PC: **Advanced Bayesian multilevel modeling with theR package brms.** *arXiv preprint.* 2017; arXiv preprint arXiv:1705.11123. 2017.
Publisher Full Text

Chang W, Cathcart C, Hall D, *et al.*: **Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis.** *Language.* 2015; **91**(1): 194–244.
Publisher Full Text

Clackson J: **The linguistic relationship between Armenian and Greek.** Oxford: Blackwell, 1994.

Dellert J: **Information-theoretic causal inference of lexical flow.** Berlin: Language Science Press, 2019.
Publisher Full Text

Dellert J, Daneyko T, Münch A, *et al.*: **NorthEuraLex: A wide-coverage lexical database of Northern Eurasia.** *Lang Resour Eval.* 2020; **54**(1): 273–301.
PubMed Abstract | Publisher Full Text | Free Full Text

Eska JF: **The celtic languages.***The Celtic Languages.* Ed. byMartin Ball and Nicole Müller. New York: Routledge, 2009; 36–41.

Greenhill SJ, Currie TE, Gray RD: **Does horizontal transmission invalidate cultural phylogenies?** *Proc Biol Sci.* 2009; **276**(1665): 2299–2306.
PubMed Abstract | Publisher Full Text | Free Full Text

Guzmán Naranjo M, Becker L: **Statistical bias control intypology.** *Linguist Typol.* 2021; **26**(3): 605–670.
Publisher Full Text

Haspelmath M, Tadmor U: **WOLD.** Leipzig: Max PlanckInstitute for Evolutionary Anthropology, 2009.
Reference Source

Heggarty P, Anderson C, Scarborough M, *et al.*: **Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages.** *Science.* 2023; **381**(6656): eabg0818.
PubMed Abstract | **Publisher Full Text**

Hruschka DJ, Branford S, Smith ED, *et al.*: **Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution.** *Curr Biol.* 2015; **25**(1): 1–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jäger G: **Global-scale phylogenetic linguistic inference from lexical resources.** *Sci Data.* 2018; **5**(1): 180189.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jäger G, Wahle J: **Phylogenetic Typology.** *Front Psychol.* 2021; **12**: 682132.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jørgensen AR: **Celtic.** In: *The Indo-European Language Family: A Phylogenetic Perspective.* Ed. by ThomasEditor Olander. Cambridge University Press, 2022; 135–151.
Publisher Full Text

McElreath R: **Statistical rethinking: a Bayesian course with examplesin R and Stan.** Boca Raton: Chapman and Hall/CRC, 2018.

Neureiter N, Ranacher P, Efrat-Kowalsky N, *et al.*: **Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer.** *Humanit Soc Sci Commun.* 2022; **9**: 1–14.
Publisher Full Text

Ranacher P, Neureiter N, van Gijn R, *et al.*: **Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact.** *J R Soc Interface.* 2021; **18**(181): 20201031.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Revell LJ: **phytools: An R package for phylogenetic comparative biology(and other things).** *Methods Ecol Evol.* 2012; **3**(2): 217–223.
Publisher Full Text

Ringe D, Warnow T, Taylor A: **Indo-European and computationalcladistics.** *Trans Philol Soc.* 2002; **100**(1): 59–129.
Publisher Full Text

Schmidt J: **Die Verwantschaftsverhältnisse der indogermanischen Sprachen.** H. Böhlau, 1872.
Reference Source

Stan Development Team: **RStan: the R interface to Stan.** R package version 2.21.7. 2022.
Reference Source

Weiss M: **Italo-Celtic.** In: *The Indo-European Language Family: APhylogenetic Perspective.* Ed. by ThomasEditor Olander. Cambridge University Press, 2022; 102–113.
Publisher Full Text

Wichmann S, Holman EW, Brown CH: **The ASJP Database. (version 17).** 2016.
Reference Source

