DATA NOTE

REVISED Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology [version 2; peer review: 1 approved, 2 approved with reservations]

Matías Guzmán Naranjo [1], Gerhard Jäger [2]

[1]Linguistics, Albert-Ludwigs-Universitat Freiburg, Freiburg, Baden-Württemberg, 79085, Germany
[2]Seminar für Sprachwissenschaft, Eberhard Karls Universitat Tubingen, Tübingen, Baden-Württemberg, 72074, Germany

v2 **First published:** 21 Jun 2023, **3**:104
https://doi.org/10.12688/openreseurope.16141.1
Latest published: 02 Jul 2024, **3**:104
https://doi.org/10.12688/openreseurope.16141.2

Abstract

It is common for people working on linguistic geography, language contact and typology to make use of some type of distance metric between lects. However, most work so far has either used Euclidean distances, or geodesic distance, both of which do not represent the real separation between communities very accurately.

This paper presents two datasets: one on walking distances and one on topographic distances between over 8700 lects across all macro-areas. We calculated walking distances using Open Street Maps data, and topographic distances using digital elevation data. We evaluate these distance metrics on three case studies and show that from the four distances, the topographic and geodesic distances showed the most consistent performance across datasets, and would be likely to be reasonable first choices. At the same time, in most cases, the Euclidean distances were not much worse than the other distances, and might be a good enough approximation in cases for which performance is critical, or the dataset cover very large areas, and the point-location information is not very precise.

Keywords

Typology, distance metrics, topographic distance, walking distance, linguistic geography

This article is included in the Linguistic Diversity collection.

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 2 (revision) 02 Jul 2024		✓ view	
version 1 21 Jun 2023	? view	? view	? view

1. **Dan Dediu**, Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
University of Barcelona, Barcelona, Spain

2. **Francesca Di Garbo**, University of Aix-Marseille, Aix-en-Provence, France

3. **Ezequiel Koile**, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

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

Corresponding author: Matías Guzmán Naranjo (mguzmann89@gmail.com)

Author roles: Guzmán Naranjo M: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Jäger G**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Software, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No [834050](Cross-Linguistic statistical inference using hierarchical Bayesian models [CrossLingference]). , assigned to Gerhard Jäger; and the Emmy Noether project 'Bayesian modelling of spatial typology', Grant number: (project number 504155622), assigned to Matías Guzmán Naranjo.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Guzmán Naranjo M and Jäger G. **Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology [version 2; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2024, **3**:104 https://doi.org/10.12688/openreseurope.16141.2

First published: 21 Jun 2023, **3**:104 https://doi.org/10.12688/openreseurope.16141.1

REVISED **Amendments from Version 1**

In this new version we have addressed most suggestions mentioned by the reviewers, these mostly boil down to: 1) we have rephrased and improved the text in several places. This should make the conclusions reported more in line with the results. We also better explain the models and the reasoning behind them. 2) we have corrected several formulae which were formatted wrongly in the first version. 3) we have improved the plots and made them easier to read as well as color-blind friendly. None of the results or calculated distances have changed since the previous version.

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

1 Introduction

Studying language contact, spatial diffusion, and typology (among others) requires having reliable distances measurements between linguistic communities. However, most work so far has used either Euclidean or geodesic distances (see Guzmán Naranjo & Becker, 2021; Guzmán Naranjo & Mertner, 2022; Ranacher *et al.,* 2021, for an example). Both these approaches, however, make some relatively simplified and unrealistic assumptions about the spatial separation of human populations. Euclidean distances assume that the earth is flat and distorted, and geodesic distances assume that the surface of the earth is a smooth sphere. While these assumptions can be warranted in some situations (e.g. communities which are very close together, or in individual islands), using these metrics leads to biased estimates of the separation of speech communities.

As a way of addressing these issues, a method for calculating approximate walking distances was recently proposed by Wichmann & Hammarström (2020), who propose a computationally efficient technique. This method, however, is not exact because it does not attempt to actually follow known walking pathways, but rather, uses population centres to route the paths. Another recent alternative for a computationally efficient approximation is proposed by Kaiping (2021), who calculates exact walking distances between the centre of geographic nodes (hexagons). Each node has an area of roughly 78 square km. Distances between languages are then calculated as the distance between the centres of these hexagons. While impressive, this method is also not exact, and we are not aware of evaluations of how good the resulting distances are for linguistic purposes.

In this paper we do several things. Our main aim is to provide a resource in the form of distance matrices, that can be used by typologists and linguists in general to study contact, areal pattern and other spatial phenomena in language. Second, we explore the question of how different distance metrics compare to each other. While there are several conceptual problems with using Euclidean and geodesic distances, there has been no attempt at quantifying how much better more realistic distance metrics are.

The structure of this paper is as follows. Section 2 gives a brief mathematical and computational description of the four distance metrics we will look at in this paper: Euclidean distances, geodesic distances, topographic distances and walking distances. Section 3 describes the materials and methods used for computing the topographic and walking distances. Sections 4 to 6 describes three case studies on modelling potential contact with three different datasets, using the calculated distances. Finally, Section 7 concludes the paper.

2 Distance metrics

2.1 Euclidean distance

The simplest type of distance metric we will discuss here is Euclidean distance. This is the distance of a straight line between two points in space. Given the points a at the coordiantes (x_a, y_a) and b at the coordinates (x_b, y_b), their distance is given by the formula:

$$d_e(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

This type of distance has been used explicitly (Guzmán Naranjo & Becker, 2021; Guzmán Naranjo & Mertner, 2022), but also implicitly, in the form of adding latitude or longitude to a statistical model (Verkerk & Di Garbo, 2022). The main reason for its relatively widespread popularity is simple: it is the simplest and fastest distance metric we can calculate for two points, and it can be obtained from latitude and longitude information. In terms of modelling, this distance metric also presents some advantages, namely the fact that popular Bayesian software like STAN or INLA provide ready-made solutions that make very efficient use of Euclidean distances for spatial models.[1]

Despite these advantages, there are several potential downsides. This distance metric assumes that the points in question lie on a plane, but this is evidently not true for two populations on the earth. For relatively short distances, and distances on the Equator, this is mostly not a problem, especially because we do not expect high accuracy in the centers from which we measure the distances. However, for larger distances, and distance far away from the Equator, Euclidean distance can produce results which are very different from actual distances along the surface of the Earth.

2.2 Geodesic distance

The geodesic distance,[2] or the distance as the crow flies, is the distance between two points on the surface of a sphere.[3] Given

[1]Particularly in the STAN ecosystem, BRMS provides ready-made functions to compute Gaussian Processes from latitude and longitude information using Euclidean distances, and there are multiple built in covariance functions that use Euclidean distances in the STAN language.

[2]Also referred to as the great-circle distance.

[3]The exact formula for computing the geodesic distance is not important, and we will not discuss it here.

that the geodesic distance takes into account the curvature of the earth, it is a likely better representation of the separation of two populations. Nonetheless, it also makes several simplified assumptions about the topography of the space between two points. Most importantly, the geodesic distance assumes a smooth surface, without hills, valleys or any other topographic barrier. While this assumption may be justified for cases like the steppe or island archipelagos, it is likely overly optimistic in places with very rugged terrains, mountain ranges, and similar geographic features.

Computationally, this distance is unproblematic, and we will not discuss the technical aspects any further. We calculated the geodesic distance between all lects in Glottolog with the `geodist` package (Padgham & Sumner, 2021) in R.[4]

2.3 Topographic distance

The topographic distance (or how the wolf runs) is the shortest distance between two points, considering the elevation changes in between both points. This is the distance along an uneven, rugged surface. To calculate topographic distances we used the `gdistance` package (van Etten, 2017) in R. Given its computational challenges, some words on the matter are necessary at this point.

Calculating topographic distances requires building an incidence matrix (a graph of connections between all points) with a digital elevation model (DEM) raster of the region containing the points of interest. Thus, the first step is to assemble a DEM for the region of interest, which, in our case, is the whole world. There are many sources for elevation data freely available, but not all datasets cover the whole planet (the northernmost and southernmost areas are often missing). For this we used the Global Multi-resolution Terrain Elevation Data 2010 (Danielson & Gesch, 2011), which does cover the whole globe, and is available at different resolutions (30-, 15-, and 7.5-arc-seconds). While ideally one would use the highest resolution possible, this is not computationally feasible for large areas. We have access to a High Performance Computing server with about 800 GB of RAM, but found this was not sufficient to build the incidence matrix for any macro area at the 7.5- or 15-arc-second resolution. For this reason, we used the 30-arc-second resolution of the data,[5] which roughly corresponds to about 1 square km per pixel (i.e. we cannot consider elevation changes that cover less that 1 km).

Given a DEM, we can calculate a graph of distances between adjacent points in the map (incidence matrices). The distance between two adjacepoints is given by $\sqrt{(h^2 + v^2)}$ where h is the horizontal distance, and v the elevation difference. We can then use this distance graph to calculate the shortest path between two points using Dijkstra algorithm, or any other similar algorithm to find the shortest path between two nodes in a graph (see Wang, 2020, for a more in-depth explanation). It is important to notice that there are alternative methods to calculate the distance between two adjacent points. Our approach assumes vertical and horizontal displacement is equally costly, but one could assign different weights to each.

However, even at a relatively low resolution, calculating topographic distances is still very resource intensive. Given these computational challenges, we only calculated distances between languages within different macro-areas. Additionally, for North America and Eurasia, we were not able to compute the incidence matrix for the whole macro-area and had to divide these into four, partially overlapping quadrants, and calculate the distance between languages for each quadrant. After having the distances for all points within each quadrant, we propagated the distances across quadrants using points in the overlapping regions.

A recent paper worth mentioning here is Koile *et al.* (2022), which makes use of travel-cost distances. The method used by the authors is similar to the one we present, but it attempts to calculate the travel time using a function to approximate hiking times, instead of taking the actual distance directly. In their study, the authors calculated the travel distances for languages spoken in 77 villages of the Caucasus, so it does not really represent the type of data we are trying to build in this study.

2.4 Walking distances

For the purposes of this paper, the walking distance between two points is the distance along mapped roads, walkways and paths that connect those two points. The idea is that road networks are a close representation of the spatial separation between populations because they are the actual pathways used for communication between communities. Of course, this assumes that moern road networks reflect the actual paths of communication one is interested in. This might be a sensible assumption when researching dialectal variation, but it might not be warranted when studying contact which is thought to have happened in the distant past.

For this paper we are using the Open Street Maps dataset (OpenStreetMap contributors, 2017) which contains information on roads and other infrastructure for most of the world.[6] For the routing we use the OSRM (Open Street Routing Machine) routing engine (Luxen & Vetter, 2011).

There are, however, some pitfalls calculating walking distances with this approach. These difficulties come from lack of connectivity between points on the map. This lack of connectivity arises in the case of islands which are not joined to the

[4]A middle way between Euclidean and geodesic distances is to use a UTM (Universal Transverse Mercator) projection for the data. Using a UTM projection splits the globe into 60 zones, and flattens each zone in a way ensures that calculating Euclidean distances on the coordinates produces near correct distances. We will not explore this approach in this paper.

[5]We used the breakline emphasis compression. This approach tries to maintain better sharp changes in terrain elevation (Gesch, 1999).

[6]We use the data dumps from https://download.geofabrik.de/ downloaded on the 22.02.2022.

mainland by ferry,[7] and some locations without roads or other transitable pathways. The later situation was especially present for languages in the Amazon.

Currently, we do not have any good way to solve this issue. A workaround one can take is to fill in the gaps by using geodesic distances when there are no roads. This approach should provide reasonable results for islands (given that communication between islands and the mainland would have been mostly as direct routes on ship), but it is only a rough approximation for unconnected points in the jungle.

We provide two datasets for the walking distances. One dataset has missing connections for these cases, and the other dataset tries to fill in these missing connections with a simple algorithm which sequentially connects the whole network to the nearest (by geodesic distance) off-network point.

Regarding computational issues, two observations are important. First, OSRM cannot build a graph (the data structure needed for the routing) for the whole world, which meant we had to work on each macro-area individually. Second, in computational terms, we found that the preprocessing steps to prepare the Open Street Maps data took a couple of weeks, but having built the OSRM graphs, calculating the actual distances is extremely efficient. Generally speaking, walking distances are easier to calculate than topographic distances on a moderately powerful server.

2.5 Taking stock

So far we have discussed four possible ways of calculating the distance between two points. It is useful to compare what the actual paths look like for the different distance metrics. Figure 1 provides an example of the paths between three points (the locations of three languages) in the Hindu-Kush area (see next section) for the four distance metrics. It is clear that the topographic, Euclidean and geodesic distances are relatively similar to each other, with the topographic path being less straight than the other two. However, the walking path is very different from the other three, and it takes a less straight route.

In computational terms, the topographic distances are the most challenging to compute. They require a considerable amount of resources, and can take several weeks for a single macro-area. Both the Euclidean and geodesic distances are the most efficient, and the walking distances sit somewhere in the middle.

The next three sections present three case studies in which we use these distance metrics to predict grammatical features of languages. The idea of these studies is not to gain linguistic insight, but to evaluate the predictive performance of the different metrics discussed here. To keep the models as simple as possible we will not consider any covariates outside the spatial term.

3 Evaluation: materials and methods

3.1 Datasets

We evaluate the four distance approaches in three datasets: Hindu-Kush (Liljegren *et al.*, 2021), South America (Carling *et al.*, 2018) and Europe (Moran & McCloy, 2019). The choice of datasets was partly opportunistic, and partly guided by theoretical considerations. Each case study presents a more detailed overview of each dataset, but in general terms each of these datasets comprises languages annotated for multiple binary features. This is important because in this study we limit ourselves to logistic regression to make all comparisons

[7]While OSM contains ferry information for many places, it is unclear how much information is missing.

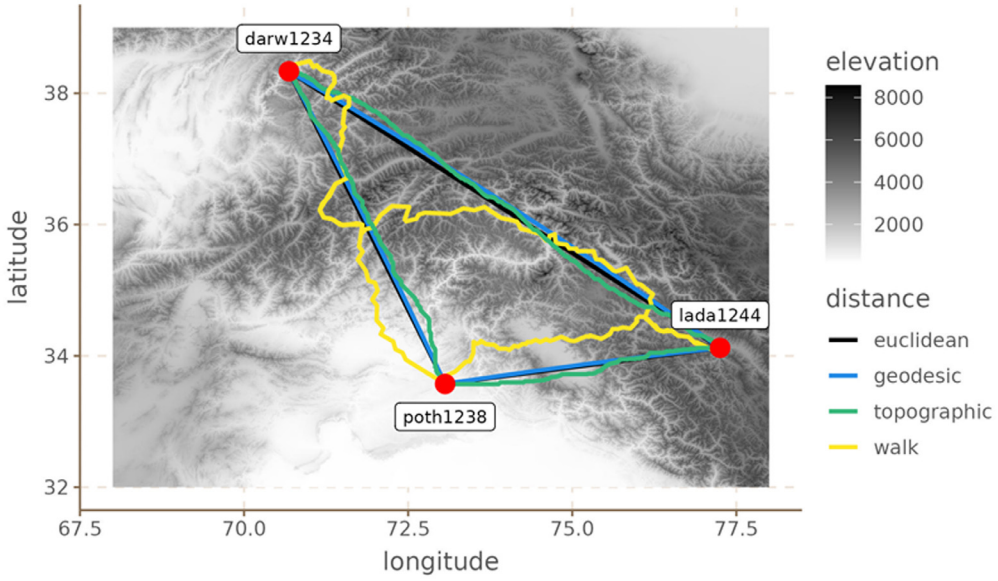

Figure 1. Example of distances' comparison for three Hindu-Kush languages.

equal.[8] Second, we chose three areas for which there are both interesting topographic features, and at least some note in the literature about contact effects. It is important to clarify that we do not expect that the features should be comparable across datasets, or that we expect the different types of features to behave similarly in terms of their spatial distribution, and correlation with the different distance metrics. On the contrary, we want to explore whether one of the distance metrics systematically shows higher correlation with different linguistic features in the different datasets.

In terms of types of linguistic features, the Hindu-Kush data combines phonology, lexical and syntactic features; the South America data exclusively comprises word order features; and the European data is made up of phoneme inventories. It is not our aim with this paper to explore any of these areas or datasets in depth. Our aim is simply to demonstrate how different distance metrics can lead to very different results when modelling contact.

3.2 The model

There are many different alternatives to model spatial relations from estimating simple correlations (van Gijn *et al.*, 2017) via autoregressive models (Murawaki & Yamauchi, 2018) to Gaussian Processes (GP) (Guzmán Naranjo & Becker, 2021; Guzmán Naranjo & Mertner, 2022). In this paper we use the latter for two reasons. First, GPs can be implemented fairly easily with Stan (Carpenter *et al.*, 2017), and second, they can use a distance matrix directly.[9] GPs are built around a kernel function which transforms the distance matrix into a covariance matrix. There are many alternatives for kernels, but we use a simple exponential kernel (see Duvenaud, 2014, for an in-depth discussion of different kernels).

While we focus here on it, these methods can also be used to estimate linguistic areas and contact areas (Guzmán Naranjo & Mertner, 2022).

In this paper we will model each feature independently from the other features. That is, we will predict each feature f_i with a logistic regression with a GP as the predictor.[10] We predict each feature individually, but there are some possible alternatives to look at all features simultaneously which could be preferable under certain circumstances (see Guzmán Naranjo & Mertner, 2022, for an example using Multiprobit models).

The model definition is as follows:

$$
\begin{aligned}
Y &\sim Bernoulli(invlogit(\mu)) \\
\mu &= \alpha + \eta \\
\alpha &\sim Normal(0,1) \\
\eta &\sim MultiNormal(0, \Sigma_{GP}) \\
\Sigma_{GP} &= K(x \vee \lambda, \delta, D) \\
\lambda &\sim InverseGamma(3,5) \\
\delta &\sim Normal(0,3) \\
K_{j,i}(\lambda,\delta,D) &= \delta^2 exp\left(\frac{-D_{j,i}^2}{2\lambda^2}\right) + \delta^2
\end{aligned}
\tag{2}
$$

(10)

Where α is the model intercept, λ is the length-scale of the GP, δ the standard deviation of the GP, and D is the distance matrix. K is the covariance function that transforms the distance matrix into a covariance matrix. The length-scale controls the distance at which two observations can influence each other significantly (a longer length-scale means a longer distance), and the standard deviation of the GP controls how strong the spatial variation can be.

To evaluate the model performance we use 10-fold cross-validation. We split the dataset into 10 groups, train the model using 9 of those groups, and predict the left out group. We then repeat this for all groups. Since we are dealing with binary features, we use balanced accuracy to measure the performance of the classifier. A balanced accuracy of 0.5 means that the classifier is performing at random chance, and thus we can conclude that there is effectively no spatial pattern to the feature in question. A balanced accuracy below 0.5, indicates an issue with the model or distance metric used. A balanced accuracy above 0.5, shows that the there is some spatial structure to the features in question, and that the model can pick up on it and use it to predict the values of the left-out observations.

4 Case study: The Hindu-Kush

4.1 Materials

This section presents a case study with languages of the Hindu-Kush. We use a dataset by Liljegren *et al.* (2021) which contains 59 languages for the Hindu-Kush area. Figure 2 shows the location of the languages in question.

The original dataset includes annotations for 80 binary features, from phonology and syntax. Since the values of many of these features were identical for all or almost all languages, we removed features with fewer than 10 or more than 49 positive values. Since we have 59 languages, this ensures

[8]This does not mean that either the distances we propose, or the method we describe in the next section only work in logistic regression. In fact, any type of likelihood would work.

[9]With models like those proposed by Murawaki and Yamauchi (2018), one first has to decide on how to build an incidence matrix of neighbours from the distance matrix. This process is not completely straightforward. However, the distance matrices we provide could also be used in those types of models.