Open Peer Review

Current Peer Review Status: ? ✓ ? ? ✗

Version 2

Reviewer Report 21 February 2025

https://doi.org/

© **2025 Chousou-Polydouri N et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✗ **Natalia Chousou-Polydouri**
University of Zurich,, Zurich, Switzerland
Nico Neureiter
Department of Geography, University of Zurich (Ringgold ID: 27217), Zurich, Zurich, Switzerland

The article introduces a method for detecting contact areas, which can also be used to control for contact effects when inferring phylogenetic relationships. The subject of the article is very interesting and it addresses a pressing need in both contact and phylogenetic studies. We find though that the goal of the paper is not clear, and the methods are not sufficiently described and motivated. Unless we are missing a key point, we think that the method's ability to detect contact areas is fundamentally flawed. At the same time, the method's utility in improving phylogenetic inference (through correcting for contact effects) is doubtful for a variety of reasons, many of which are already acknowledged in the paper.
We still want to commend the authors for approaching this important problem in a novel way. Gaussian processes provide a flexible way to explain data probabilistically and they can be naturally generalized to a mixture of multiple processes (inheritance and contact). The fact that they can be implemented in existing probabilistic programming languages like Stan is another advantage. In light of this, we hope that an adaptation of the model can resolve the issues brought up in this review. However, we cannot approve the article in its current form.

Geographic part of the model - confound strength
The most crucial point of our review concerns the geographic part of the model and the calculation of the confound strength. If we are not missing something, the geographic part of the model constrains the covariance matrix G to be a transformation of the geographic distances. Since confound strength values are normalized entries of G, they would also be a scaled version of the geographic distance. Since G and G are fixed across languages, Gtj=G epx(-G,tjG) is a fixed and continuously decreasing function of G,tj. It is not clear if and how the linguistic data are reflected here. Looking at the maps included in the paper reinforces the interpretation above. Darker edges reflect shorter distances and we could not find a counterexample.
If the above understanding of the geographic aspect of the model is correct, it would unfortunately make any interpretation of the confound strength essentially meaningless.

Resolving this issue would require changing the model (e.g. estimate G as a latent variable in Stan, and only use geographic distances as a prior).

Goal of the paper/method
It is not clear how the method is proposed to be used. In the introduction, the focus seems to be on improving phylogenetic inference by controlling for contact which is treated as just a confounder. However, later, the focus shifts to detecting contact areas, languages pairs that have participated in contact events and even characters that are plausibly affected by contact.
As we explained above, we are not convinced that the proposed method can be used to detect contact. It could be argued though that the current methodology could be used to calculate a contact proxy in order to improve phylogenetic inference, in line with the goal of the paper as stated in the beginning. In this case though, the phylogenetic methodology chosen is puzzling. UPGMA, apart from being a distance-based method, performs usually worse than Neighbor-Joining and assumes a strict clock, an assumption that is most often violated in linguistic evolution.

Description of methods
The paper is very short in explaining the methods. The Stan model is mostly clear (some more details and motivation of choices would be useful, see below), but a lot of the analysis actually happens in post processing (UPGMA, consensus tree, computing the confounding strength,...). These post processing aspects are not described fully and required speculation in the review.
Page 4:
Is there a reason why a fixed mean of 0 makes sense here? Could you estimate it? A general preference for a certain feature state could be represented by a mean which is non-zero, but constant across languages (i.e. logit(ps,l)=s,l+s,l+s and estimate s as a parameter in Stan).
You use UPGMA to infer phylogenetic trees from the samples of the covariance matrix P later on, but it seems that in the Stan model P is not constrained to follow the correlation structure implied by a phylogenetic tree. What could be the effect of this? E.g. it seems that any contact effects that are above/below geographic expectations (i.e. distant languages with more contact than expected or proximate languages without contact effects) would be interpreted as closer/more distant relatives in P.
What is the motivation for using these particular GP kernels? How are they appropriate for linguistic evolution?
You are explaining that δP are sampled latent variables, but you don't mention how you obtain δG. Is it just the geodetic distance between the locations of the languages?
Page 5:
It is not entirely clear how the results of Stan are used as input for UPGMA. Is it the case that the P matrix is used as a similarity matrix for UPGMA?
Page 6:
It is unclear how the consensus tree was calculated. On page 5, it is described as an MCC tree, while on figures 2 and 3 a majority-rule consensus tree is described. This choice has potential ramifications for the calculation of the average standard deviation of split frequencies when it is calculated for all the clades present in the consensus tree (as different types of consensus trees can include different clades).
Page 7:
How are phylogenetic distances defined/calculated? Are you working directly with the δP values here?
More explanation/motivation is needed for the normalization choice here. The extreme values have a very strong effect on all other values and make a comparison with other values difficult. It's