[10]There is an important caveat to our modelling. Because of how the routing works, the walking distances do not necessarily satisfy the triangle inequality (in most cases by a couple of meters). This can happen, for example, because a street is one direction only, which means that the distance from A to B is different than the distance from B to A, because the routing algorithm needs to find a different paths. The consequence of this is that the estimation of the parameters in the models is biased. The problem arises because under some values of the length-scale, the resulting covariance matrix needs to be symmetric and positive definite. The effect on the sampler is that there are values of the length-scale which cannot be sampled and the posterior of the length-scale will not be correct. Other parameters will also be biased and should not be interpreted. Since for this paper we are only looking at predictive performance, this does not matter for our results. However, a researcher interested in understanding the actual spatial structures in the data should take extra care in fixing the distances.

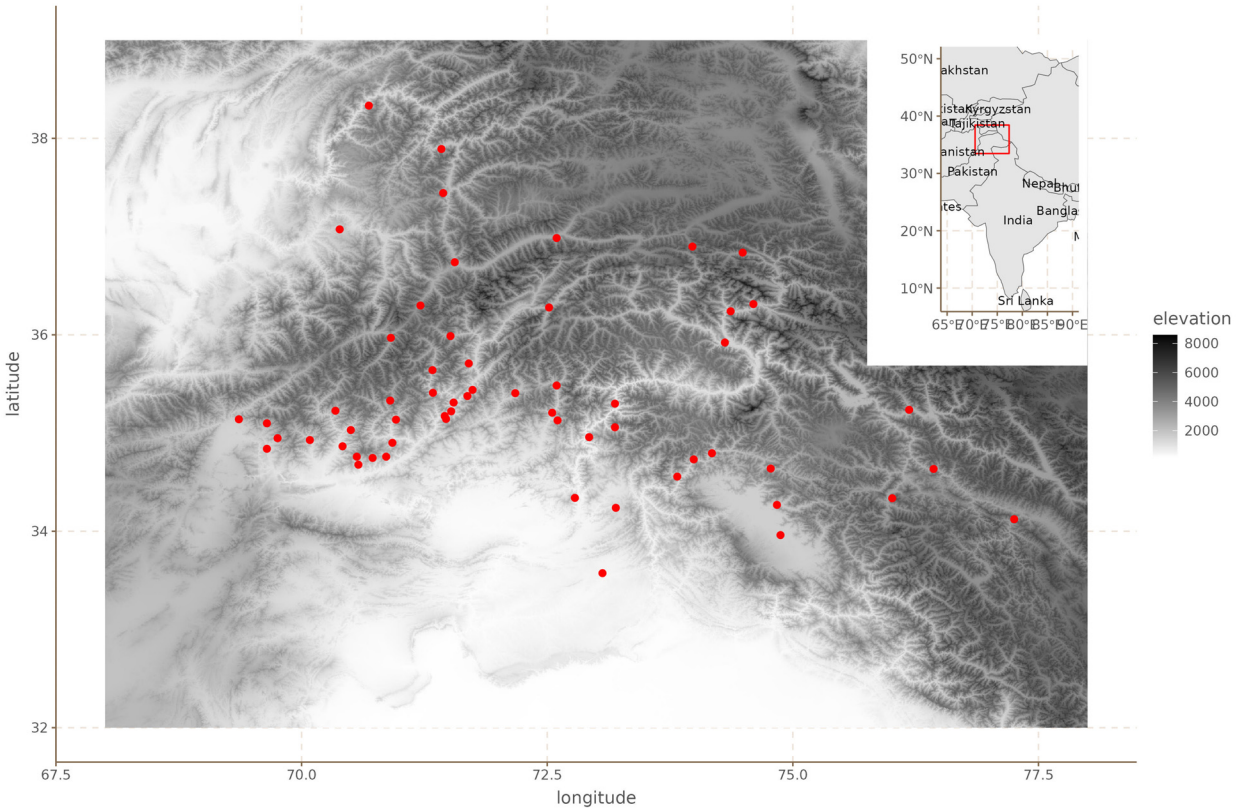

Figure 2. Hindu-Kush languages.

that for each feature, at least 10 languages have a different value than the majority value. This left us with a total of 48 features.[11] Some features have missing values for some languages. The model simply ignores those languages in the case of missing values.

Looking at the Hindu-Kush region is particularly interesting for two reasons. First, contact effects have been extensively documented for this area (Liljegren, 2019; Liljegren, 2020; Liljegren, 2022, and references therein),[12] which leads us to expect positive results at least to a certain extent. Second, and perhaps more importantly, the region is extremely mountainous as can be seen in Figure 2, which means that simple Euclidean and geodesic distances are likely biased estimates of the actual separation between communities (see also Figure 1).

4.2 Results

Figure 3 shows the balanced accuracy of each model for each grammatical feature.

These results are somewhat unexpected in that they show that there is no clear, systematic winner among the four distance metrics. Even the Euclidean distance, which we could expect to be the least accurate of the four, has the highest accuracy for some of the features like Possessive Suffixes and Oblique Object Word Order. Similarly, the topographic and walking distances perform rather poorly in features like Retroflex Fricatives and P Agreement (respectively). What this shows is that it is not instantly clear that one distance metric is better than the others in all cases.

Table 1 shows the mean accuracy for each distance across all features, and the aggregate counts for how many features each distance produces at the highest accuracy.

Going by these results, the walking distance outperforms in terms of the model the other distance metrics in 17 of the 48 features, followed by the topographic distance, then geodesic distance and finally Euclidean distance, which perform considerably worse. In terms of average balanced accuracy, the walking distances also seems to perform slightly better than the others.

We can visualise the differences in the models by plotting the conditional effects of a couple of these models. The conditional

[11] See the supplementary materials.

[12] See also the "Language contact and relatedness in the Hindu-Kush region" project: https://hindukush.ling.su.se/.

Figure 3. Balanced accuracy by feature and distance metric for the Hindu-Kush.

effects of a spatial model are predictions of a grid of points on the area in question (from 69 to 77.5 degrees longitude, and from 33.2 to 38.5 latitude, with steps of 0.05, for a total of 18297 points). To build these predictions, we need to calculate the distance from each point in the grid to each of the languages in the dataset. Because walking distances are sensitive to the existence of an accessible road from the point in question, it is not possible to build the required matrix for the conditional effects of walking distances, at least not for this area of the world.[13] For this reason, we only present the conditional effects of Euclidean, geodesic and topographic distances.

[13]A (computationally very costly) alternative would be to mix topographic and walking distances, defaulting to topographic distances for cases in which there are no available roads for the walking distances. It is however not obvious how we could go about doing this, and we leave this question open for future research.

Table 1. Aggregated accuracy by distance metric for the Hindu-Kush.

distance	mean accuracy	sd accuracy	n. times best accuracy	n.times best accuracy > 0.5
euclidean	0.62	0.14	6	3
geodesic	0.67	0.13	15	10
topographic	0.66	0.13	16	10
walking	0.68	0.13	17	9

For illustration we select the `Unique S Case` and `Zero Copula for Predicate Adjectives`, since these two seem to show large differences in the predictive power of the Euclidean distances. These are shown in Figure 4 and Figure 5.

In both cases, the difference is that the Euclidean distances produce a much stronger areal effect structure, with more extreme probabilities at the peaks. In contrast, both the areal patterns of the topographic and geodesic distances are smoother, and less extreme. This arises because Euclidean distances are overall shorter than either geodesic or topographic distances, which makes the model infer stronger spatial dependencies. However, in this case, inferring stronger spatial relations leads to overgeneralization and incorrect predictions.

One thing which this modelling approach fails to capture is the fact that a higher accuracy might not reflect the real contact situation. That is, the fact that in some cases the Euclidean distances produced better predictions, does not necessarily mean that the model reflects the actual contact scenario, and it is only finding spurious spatial correlations. A thorugh exploration of this scenario is outside the aim of this paper, but we further mention some considerations in the discussion.

5 Case study: European phoneme inventories

5.1 Materials

We now turn to European phoneme inventories. For this case study, we limit ourselves to languages found in the upper left quadrant for Eurasia (Western Eurasia), between -19.0212 and 82.3004 longitude, and 38.6147 and 68.8326 latitude. This area contains 118 languages in Phoible 2.0.[14,15] Because Phoible lists multiple phoneme inventories for various languages, we randomly chose only one phoneme inventory for each language. We then removed phonemes which were either too rare (present in fewer than 20 languages), or too common (present in more than 88 languages). This left us with a total of 55 phonemes.

Figure 6 shows the distribution of languages in our dataset together with the elevation.

5.2 Results

Figure 7 shows the balanced accuracy for each phoneme. Table 2 shows the mean balanced accuracy, and the number of times each distance metric achieved best accuracy, and best accuracy above 0.5. It is clear in this case that most features are hard to predict, and that they do not show areal patterns. However, for those features that do show areal patterns, both the Euclidean, geodesic and topographic distances outperform in terms of the model the walking distance metrics most of the time.

6 Case study: South American features

6.1 Materials

We are using the data for South American languages provided in DIACL (Carling *et al.*, 2018). This dataset contains data for 70 languages, across 18 binarized word-order features like whether the languages are A(gent)VO or not. As before, we removed features which were either too common (appear in 55 or more languages), or too rare (appear in 15 or fewer languages). The final dataset contains 10 features.[16] Figure 8 shows the spatial distribution of the languages in our sample.

6.2 Results

Figure 9 shows the balanced accuracy for each grammatical feature and Table 3 the mean balanced accuracy, and the number of times each distance metric achieved best accuracy, and best accuracy above 0.5. For this dataset we only find evidence of areal patterns for three of the features: VSa, So=Sa and the order AOV. Interestingly, for all three cases, the walking distances had either an at chance performance, or worse than chance. It appears that walking distances for South America are not reliable, or at least not for these data.

7 Concluding remarks

We have presented an overview of four distance metrics for typological research, two of which had not been computed

[14] https://zenodo.org/badge/latestdoi/19120525 (Moran & McCloy, 2019)

[15] An alternative database one could look at is Nikolaev (2018), but we chose Phoible because it seems to be more popular among typologists.

[16] Like for the Hindu-Kush dataset, DIACL contains some missing values for some features for some languages. In cases of missing data we simply omitted the languages with missing values for any given feature.

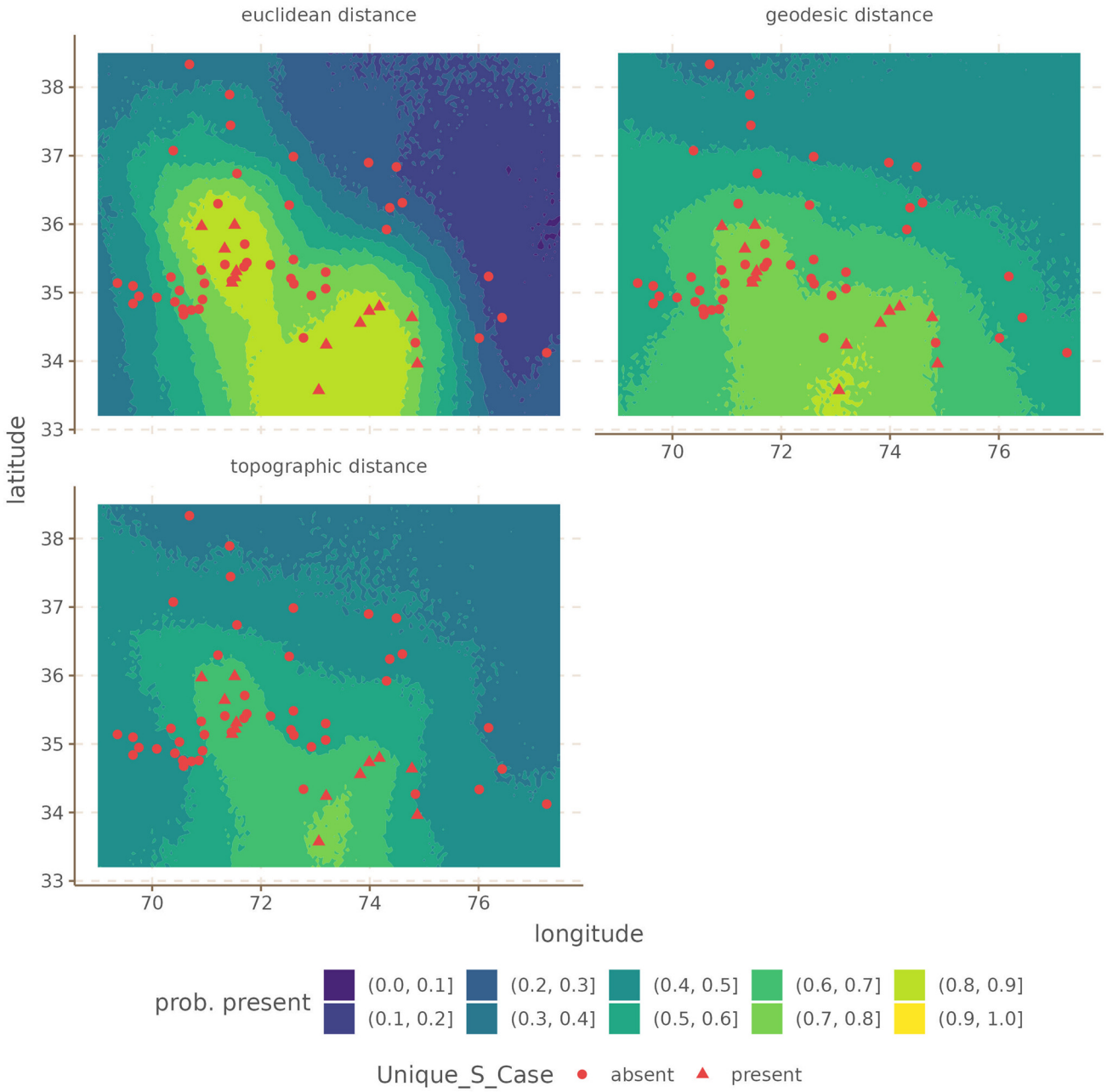

Figure 4. Conditional effects for `Unique S Case` for the Hindu-Kush.

before on a large scale. We show that it is in fact possible to compute topographic and walking distances for the world's languages. The resource we provide should be interesting to linguists working on topics related to the geography of languages and dialects, as well as researchers interested in language contact or other spatial phenomena.

We have presented one possible way of using these distances, namely with a Gaussian Process. Our modeling is not meant as an exhaustive exploration of the linguistic areas we use as examples, or the involved phenomena, they are meant as illustrations of how different distance metrics can lead to different results. It is important to emphasize that our results do not necessarily represent spatial patterns which are the result of language contact or diffusion, it is possible that some of the structures we observe emerge by chance distribution.

The results in terms of predictive performance are not completely clear, however. For the Hindu-Kush dataset, the walking distances showed a very small advantage over the

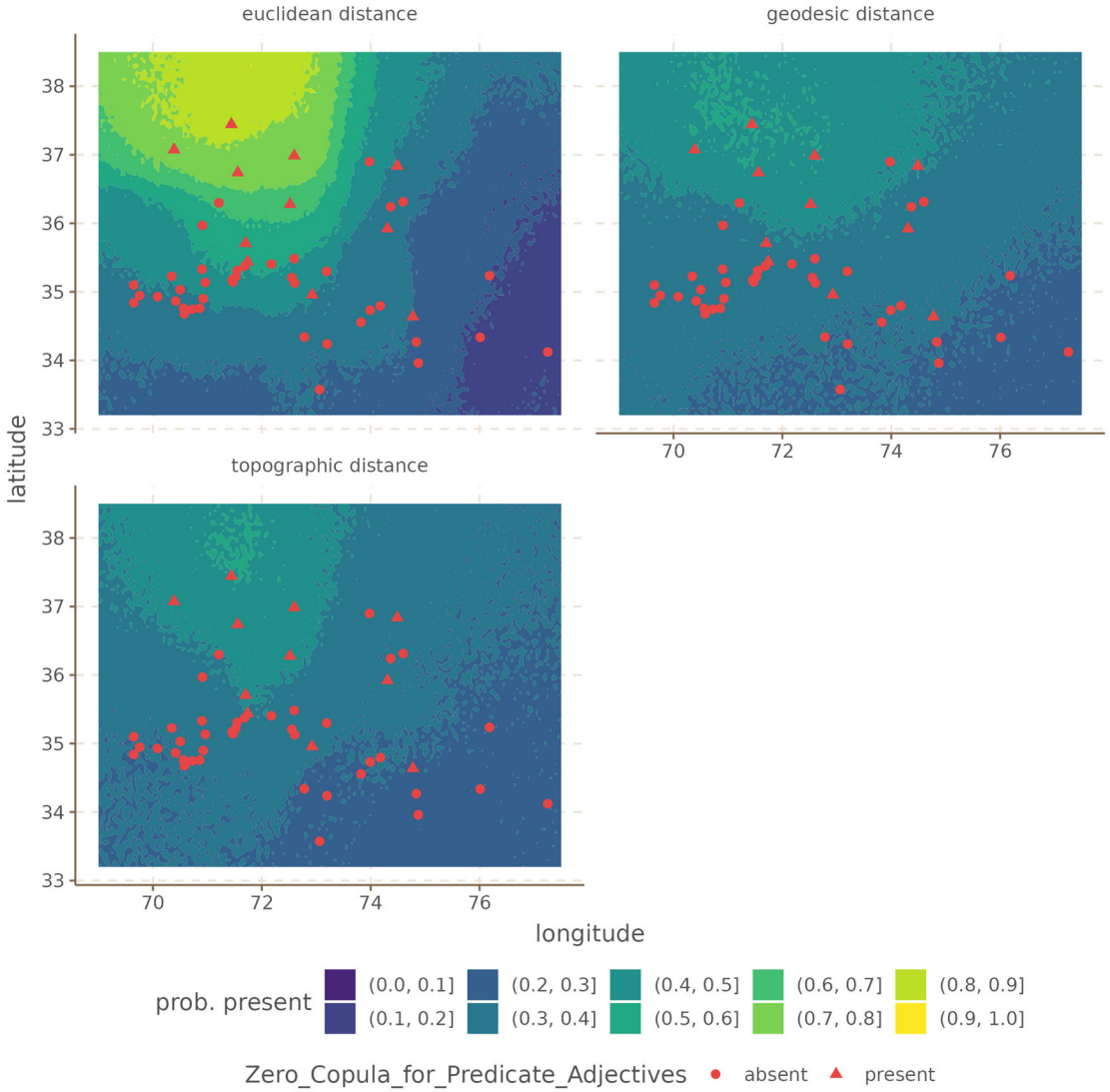

Figure 5. **Conditional effects for `Zero Copula for Predicate Adjectives` for the Hindu-Kush.**

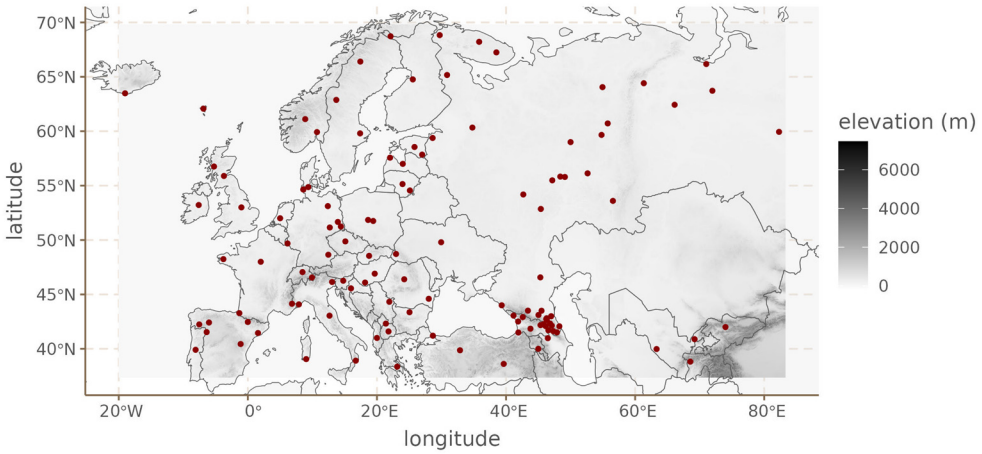

Figure 6. **European languages.**

Figure 7. Balanced accuracy by feature and distance metric for Europe.