not entirely clear why normalization is necessary or whether other ways of achieving it were considered.

Convergence issues
The paper's explanation for the lack of convergence is not convincing to us. The authors describe a scaling factor of the covariance matrix as a possible unidentifiability, which would hinder convergence. However, we do not understand why this scale factor would be unidentifiable: while the scale factor does not affect the tree topology, it does affect the likelihood and the prior probability of the Gaussian process model. Otherwise, the scaling factor only represents a single added dimension in the parameters spaces and we don't see why it would be particularly difficult for Stan to sample.
Even though non-convergence of parameters we are not interested in (like the scale factor of the matrix) is indeed not a primary concern, it might be indicative of deeper convergence problems. More details should be presented regarding convergence statistics. E.g. The authors mention that they ran four chains but provide only one value for the average standard deviation of split frequencies. On page 14 they mention that they calculated one value per chain, while one needs a pair of chains to calculate this convergence statistic. Additionally, it does not seem that the values of this convergence statistic are sufficiently low (especially the values for clades with higher support than 10% are very high), cf the guidelines of MrBayes recommending the statistic to be lower than 0.01 in order to consider that two runs have converged.
It seems also that the reported values for the average standard deviation of split frequencies is calculated only for the clades present in the consensus tree (for the low values of the statistic - 0.016 and 0.031), instead of all the splits. This would indicate that only the consensus tree has converged, rather than the whole posterior distribution. Even the higher values of the statistic are not calculated using all splits, leading us to suspect that the real value is even higher.
Unfortunately the simulation study did not seem to provide any clear insights into the convergence issues.

Minor points
Page 3:
Model architecture: Describing M as "the statistical model" and including it as a node in the graphical model seems odd. Usually the nodes represent variables and the whole graph is a description of the model.
The figure should just be this: G → D ← P
Typo: simulatenous → simultaneous
Page 10:
Typo: that themodel infers the the geographical –> that the model infers the geographical
Page 14:
Typo: randomtree –> random tree
Typo: majoritysupported –> majority-supported

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by evidence?
No

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: phylogenetic methods, bayesian statistics, inference of linguistic contact and linguistic areas, linguistic evolution

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 03 June 2024

https://doi.org/10.21956/openreseurope.18448.r40296

© 2024 Clarté G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Grégoire Clarté
Mathematics, The University of Edinburgh, Edinburgh, Scotland, UK

The article uses UPGMA method to reconstruct phylogenetic trees. The article is interesting in that it tries to include the possibility of lateral transfer explained in terms of geographical proximity.

The first remark, which is not mentionned in the article if I read correctly, is that the UPGMA method is criticable for building phylogenies. Although this method allow to compute a "distance" the process of building the tree with this distance is dubious (why middle point for the ancestor of two languages? Other questions arise in the model itself). It seems this model leads to strange results in the experiments, for example the position of Greek in Fig 6, which is not discussed. For more discussion on the subject there are articles for example Steel, M., Hendy, M. D., & Penny, D. (1988) ‘Loss of information in genetic distances.’ on the use of distances, and Greenhill "Language Phylogenies: Modelling the Evolution of Language". Overall, I would say that the model presented lacks interpretability and support from the process usually used to model linguistic evolutions.