Table 2. Aggregated accuracy by distance metric for Europe.

distance	mean accuracy	sd accuracy	n. times best accuracy	n.times best acc > 0.5
euclidean	0.54	0.09	18	1
geodesic	0.57	0.1	34	7
topographic	0.54	0.09	20	2
walking	0.53	0.08	18	1

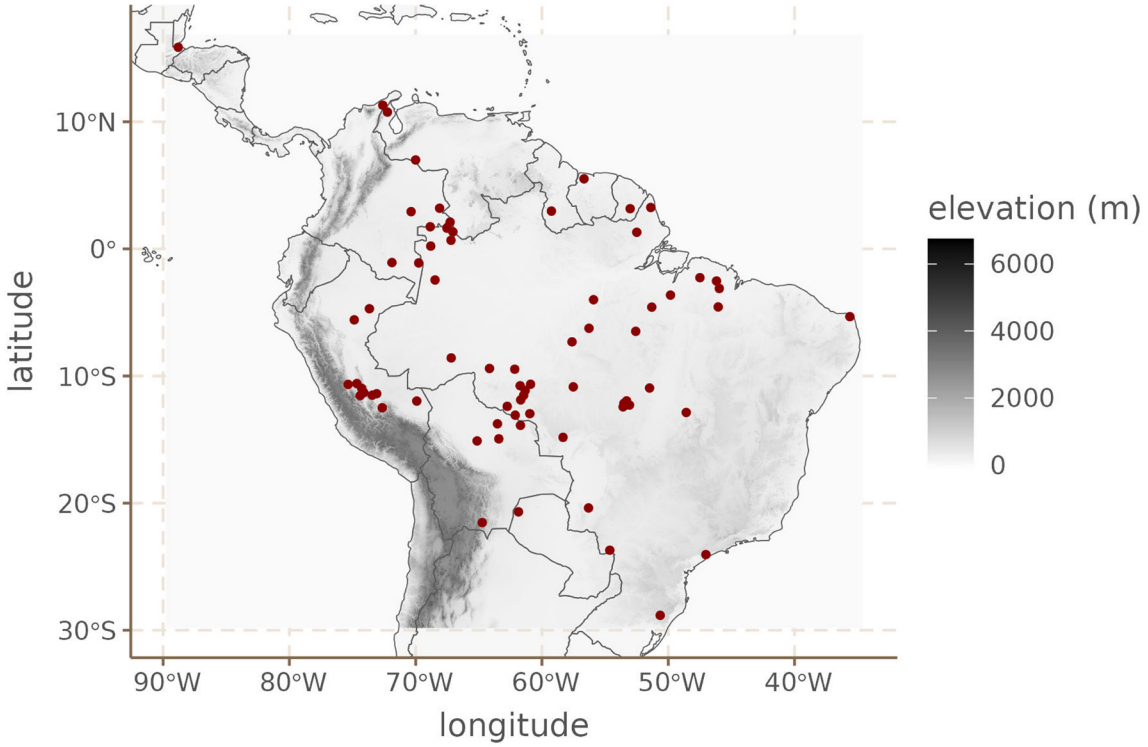

Figure 8. South American languages.

other distance metrics, and the topographic distance performed more or less at the same level as the geodesic distance, and the Euclidean distance performed worse of the four. However, for the European dataset, these results are reversed. Both the geodesic distances performed best, and topographic and Euclidean distances performed slightly better than the walking distance. Given the fact that the mapping system for Europe in OSM is more developed than for the Hindu-Kush area, we expected the results to be the other way around.

One possible explanation for our results is that walking distances are not very accurate representations of the spatial separation of populations very far apart. Additionally, the Hindu-Kush data has relatively good point accuracy of the location of the languages in questions, while the European data uses very rough approximations. These two factors could be causing the walking distances to perform poorly.

It is worth discussing the second factor in some more detail. While it is common to use point representations of language locations, we know that this is only an approximation of the real territorial extent of a language. While this approximation might be relatively accurate for languages with few speakers, languages with many speakers (e.g. Swahili, Russian, Mandarin, etc.) cannot be properly represented as single points in space, and a point representation will inevitably fail to capture the real contact dynamics of the language.For the South American dataset the results are somewhat more difficult to interpret. It is possible that modern roads and paths are a poor representation of migration paths and trade routes for the languages of the continent. Alternatively, it might just be that our route information for the region is suboptimal.

Overall, we can say for certain that the choice of distance metric can have a very large impact on the models. For some

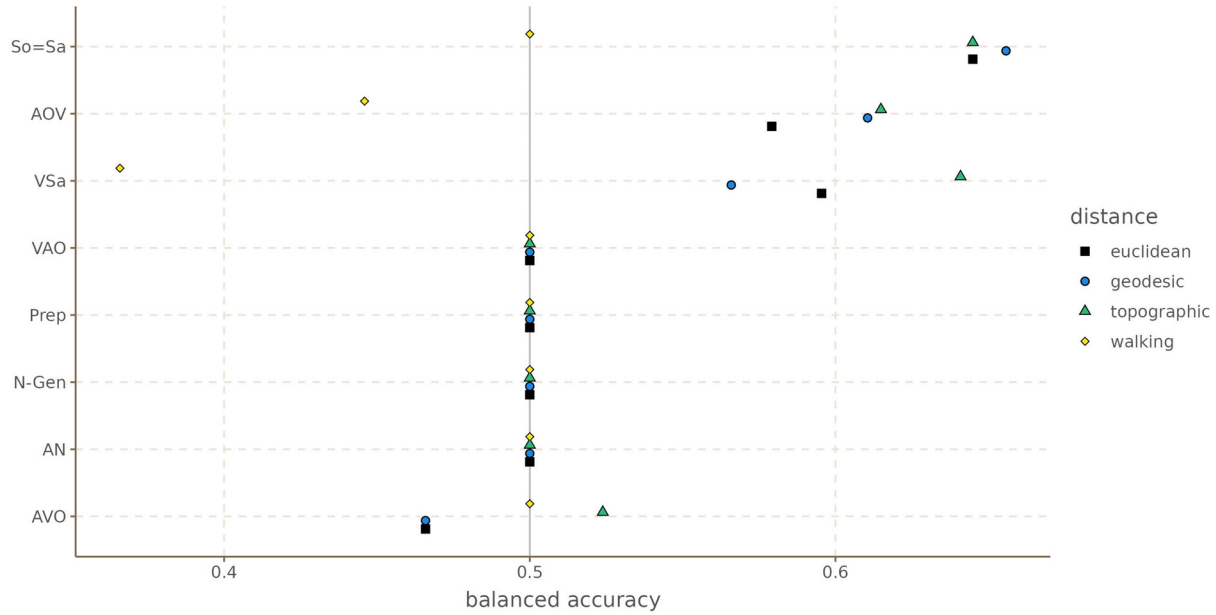

Figure 9. Balanced accuracy by feature and distance metric for South America.

Table 3. Aggregated accuracy by distance metric for South America.

distance	mean accuracy	sd accuracy	n. times best accuracy	n. times best acc > 0.5
euclidean	0.54	0.06	4	0
geodesic	0.54	0.07	5	1
topographic	0.55	0.07	7	2
walking	0.48	0.05	4	0

features, we saw upwards of 10% difference between the best and worse distance metric (e.g. Zero Copula for Predicate Nominals in the Hindu-Kush dataset). However, we cannot know a-priori which distance metric will better capture spatial patterns in any one case. From the four distances, the topographic and geodesic distances showed the most consistent performance across datasets, and would be likely to be reasonable first choices. At the same time, in most cases, the Euclidean distances were not much worse than the other distances, and might be a good enough approximation in cases for which performance is critical, or the dataset cover very large areas, and the point-location information is not very precise.

8 Data and software availability

All distance matrices for both walking and topographic distances are freely available and archived with Zenodo under CC-BY license: 10.5281/zenodo.7973820. The code for building the topographic distances is also available, as well as the code for running the test cases. We also include an environment file which should facilitate replication. See Guzmán Naranjo and Jäger (2023).

All Open Street Maps data used for the walking distances calculations can be downloaded from: https://download.geofabrik.de/. We used the versions as of 22.02.2022.

9 Author contributions

Matías Guzmán Naranjo: Conceptualization, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation

Gerhard Jäger: Conceptualization, Methodology, Software, Funding Acquisition, Writing – Review & Editing

10 Acknowledgments

We thank all members of the CrossLingference project for their helpful comments and suggestions.