The model itself is a little strange, it seems quite distant from the usual modelling in linguistics. It is not entirely clear how the cognates are used. Does it correspond to the "features" in the model? That is, can a cognate appear several times in the same language and different languages? I think this model is very far from reality, and can lead to strange results when the number of cognates increase (with the model, we would expect the cognates to have independent presence/anbsence in unrelated far away languages, say Sapmi languages and Germanic). The data from Northeuralex is even more strange, as it mixes several families which are known to be unrelated. Does this induce bias in the analysis? Are the cognates from e.g. Estonian that are actually borrowed from Latvian meaningful? I think independent studies would have been advisable. As a side note: the use of the Northeuralex data involves a computation of cognates using automated tools. There were warnings on this subject by List if I remember well.

Concerning the use of Stan, the analysis of the output is an interesting addition (and useful).

I think the model is interesting but further modelling assumptions should be discussed in the article. Additional comparison with other methods could be interesting as well (for example other methods only based on phylogenetic signal, such as the Dollo model). Nevertheless, it is interesting to see both geographical and evolutionary parameters learnt jointly.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Partly

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Partly

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian phylolinguistic, Bayesian modelling, Numerical methods.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 20 May 2024

https://doi.org/10.21956/openreseurope.18448.r40298

© 2024 Blum F. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Frederic Blum
[1] Max Planck Institute for Evolutionary Anthropology, Leipzig,, Saxonia, Germany
[2] Chair for Multilingual Computational Linguistics, University of Passau, Passau, Bavaria, Germany

The paper contributes a workflow of using Gaussian Processes for controlling for spatial biases during the inference of phylogenetic trees in linguistics. Based on two case studies, the authors show how the Gaussian Processes can identify contact between language pairs and its implications on the branch structure of the inferred trees. The method is a useful contribution, since spatial bias was often discussed, but rarely addressed quantitatively in phylolinguistic studies (cf. Neureiter et al. 2022). Considering the limitations (which the authors make explicit), the method will need to be verified on additional datasets in the future. That being said, typological studies should find a way for controlling for spatial and genealogical biases, and the present paper adds another important technique to the typologists toolbox.

The paper provides a clear line of argumentation that relates to the relevant previous research. One limitation that is not (yet) discussed in the paper is the focus on two datasets that consist mostly of European languages. It will be interesting to see how the method applies to other contact areas. Another possible comparison would be to test this method directly against similar approaches, such as Neureiter et al's (2022) contacTrees. It is likely that future phylogenetic studies adopt one of either method to implement such a spatial control, and the present study succeeds in highlighting this importance very well. Future phylogenetic studies should use some of the available methods to implement such a spatial control, and the study succeeds in highlighting this importance very well.

The authors could improve on two points. First, it is not clear to me how the distance measure between languages was computed. In case I have simply missed this, it should probably be highlighted a bit more, but it seems like this has only been made explicit for the simulated data of the Section 6.3 appendix. Considering that different ways exist to compare distances between languages (topographic, geodesic, etc), the processing of distances should be made explicit. This includes the source of the coordinates for each language (assuming a single-point location was used, e.g. Glottolog coordinates, or the individual datasets), and the publication of any pre-processing scripts for computing the distances, along the other scripts. Since one of the authors is also part of another paper that relates to linguistic geopgrahy and distance measures [Refer ref 1], this should be reasonable to elaborate on.

Second, some further improvements could be made with respect to the presentation of the data and code. Please make the code fully reproducible.
- Since the data is not part of the versionized repository, the download link might fail to work in the future. Given that the data itself is quite small, it would be recommendable to add the data to

the repository, so that it gets a DOI.
- Please upload the fitted models to some repository like OSF. Since the model takes long to run, it is easier for the reviewer and audience to reproduce the results if the model files are provided from the start.
- It would have been super useful to have a small Readme.md that introduces the individual files, introducing the dependencies between individual scripts.
- Since R packages are notoriously difficult to versionize, it is recommendable to use a tool like `groundhog` that stores the package versions and enables their installing. Without this, the code might break in the future.
- Hard-coded paths needed to be changed in order to run the scripts locally. Using relative paths and `paste` would ensure reproducibility across platforms.
- The code did not run without fixing individual lines, e.g. the unzip-command had to be changed in order to correctly read the file, and for the Stan-model the download only referred to the metadata. Alternatively, provide a How-To for the download in the Readme file.

```bash
> untar("jdellert-diss-supplements.tar.gz",files="inferred-cognates.tsv")
tar: inferred-cognates.tsv: Not found in archive
tar: Error exit delayed from previous errors.
Warning message:
In untar("jdellert-diss-supplements.tar.gz", files = "inferred-cognates.tsv") :
‘/usr/bin/tar -xf 'jdellert-diss-supplements.tar.gz' 'inferred-cognates.tsv” returned error code 1
```

Small typos:
- Page 4 (Section 3)> 'Fromboth'
- Page 10 (Section 4.5): 'themodel'
- Page 11 (Section 4.5): 'Rpackage'
- Page 12 (Section 5.2): 'are stationary across time can.'

References

1. Guzmán Naranjo M, Jäger G: Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology. *Open Research Europe*. 2023; **3**. Publisher Full Text

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

Yes

Is the work clearly and cogently presented?

Yes

Is the argument persuasive and supported by evidence?

Yes

If any, are all the source data and materials underlying the results available?

Partly

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian statistics; computational modeling of linguistic data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 25 January 2024

https://doi.org/10.21956/openreseurope.18448.r37655

© 2024 Brochhagen T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas Brochhagen
Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

The authors have convincingly addressed the concerns I raised in my earlier assessment. I thank them for the informative investigation of the convergence issues and for sharing their code.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational cognitive modelling; Bayesian methods; lexical typology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 20 July 2023

https://doi.org/10.21956/openreseurope.16746.r33777

© 2023 Brochhagen T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas Brochhagen
[1] Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[2] Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

The article "Gaussian process models for geographic controls in phylogenetic trees" introduces a new approach to phylogenetic inference that factors in potential geographic biases, which may otherwise skew the inference (e.g., contact zones that explain why linguistic features are shared add noise to phylogenetic inference that may mistake this commonality as signal of common ancestry). Geographic bias is a common problem in phylogenetic inference, and a central topic in typological research more generally. The proposal is a Gaussian process model that infers latent phylogenetic similarity after accounting for observed geographic similarity (closeness of point estimates of where a language is spoken). The model is evaluated on two datasets where geographic influences have already been studied so that model predictions can be benchmarked against past research. Results suggests that the model succeeds in characterizing geographic bias, as reflected by inferred phylogenetic similarities; which is to say that, as desired, geographic confounds are factored out of the inference.

The model has a few limitations. These include its failure to make accurate fine-grained predictions about individual linguistic features with respect to geographic bias; and chain convergence issues by the sampler. That being said, issues and limitations are clearly highlighted and discussed. They do not detract from the article's main contributions but rather contextualize them.

This is a very clear and well-executed study. The idea of using Gaussian processes to model geographic biases is becoming well-established, and adding them as a component of phylogenetic inference makes a lot of sense, as also demonstrated by the results. The evaluation on the two datasets, and additional checks on loanword identification, convincingly show that the proposed model is also of practical import; as well as where its limitations are. It would be interesting to see if the model scales to larger datasets, and to which extent divergences between a model that does not factor out potential geographic bias and one that does can be used to explore heretofore less prominently researched contact areas. Overall, I think this is a nice and solid contribution that will surely be further stress-tested in future research.

I have two main comments to offer. The first has to do with chain convergence (Results section). The authors mention that, with this model, chain convergence is not guaranteed. The reasoning offered makes sense to me, but I think that this methodological deviation from the standard diagnostics (in this case, checking that the same tree topology is expressed instead of computing, say, split R-hat as one would do normally) warrants further discussion and experimentation in the appendix so that researchers that want to use this approach know whether to trust their estimates or not. This brings me to my second and more important comment: the (re-)usability of the methods. As far as I can see, the Stan implementation is not shared. I would highly encourage the authors to make it openly-available so that the approach can easily be put to use and evaluated by others. Auxiliary scripts used to, e.g., compare the tree topologies obtained from different chains would also be useful to have. While I think the first issue can also be tackled in future research, I think the second (code availability) should be addressed in the current version. This is why I went for "Approval with Reservations" instead of "Approval"