References

Carling G, Larsson F, Cathcart CA, *et al.*: **Diachronic Atlas of Comparative Linguistics (DiACL)—a database for ancient language typology.** *PLoS One.* 2018; **13**(10): e0205313.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Carpenter B, Gelman A, Hoffman MD, *et al.*: **Stan: a probabilistic programming language.** *J Stat Softw.* 2017; **76**(1): 1.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Danielson JJ, Gesch DB: **Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010).** US Department of the Interior, US Geological Survey Washington, DC, USA, 2011.
Reference Source

Duvenaud D: **Automatic model construction with Gaussian processes.** PhD thesis. University of Cambridge, 2014.
Reference Source

Gesch DB: **The Effects of DEM Generalization Methods on Derived Hydrologic Features.** In: *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ed. by Kim Lowell and Annick Jaton. New York: Routledge, 1999; 255–261.
Reference Source

Guzmán Naranjo M, Becker L: **Statistical Bias Control in Typology.** In: *Linguistic Typology*. 2021.
Publisher Full Text

Guzmán Naranjo M, Jäger G: **Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology - Dataset.** Version 1. Also supported by the Emmy Noether project 'Bayesian modelling of spatial typology', Grant number: (project number 504155622). *Zenodo.* 2023.
http://www.doi.org/10.5281/zenodo.7973820

Guzmán Naranjo M, Mertner M: **Estimating areal effects in typology: a case study of African phoneme inventories.** In: *Linguistic Typology.* ahead of press. 2022.
Publisher Full Text

Kaiping G: **A network for simulating pre-colonial migration in the Americas.** In: *Santa Barbara: UC Santa Barbara: Center for Spatial Studies.* 2021.
Publisher Full Text

Koile E, Chechuro I, Moroz G, *et al.*: **Geography and language divergence: the case of Andic languages.** *PLoS One.* 2022; **17**(5): e0265460.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Liljegren H: **Gender typology and gender (in) stability in Hindu Kush Indo-Aryan languages**. In: *Grammatical Gender and Linguistic Complexity*. Ed. by Francesca Di Garbo, Bruno Olsson, and Bernhard Wälchli. Vol. 1 General issues and specific studies. Studies in Diversity Linguistics 26. Berlin: Language Science Press, 2019; **1**: 279–328.
Reference Source

Liljegren H: **The Hindu Kush-Karakorum and linguistic areality.** *J South Asian Lang Linguist.* 2020; **7**(2): 187–233.
Publisher Full Text

Liljegren H: **Kinship terminologies reveal ancient contact zone in the Hindu Kush.** *Linguist Typol.* 2022; **26**(2): 211–245.
Publisher Full Text

Liljegren H, Forkel R, Knobloch N, *et al.*: **Hindu Kush Areal Typology.** Version v1.0 [Data set]. Zenodo. 2021.
Publisher Full Text

Luxen D, Vetter C: **Real-time routing with OpenStreetMap data.** In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS '11*. Chicago, Illinois: ACM, 2011; 513–516.
Publisher Full Text

Moran S, McCloy D: **PHOIBLE 2.0**. Jena: Max Planck Institute for the Science of Human History, 2019.
Reference Source

Murawaki Y, Yamauchi K: **A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features.** *J Lang Evol.* 2018; **3**(1): 13–25.
Publisher Full Text

Nikolaev D: **The Database of Eurasian Phonological Inventories: a research tool for distributional phonological typology.** *Linguistics Vanguard.* 2018; **4**(1).
Publisher Full Text

OpenStreetMap contributors: **Planet dump.** https://planet.osm.org. 2017.
Reference Source

Padgham M, Sumner MD: **geodist: Fast, Dependency-Free Geodesic Distance Calculations. R package version 0.0.7**. 2021.
Reference Source

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**

van Etten J: **R Package gdistance: distances and routes on geographical grids.** *J Stat Softw.* 2017; **76**(13): 21.
Publisher Full Text

Van Gijn R, Hammarström H, van de Kerke S: **Linguistic areas, linguistic convergence and river systems in South America**. In: *The Cambridge Handbook of Areal Linguistics*. Ed. by Raymond Hickey. Cambridge Handbooks in Language and Linguistics. Cambride: Cambridge University Press, 2017; 964–996.
Reference Source

Verkerk A, Di Garbo F: **Sociogeographic correlates of typological variation in Northwestern Bantu gender systems.** *Language Dynamics and Change.* 2022; **12**: 155–223.
Reference Source

Wang IJ: **Topographic path analysis for modelling dispersal and functional connectivity: calculating topographic distances using the topoDistance R package.** *Methods Ecol Evol.* 2020; **11**(2): 265–272.
Publisher Full Text

Wichmann S, Hammarström H: **Methods for calculating walking distances.** *Phys A Stat Mech Appl.* 2020; **540**: 122890.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 2

Reviewer Report 09 July 2024

https://doi.org/10.21956/openreseurope.18965.r41849

✓ **Francesca Di Garbo**
Department of Language Sciences; CNRS Laboratory Parole et Langage (UMR 7309), University of Aix-Marseille, Aix-en-Provence, France

I am satisfied with the revisions implemented by the Authors and approve of this version of the article.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity, linguistic typology, language contact

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 05 October 2023

https://doi.org/10.21956/openreseurope.17425.r33111

? **Ezequiel Koile**
Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

This paper explores different methods measuring geographic distances, calculates these distances

for all pairs of languages at a global scale, and tests their correlation with linguistic features. It provides a valuable resource for the researchers in linguistic and cultural evolution, and discusses which of these definitions are more suitable for studying different geographic areas and linguistic features.

I think this is an interesting study which should be indexed after considering the following comments.

Major comments

1 - Euclidean distance

I do not understand why Euclidean distance is considered here, at least with the current definition. I would understand Euclidean distance as a rectangular distance on a plane, where both coordinates (x,y) are defined in the same units. On a spherical surface, the closest equivalent to Euclidean would be geodesic/great-circle distance. Calculating distances like in Eq. (1), just as the squared sum of latitude and longitude in degrees, treating them as distances, is simply wrong. I understand that this calculation is computationally non-intensive, but so is geodesic distance, and much more exact.

Euclidean distance as defined here can be useful at relatively short scales, especially near the equator (and it is used in this way inside each area studied in the present paper, except for Northern Europe), but there should be at least some approximation such as the haversine formula, or UTM projection, as mentioned in Footnote 4.

I disagree with the first statement in the sentence “Euclidean distances assume that the earth is flat, and geodesic distances assume that the surface of the earth is a smooth sphere.” for Euclidean distance as defined in Eq. (1) does not (only?) assume a flat surface of the earth, but a deformed one, where the E-W extension of Greenland at its maximum, from Etah at (long = -72°, lat = +78°) to Danmarkshavn (long = -18°, lat = +77°) is larger than the maximal E-W extension of South America, from Piura in Peru (long = -80°, lat = -5°) to Natal in Brazil (long = -35°, lat = -5°), while the first is 1,300km (54°) and the second one 5,000km (45°).

The authors acknowledge this fact: “However, for larger distances, and distance far away from the Equator, Euclidean distance can produce results which are very different from actual distances along the surface of the Earth.”, but still use these metrics.

The authors mention 4 references, two using either Euclidean or geodesic distances, and then two more using Euclidean distances either explicitly or implicitly. The first two do not use Euclidean distance as far as I understand. Murawaki and Yamauchi (2018) mentions:

“Languages were found to be associated with single-point geographical coordinates (longitude and latitude), as shown in Fig. 1. We constructed a spatial neighbor graph by linking all language pairs that were located within a distance of R km.”

Which I understand as geodesic distances, since these are measured in km.

In Ranacher et al (2021):

"Distance is modelled as a cost function C, which assigns a non-negative value ci,j to each pair of locations i and j. Costs can be expressed by the Euclidean distance, great-circle distance, hiking effort, travel times, or any other meaningful property quantifying the effort to traverse geographical space."

It is not clear to me what they mean by Euclidean distance here, but could not find this used as in Eq.(1) in the further calculation of costs.

As for the other references, it is true that Guzmán Naranjo and Becker (2021) use it, although pointing out that this is a simplification:

"For simplicity, we use Euclidean distances between languages. This is, of course, a simplification. It would be fairly easy to use other distance metrics with a GP, e.g. geodesic distances or walking distances (Wichmann and Hammarström 2020)."

Finally, in Verkerk and Di Garbo (2022), I do not see this so clearly. Latitude and longitude are used as independent predictors (one would expect that latitude is more determinant for it is more directly related with temperature differences), e.g.:

"The languages with the four different types of gender system differ in their mapping onto these four nonlinguistic variables, with the differences between types being larger for longitude and current forest overlap, and smaller for number of L1 speakers and latitude."

In addition, these measures are used in the African continent, which is relatively close to the equator.

2 - Cross-comparison of case-studies

I understand that the data available is not completely homogeneous for different areas, but I find it confusing that different linguistic measures are used in the different geographic areas:

"In terms of types of linguistic features, the Hindu-Kush data combines phonology, lexical and syntactic features; the South America data exclusively comprises word order features; and the European data is made up of phoneme inventories."

I do not see why these different features would be expected to behave in the same way. If there are divergences from the expected correlations with geographic distance, is this because of the topography of the region or because the features studied are affected to a greater/lesser extent by language contact? It would be interesting to test at least one group of features in all 3 geographic scenarios, or split the features in a comparable way (e.g. those in the Hindu-Kush into phonological, lexical and syntactic, to check which ones are more affected by contact).

More generally, the use of vocabulary like "...topographic distance tends to *outperform* the other distance metrics..." gives the idea of one distance being "better" than others in general. It should be clear in the writing that this outperformance is always a correlation with linguistic features, assuming that shorter geographic distances (in any of its flavors) is responsible for higher similarity among these features among languages.

3 - Report of main results

I do not agree with statements such as (Abstract):

"We evaluate these distance metrics on three case studies and show that topographic distance tends to outperform the other distance metrics, but geodesic distances can be used as an adequate approximation in some cases."

Or, more specific outputs for each case study such as:

"For the Hindu-Kush dataset, the walking distances showed a clear advantage over the other distance metrics..."

This is not so clear to me from Table 1. It seems to have performed as well as topographic and geodesic distances.

"However, for the European dataset, these results are reversed. Both the topographic and Euclidean distances performed considerably better than the walking distance and geodesic distance."

This is also not clear from Table 2. Geodesic distances perform clearly better than the other 3, which perform approximately equally in all 3 metrics.

Minor comments

General:

Numbering of sections: this is a little confusing along the paper. All sections start with the number 5: 5.1, 5.2,... Unless this is the fifth chapter of a collection, please correct. At the end of the paper, Sections 6-8 appear for Data and Software availability, Author contributions, and Acknowledgments respectively.

As for the subsections (sections after the 5.), it seems that some of them are incorrectly labeled. For example, at the end of the Introduction, Sections 5.2, 5.4, and 5.7 are mentioned only, while all should be referred: 5.3 is Materials and methods, and I believe 5.4-5.6 should be sub-sections of the same section (Case studies 1, 2, and 3).

Abstract:

Delete the first instance of "We evaluate these distances." and its corresponding line break.

Page 3, left column (in the following, P3L):

Last paragraph Introduction: Rephrase according to new numeration (see above).