Minor comments:

* Footnote 1 is incomplete (the section number is missing);

* "Stan" is named after Stanislaw Ulam. It's conventionally written with only the first letter capitalized.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Partly

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational cognitive modelling; Bayesian methods; lexical typology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Dec 2023

Frederik Hartmann

We thank the reviewer very much for their helpful comments and we attempted to make the paper and our arguments clearer following the reviewer's suggestions. The main changes are: (1) publishing the source code on Zenodo (see link in the paper body) and (2) providing a detailed investigation into the convergence issues and how they affect clade credibility (see online Appendix). Specific responses to individual points: I have two main comments to offer. The first has to do with chain convergence (Results section). The authors mention that, with this model, chain convergence is not guaranteed. The reasoning offered makes sense to me, but I think that this methodological deviation from the standard diagnostics (in this case, checking that the same tree topology is expressed instead of

computing, say, split R-hat as one would do normally) warrants further discussion and experimentation in the appendix so that researchers that want to use this approach know whether to trust their estimates or not. > We ran additional tests on simulated data to be able to better pinpoint the cause of the convergence issues. The appendix (Section 6.3) now contains a detailed discussion of those tests and their results. This brings me to my second and more important comment: the (re-)usability of the methods. As far as I can see, the Stan implementation is not shared. I would highly encourage the authors to make it openly-available so that the approach can easily be put to use and evaluated by others. Auxiliary scripts used to, e.g., compare the tree topologies obtained from different chains would also be useful to have. While I think the first issue can also be tackled in future research, I think the second (code availability) should be addressed in the current version. This is why I went for "Approval with Reservations" instead of "Approval" > The code files have now been uploaded to the repository as indicated in the paper (10.5281/zenodo.10247032). Minor comments: * Footnote 1 is incomplete (the section number is missing); * "Stan" is named after Stanislaw Ulam. It's conventionally written with only the first letter capitalized. > These errors were corrected

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 June 2023

https://doi.org/10.21956/openreseurope.16746.r32026

© 2023 Tresoldi T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Tiago Tresoldi
[1] Uppsala University, Uppsala, Sweden
[2] Uppsala University, Uppsala, Sweden

The paper significantly contributes to computational historical linguistics ("phylogenetic linguistics") addressing a key challenge in the field, as is very welcomed. The authors present a novel approach for modelling language change not exclusively as vertical transmission with random mutations, considering the significant role of horizontal transmission via language contact which is promoted by geographic proximity. They propose a hybrid approach combining spatial with phylogenetic modelling to address this. The probability of a language having a particular feature is assumed to depend on both the presence or absence of the feature in the language's spatial neighbours and in its genetic relatives. The paper then tests this model on datasets of known language families with known contact effects.

The analysis results show that the Gaussian process model approach yields mostly accurate inferences of contact zones and their effects on phylogenetic inference, yielding "deconfounded" trees. Note, however, that the model also appears to overestimate certain confounds, especially with languages that are both highly phylogenetically related and geographically close. The

authors acknowledge several limitations of the approach, including its reliance on a distance-based architecture and the inability to use it for loanword classifications. Despite these limitations, the paper concludes that the method provides valuable insights at different levels of analysis, which is useful for investigating micro-level effects such as pairwise distances and macro-level effects on a tree topology or individual clades.

The results should be interpreted as an initial exploration in this direction. While it would have been beneficial to evaluate the performance on other datasets, including synthetic ones, or to allow their system to infer the geographical similarity matrix 'G' just as `P`, these are not limitations that detract from the value of the paper. The authors' transparency in highlighting the biases and limitations of their approach is commendable. The results presented in the "Clade-level analysis" and "character-level analysis" sections are auspicious, and it seems evident that they are considering the generalization of the method with additional matrices, such as one for "cultural power" (addressing issues such as those of English of French).

Simultaneously, the paper introduces a methodology that deviates from the conventional approach in the field, which typically involves following tutorials on BEAST2. The authors' use of STAN and their construction of UPGMA trees for the posterior is refreshing, providing diversity in the field's methodology. However, the impact of this approach is diminished by the need for public access to the software pipeline, making it challenging for the reviewer to assess the method and reproduce the results. The source for the G matrix is not provided, and the contents of Figure 1 and the "Methods" session are too limited to a re-implementation aiming for reproducing results. It is hoped that the authors plan to release and refine the software pipeline post-publication. Furthermore, without access to the source code and all data, it is challenging to investigate potential issues related to the 'G' matrix and its subproducts, such as the apparent lack of linkage between Persian and Urdu/Hindi, as indicated in Figure 4. The availability of the software pipeline as part of the submission would significantly enhance the paper's value. While the decision to require the complete source code and all data ultimately lies with the editor, it is strongly recommended that the authors consider releasing it.

Specific issues:

- Figure 2 could be improved by indicating the support level used to accept branches. It appears to be 0.5, but confirmation would be helpful.

- In Figure 3, it would be beneficial to highlight that we are ultimately examining a cladogram. For instance, the Indo-European and Uralic members could be emphasized.

- The discussion of Greaco-Armenian and Italo-Celtic could benefit from citing more recent sources, such as Olander et al. 2022's "The Indo-European Language Family: A Phylogenetic Perspective".

- The first mention of "phylogenies" or "phylogenetic linguistics", as well as "cognate", could be explained for readers less familiar with the field. This is particularly relevant as we are discussing "cognates" as "root-meaning pairs" (i.e., more as in computational historical linguistics than traditional comparative linguistics).

- The statement "A well-known potential problem of phylogenetic..." could use a citation for support.

- The claim that "language contact almost always happens under spatial proximity" should be backed up with a citation.

- The statement "The probability is calculated as the logit sum of the variables ρ and γ." would benefit from a prior definition of what ρ and γ model.

- The repetition of "investigate" in "we investigate the strength of the geographical confound between languages on a general level, before investigating" could be avoided for better readability.

- The sentence "The clades the models identifies correctly..." could be restructured for clarity. For instance, "The models correctly identify several clades with high support. In the ASJP dataset, these include Slavic, Indo-Iranian, Romance, Germanic, and Celtic. In the Northeuralex dataset, they include Germanic (subdivided into North and West Germanic), Balto-Slavic, Slavic (split into East and West Slavic), Baltic, and Uralic (further divided into Finnic and Sámi)."

- The sentence "The level of support that well-established clades exhibit in this model indicates that this inference technique was successful and despite the drawbacks of distance-based methods (see Section Advantages, biases and limitations of the approach), the method can be deemed accurate enough to continue analyzing the results." should be split into two sentences for better readability.

- "This points towards the interpretation that the model over-corrects for geospatial effects in this case." -- I would say something like "This suggests that the model might be over-correcting for geospatial effects in this instance."

- "A reduction in clade support from the confounded to the deconfounded model means that adding the geographical control variable causes the model to construe some of the similarities between the constituent languages as contact-induced rather than phylogenetic." -- I would say something like "A reduction in clade support when transitioning from the confounded to the deconfounded model indicates that adding the geographical control variable leads the model to interpret some of the similarities among the constituent languages as contact-induced rather than phylogenetic."

- The sentence "Table 4 and Table 5 show the top 10 of these these distances for each dataset." has a repeated "these". A better phrasing could be "Table 4 and Table 5 display the top 10 distances for each dataset, demonstrating the most significant changes."

- There appear to be some formatting issues with the citations, such as nested parentheses in "(Greenhill et al., 2009)", "Chang et al. (2015))", "(Hartmann & Jäger, 2023)", and the missing parenthesis at the end of footnote 7. These should be corrected.