First paragraph Section 5.2:

(x1,y1) and (x2,y2) : replace 1 → a, 2 → b for consistency with Eq. (1)

Eq.(1): missing exponent 2 in the second term on the right hand side

P3R:
Footnote 3: why is the formula for computing the geodesic distance convoluted? Clarify and add a reference to its definition.

Last line: “but we include” → “considering”

P4L:

First paragraph:

“northern most” → “northernmost” (same for southernmost)

“ram” → “RAM”

Second paragraph:

Is this formula correct, symmetric for horizontal and vertical distances? It would be more realistic to add a factor, v → a.v (a >1), and probably test the case studies with different values of a, compared to a = 1?

Complete the symbol for square root inline (if typographically possible).

Last paragraph:

“… but it attempts to calculate the travel time using a function to approximate hiking times, instead of taking the actual distance directly”. Related to the previous comment, there is only a dimensional factor between time and distance, the latter weighted by a higher cost of moving vertically.

“77 languages in the Americas” → “languages spoken in 77 villages in the Caucasus”

P4R:

1st paragraph:

“For the purposes of this paper, the walking distance between two points is the distance along mapped roads, walkways and paths that connect those two points. The idea is that road networks are a close representation of the spatial separation between populations because they are the actual pathways used for communication between communities.”

This might be reasonable for current language contacts, but be aware that in any study involving historic changes, even recent ones, these mapped pathways can differ to a great extent

2nd paragraph:

Open Stree Maps → Open Street Maps

OSRM = Open Street Routing Machine. Define this when first introduced.

Footnote 4: Explain the meaning of UTM.

P5:

Figure 1:

It should have a different caption, such as “Example of distances’ comparison for three Hindu-Kush languages”

Figure 2:

The caption should be “Hindu-Kush languages”.

Figures 1 and 2 (and probably 4 and 5):

Add an inset map, it is not very clear what part of the world we are considering beyond the lat and long values.

Add units for elevation (meters?)

P5L:

Last paragraph: “The choice of dataset was…”: dataset → datasets

P6L:

First paragraph:

“…a logistic regression with a GP as predictor.” → as a/the predictor ?

Footnote 10:

“…the walking distances do not necessarily satisfy the triangle inequality.”

Could you please expand on this? If d(A,C) < d(A,B) + d(B,C), why not move from A to C through B instead, reducing the magnitude of d(A,C) and satisfying the triangle inequality always?

P6R:

“from phonology to syntax” → “from phonology and syntax”.

P7L:

Top: feature → features

Figures 3, 7, and 9:

Add a dashed vertical line at balanced accuracy = 0.5 (all figures have different x-axis scale, so this would help comparability as well).

Rename the x axis as “balanced accuracy”.

The coloured dots are not easily readable in the current format. Even in Fig 9, where there are few features in the y-axis, the jittering applied does not help to understand which dot belongs to each feature. I suggest trying a different visualization, e.g. horizontal lines matching all 4 dots for each feature, and different icon shapes for these 4 dots (the colors can be kept besides the shapes). Also, reordering the y-axis by their value in the x-axis (see next item) can help in cases such as Fig 7, with a great proportion of dots on the x=0.5 line.

Features in each case are ordered in (inverse) alphabetical order. It might be easier to read if they were ordered differently, either by (average? minimum?) value of accuracy, or by some meaningful grouping (e.g. phoneme types in Fig 7).

Tables 1-3:

The row order seems arbitrary. Reorder either from simpler to more complex (euclidean, geodesic, topographic, walking) or by some of the values featured, e.g. “n. Times best accuracy”).

P8L:

Last paragraph:

“from 69 to 77.5 longitude…” add “degrees” or the degrees symbol.

“Because walking distances are sensitive to there being an accessible road from the point in question,” → “Because walking distances are sensitive to the existence of an accessible road from the point in question,”

P8R:

“These are shown in Figure 4 and Figure 5”: I think this sentence belongs to the previous paragraph.

“…Euclidean distance build…” → “…Euclidean distances build ⋯”

Figures 4 and 5:

It is not very straightforward to understand these figures. Why would conditional effects give predictions on presence or absence of a feature? Shouldn’t they only predict similarity among closer locations? If they are based on the distances to each language, isn’t it circular?

Also, why is the value “Indeterminate” present in Fig 5 but not in Fig 4? If it means NA, it should be

included in Fig 4 as well (unless there are no NA's in this variable). If this is a third category, is this not a binary feature? Also, please be consistent in the choice of symbols (the triangle means present in Fig 4 and indeterminate in Fig 5).

P8R:

"...in the upper left quadrant for Eurasia..." Do you mean in the Northwestern quadrant of Eurasia? If so, according to Fig 6, this looks more like Western Eurasia (= Europe mostly).

Footnote 15: Phoibl → Phoible

"It is clear in this case that most features are hard to predict, and that they do not show real patterns." Do you mean areal patterns?

P9R:

"...for the worlds languages." → world's.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity, quantitative 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.

Author Response 05 Apr 2024

Matías Guzmán Naranjo

We have carried out most of the corrections the reviewer suggests. Several formulas were miss-formatted, and several descriptions of the results were not correct, as the reviewer pointed out. We have fixed these. We have also improved several figures according to the reviewer's suggestions. There is one major point the reviewer mentions with which we're not in agreement. Below some specifics: * Regarding the comparison with Euclidean

distances: -> There are two things important here. First, and minor, some of the studies like Ranacher et al. do use in practice euclidean distance computations on a particular projection (personal communication with Gereon Kaiping), and in the case of Verkerk and Di Garbo, using the linear latitude and longitude independently, should be equivalent to using euclidean distances as linear predictors on each dimension. There are also some new studies that came out recently that use linear distances like Guzmán Naranjo and Mertner (2022), we have added this reference. The Second point is technical and comes from two issues we see. First, because brms is gaining popularity as a library for fitting models, it is becoming more and more common for researchers to use the in-built gp() function. This function can only compute euclidean distances. There is no alternative yet. To use other types of distances the researcher needs to write stan code directly, which is still not very widespread. The second factor that plays a role are what's called approximate Gaussian Processes, which are used for efficiency purposes. While fitting an exact GP to 1000 data points is slow but feasible, it is not possible to fit an exact GP to 2000 data points (at least not in normal research times and consumer hardware). Approx GPs are an alternative that comes very close to estimating the parameters of an exact GP, at the cost of minor exactness. However, approx GPs require covariance functions which are only defined for Euclidean distances (there are also some covariance functions for periodic kernels, but that's something else). Until someone defines an approx GP for geodesic distances (which is hard), Euclidean distances are the only real alternative here. So, to sum up. We think Euclidean distances are still used, and we do not really see a good reason to not include them in the study. We do not see that comparison as detracting from our main results. We even find that they are an ok approximation in some cases. * Regarding cross-comparisons: -> We do not expect these different features to behave the same way. We are also not trying to compare these features across areas. We only want to illustrate one of several possible uses for the calculated distances, and one of several ways of comparing them with respect to how well they capture the geographic distribution of linguistic features. We have clarified. * Regarding the formula for horizontal and vertical distance: -> It's correct. It is possible to use different formulas, but the distances take too long to compute to make this process feasible, at least for us. Additionally, you would actually need to use the angle because steepness is the key factor influencing speed and difficulty of travel. Finding better topographic-style distance is an ongoing project. * Regarding how walking distances calculated on modern pathways may not be valid for past contact: -> Indeed. As mentioned in the paper, we believe this might be one of the reasons that walking distances work so poorly for Europe. We have added a note at this point. * Regarding triangle inequality and OSRM: -> We have expanded a bit. The problem is very technical. The issue comes from how OSRM works internally. The routing algorithm in OSRM follow road constraints like allowed direction of travel as well as vehicular type allowed on the roads. This leads to a combination of factors which result in non-triangular inequality situations. For example, very often going from A to B requires following a single direction road on which there is no sidewalk, and thus, the direction of the road must be respected, which means that A->B and B->A cannot be done on that same route. This sort of situation can be corrected on small matrices with well understood algorithms, but these did not work on our larger matrices. What seems to be happening is that there are too many interacting edges, and if we fix A-B-C, some new issue arises for A-B-D and A-B-E, etc. The algorithm (and script) we tried: /* * Triangle fixing algorithm - implementation of algorithm 3.1 * (Metric_Nearness_L2) in Brickell, J., Dhillon, I., Sra, S., and * Tropp, J. (2008). The Metric Nearness Problem. SIAM. J.

Matrix * Anal. & Appl. 30, 375-396. * * (c) Toni Giorgino 2016 * Distributed under GPL-2 with NO WARRANTY. * * $Id: triangleFixing.c 424 2016-08-25 19:45:42Z tonig $ * */ * Regarding or ME plots: -> The model uses the GP to predict the expected value of a feature at some point, given the distance of that point to the points the model was trained on. It is the same thing with the CV, the difference is that for the CV we predict locations of languages we left out of the training, and for the ME we predict a dense grid of points.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 October 2023

https://doi.org/10.21956/openreseurope.17425.r33110

?

Francesca Di Garbo
Department of Language Sciences; CNRS Laboratory Parole et Langage (UMR 7309), University of Aix-Marseille, Aix-en-Provence, France

Summary and general comments

This paper introduces two novel techniques for computing distances between languages. The two distance metrics created through these techniques are a measure of topographic distance and one of walking distance between language communities. The paper illustrates the technical implementation of the two measures and shows how they can be used to study linguistic areality and contact effects from a typological perspective. The accuracy of the two metrics is compared against two distance metrics previously used in typological studies of language contact: Euclidean distance and geodesic distance. The four metrics are applied to three existing typological databases, each focusing on linguistic features of three established linguistic areas: the Hindu-Kush, Europe, and South America. The studies show that the choice of distance metric can significantly affect the type of generalizations that be drawn from a model, and that it is hard to establish a priori which measure would work best to capture areality effects in a given dataset. The results presented in the paper suggest that, in general, topographic and geodesic distance perform most consistently across the three datasets.

The paper makes an important methodological contribution to the study of spatial effects on the distribution of linguistic diversity. Firstly, because it enriches the range of methods typologists can use to control for areal biases in the distribution of structural properties of languages. Secondly, because the method offers a promising new way to account for areality bottom-up. This is especially crucial not only for the purpose of bias control, but also for detecting linguistic areas and/or areality effects in an exploratory fashion, for instance by testing hypotheses about the areal spread of given linguistic phenomena independently of pre-established areal groupings.