References

1. The Indo-European Language Family. 2022. Publisher Full Text

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Partly

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I have been doing research in computational historical linguistics and have interested in modelling geographic controls in phylogenies.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Dec 2023

Frederik Hartmann

We thank the reviewer very much for their helpful comments and we attempted to make the paper and our arguments clearer following the reviewer's suggestions. The main changes are: (1) adding references and (2) publishing the source code on Zenodo (see link in the paper body). Specific responses to individual points: The availability of the software pipeline as part of the submission would significantly enhance the paper's value. While the decision to require the complete source code and all data ultimately lies with the editor, it is strongly recommended that the authors consider releasing it. > The code and data have been shared on Zenodo Figure 2 could be improved by indicating the support level used to accept branches. It appears to be 0.5, but confirmation would be helpful. > The threshold of 0.5 is now indicated in the figure caption. In Figure 3, it would be beneficial to highlight that we are ultimately examining a cladogram. For instance, the Indo-European and Uralic members could be emphasized. > It is now indicated that this is a cladogram of Indo-European and Uralic languages. The discussion of Greaco-Armenian and Italo-Celtic could benefit from citing more recent sources, such as Olander et al. 2022's "The Indo-European

Language Family: A Phylogenetic Perspective". > The references were updated according to the reviewer's suggestions. The first mention of "phylogenies" or "phylogenetic linguistics", as well as "cognate", could be explained for readers less familiar with the field. This is particularly relevant as we are discussing "cognates" as "root-meaning pairs" (i.e., more as in computational historical linguistics than traditional comparative linguistics). > After the first mention of "phylogenies", the subsequent text makes it sufficiently clear, in our opinion, what it means. We further added a footnote explaining our usage of "cognate". The statement "A well-known potential problem of phylogenetic..." could use a citation for support. > A citation to this effect has been added. The claim that "language contact almost always happens under spatial proximity" should be backed up with a citation. > We did not find a suitable quotation, so we added a footnote justifying the claim. The statement "The probability is calculated as the logit sum of the variables ρ and γ." would benefit from a prior definition of what ρ and γ model. > The paragraph was restructured and clarifications were added. The repetition of "investigate" in "we investigate the strength of the geographical confound between languages on a general level, before investigating" could be avoided for better readability. > The sentence was changed accordingly. The sentence "The clades the models identifies correctly..." could be restructured for clarity. For instance, "The models correctly identify several clades with high support. In the ASJP dataset, these include Slavic, Indo-Iranian, Romance, Germanic, and Celtic. In the Northeuralex dataset, they include Germanic (subdivided into North and West Germanic), Balto-Slavic, Slavic (split into East and West Slavic), Baltic, and Uralic (further divided into Finnic and Sámi)." > The sentence was changed accordingly. The sentence "The level of support that well-established clades exhibit in this model indicates that this inference technique was successful and despite the drawbacks of distance-based methods (see Section Advantages, biases and limitations of the approach), the method can be deemed accurate enough to continue analyzing the results." should be split into two sentences for better readability. > The sentence was changed accordingly. "This points towards the interpretation that the model over-corrects for geospatial effects in this case." -- I would say something like "This suggests that the model might be over-correcting for geospatial effects in this instance." > The sentence was changed accordingly. "A reduction in clade support from the confounded to the deconfounded model means that adding the geographical control variable causes the model to construe some of the similarities between the constituent languages as contact-induced rather than phylogenetic." -- I would say something like "A reduction in clade support when transitioning from the confounded to the deconfounded model indicates that adding the geographical control variable leads the model to interpret some of the similarities among the constituent languages as contact-induced rather than phylogenetic." > The sentence was changed accordingly. The sentence "Table 4 and Table 5 show the top 10 of these these distances for each dataset." has a repeated "these". A better phrasing could be "Table 4 and Table 5 display the top 10 distances for each dataset, demonstrating the most significant changes." > The sentence was changed accordingly. There appear to be some formatting issues with the citations, such as nested parentheses in "(Greenhill et al., 2009)", "Chang et al. (2015))", "(Hartmann & Jäger, 2023)", and the missing parenthesis at the end of footnote 7. These should be corrected. > These errors were corrected, provided the typeset version reflects our changes.

Competing Interests: No competing interests were disclosed.