I would encourage the authors to develop a bit more on this second point, which does not emerge so neatly from the current version of the text, but, which, I believe, is of great potential for linguistic typology. Have the authors thought, for instance, about applying their new measures to the newly published Grambank dataset? This newly released database, the largest worldwide typological database to date, could offer an important test ground to further test the validity of the two distance metrics and, more generally, to detect areality effects bottom-up.

In addition to the above-mentioned suggestion, I have a number of more minor comments and suggestions of revision, which are highlighted below.

Minor comments and suggestions

Abstract

"We evaluate these distance metrics on three case studies and show that topographic distance tends to outperform the other distance metrics, but geodesic distances can be used as an adequate a pproximation in some cases."

This passage can be a bit misleading (when compared to the rest of the paper). It seems to imply that topographic distance is consistently better than the other three measures, while, in fact, depending on the model and the phenomena at stake, things can vary. I would suggest the authors to rephrase this passage by framing things in line with what stated in the concluding section (cf. *From the four distances, the topographic and geodesic distances showed the most consistent performance across datasets, and would be likely to be reasonable first choices. At the same time, in most cases, the Euclidean distances were not much worse than the other distances, and might be a good enough approximation in cases for which performance is critical, or the dataset cover very large areas, and the point-location information is not very precise.*)

Page 3

*"Section 5.4 describes a case study on
modelling potential contact with the different datasets."*

In the abstract, and in the rest of the paper, the modelling study is presented as a series of three independent case studies. Please rephrase in order to make things consistent.

Page 6

Figure 2. European languages > Hindu-Kush languages

Page 8, about the interpretation of Table 2 and of the European case study

"However, for those features that do show areal patterns, both the Euclidean and topographic distances outperform the geodesic and walking distance metrics."

Is this really the case? Unless I interpret things wrongly, judging from Table 2, it seems that it is the topographic and geodesic distances that outperform the other distance measures in this case.

Could you please clarify?

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity, linguistic typology, language contact

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 05 Apr 2024

Matías Guzmán Naranjo

We have implemented most of the reviewer's suggestions, and rephrased accordingly. Regarding the specific point on exploring areality with Grambank: -> We have and this is currently work in progress. Since it is a bit off-topic we don't go much into detail. But we now make a brief mention of it.

Competing Interests: No competing interests were disclosed.

Reviewer Report 29 June 2023

https://doi.org/10.21956/openreseurope.17425.r33113

?

Dan Dediu
[1] Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
[2] Department of Catalan Philology and General Linguistics, University of Barcelona, Barcelona, Catalonia, Spain

This paper introduces two new ways of computing between-languages distances and provides directly usable datasets for these (plus two “older” methods). Given the general importance of ways of quantifying space in many cross-linguistic (and, more generally, cross-cultural) studies, I think this is an important addition to the literature, especially given the massive computational resources needed to compute one of these new distances, resources that are now readily available to the “average” language scientist.

However, I do have a few suggestions and questions that might help improve the paper. In their (approximate) order of appearance in the paper:

Abstract:

- Remove 1st “We evaluate these distances.” (and, in general, the papers seems to in need of a careful reading).

- After having read the paper, I think the last sentence is a bit misleading, in that it is far from clear to me (and to the authors themselves, in fact) that “topographic distance tends to outperform the other[s]” in a meaningful and general way – I think this should be toned down as the paper is not about results but about the method and the distance matrices IMHO.

English language:

- Please check prepositions, spelling, plural, etc in: "warranted in some situations"; "leads to bias estimates"; "used for typologists and linguists"; "simplest and fasted distance metric"; "Open Stree Maps"; "show real patterns"; "but we chose Phoibl"; "before in a large scale"; "for the worlds languages"; "The simplest type of distance metric we will discuss here are Euclidean distances" -> “is”?; "the northern most and southern most areas" -> write connected (“northernmost”)?; "distances, faulting to topographic" -> defaulting?; "Euclidean distance build" -> “distances build” or “distance builds”

P3:

- missing ^2 after (y1-y2) in d_e formula.

P4:

- Para “Given a DEM…” -> please reformulate to make understanding this process easier the first time one reads this. My suggestion would be to 1st clarify that you compute the graph between all points in the DEM and then you use these to compute the minimum distance between the points one is actually interested in.

- Same para: please justify why sqrt{h^2 + v^2} as it is not immediately obvious why the h and v dimensions should be treated (a) separately and (b) have equal weight. Naively I would have expected something more like sqrt(x^2 + y^2 + w*v^2} where w is an (empirical) weight qualifying the relative “costs” of vertical vs horizontal displacements?

- Para “However, even at a …“ -> I was wondering if this would not work as a general solution in that maybe one can consider overlapping neighborhoods of the points of interest, but of course defining these neighborhoods brings in assumptions about the relative costs of vertical vs horizontal displacements (see above), but I guess it would be something worth discussing in the paper (if not testing)?

P5:

- “PGs are built” -> GPs?

Figure 1 (and figures in general):
- The color palette is atrocious for people with abnormal color perception! It is impossible the distinguish the first 3 distances on the map. Please try to use something like viridis or equivalent.

Footnote 10:
- Please explain why the triangle inequality is important for the GP and how bad their violation is (to footnote hints but I think this needs a better development).

Eq 9:
- The "()" after "GP" are meaningful?

Figure 2:
- That dark red (I guess) for the dots is very hard to see (for me) against the grayscale of the elevation.

- "European" -> "Hindu-Kush".

P7 (and elsewhere):
- please justify these cut-off points (here, <10 and >49 – especially 49 seems a bit weird).

- 2nd sentence and following of 1st para of section 5.4.2 "This metric..." -> this is partially repeating stuff that was already mentioned -> please move above.

Figure 3:
- could you also use symbols as well as colors? Highlight the 0.5 vertical line. Order the features in a meaningful way? Connect the same-color points by lines as well (resulting in an easier-to-follow "profile" but make it clear this does not imply any "linkage" between features). For the vertical jitter, either use the same jitter for all features and/or draw alternating colored bands by feature to make it clear which point belongs to each feature (now it is pretty hard to see, especially combined with the color scheme).

P8
- "What this shows is that it is not instantly clear that one distance metric is better than the others in all cases." -> I think the message is less optimistic here, namely that it really seems to depend on the feature considered, which is interesting in itself.

- Table 1 (and the others): the mean accuracy is hard to judge without some measure of variation (IQR, stddev...) – maybe better show the actual distribution (histogram/density)? (also, they look very similar to my eye – some statistical tsting would be useful?) Also n.times best accuracy is not very informative without knowing by what margin these were "best".

- Para "One thing which..." I think this is *very* important and should be discussed in the Discussion, together with the idea that languages are not points.

- "languages, we randomly chose only one phoneme inventory for each language." -> I know this is sometimes done, but it does not mean it is a good idea: sometimes, depending on the question, different inventories for the same language in PHOIBLE give very different answers...

- Again, justify the cut-off points, but here this might be even more important because, as argued for SegBo (see, e.g. Eisen, E. (2019)[1]) when looking at segment borrowing it is the

segments with around 50% frequency that might carry the best signal.

Footnote 16:

- I am confused as I think you mention that for H-K languages there was also some missing data?

P9:

- “For the Hindu-Kush dataset, the walking distances showed a clear advantage over the other distance metrics” -> I am not sure this is the message I got from the results…

Figure 6:

- add elevation legend (also make it clear that the legend’s values are *global*).

Finally, two general comments:

1. Language are not points: I clearly agree with the authors’ take in the paper (I do the same myself) but I think this point should at least be discussed and maybe ways of addressing it wrt the two new distances mentioned (if any)?

2. The right benchmark for these distances: I agree with the authors that this is not the point of this paper, but still choosing the right benchmark might make the difference between these two new distances being actively embraced by the community or lingering in the literature… Given that *most* actual uses of distances is to *control for* (i.e., remove) contact, maybe this would be a more appropriate benchmark? Also, (balanced) accuracy might not be the best (or, at least, the only) quantification of success here, but instead some sort of formal model comparison as well? And, finally, looking at known linguistic areas (e.g., the Balkans, which has non-trivial topography) might help? These are just ideas (and half-digested at that), of course, but I think the authors might want to at least discuss them given the amount of work (and computational power) already expended on this project?

References

1. Eisen E: The Typology of Phonological Segment Borrowing. *Thesis for: MA in linguistics. Hebrew University of Jerusalem*. 2019. Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity, cross-linguistic studies, statistics

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 05 Apr 2024

Matías Guzmán Naranjo

We have corrected most issues the reviewer highlights, and added further explanation where things were not clear, as well as reformulated claims which were too strong/not quite in line with the results. We have also redone the plots to make them more color-blind friendly. There are a couple of things where we do not quite agree on with the reviewer's suggestions, or consider it to be for future work. Here the concrete points: * Regarding ways of calculating travel times/distances: -> There are, of course, many alternative methods to calculate travel costs over a surface with elevation differentials. The formula we use is just Pythagoras theorem, it assumes that to go between two adjacent points, one has to travel in a straight line between them (the hypotenuse). It effectively reduces a complex surface to a surface made up of triangles. The result is a distance that approximates the real total distance traveled. The quality of the approximation depends on the resolution of the DEM. There are alternatives using formulas that approximate walking times which look similar to what you propose. We chose this solution because it avoids having to justify some value for w, which will inevitably change depending on a multitude of factors like terrain unevenness, steepness, flora, etc.. Testing travel time/travel cost metrics is an interesting idea we do plan to pursue in the future. * Regarding the suggestion of overlapping neighbors: -> Our guess is that the reviewer means something like what Kaiping did for the Americas? It’s an alternative method, but then you lose resolution for the distance between the neighbors. Depending on the purpose of the distance, that would make the calculation faster, but we already did most of the heavy lifting here. We are not sure that such an approach would help with higher resolution DEMs. Increasing DEM resolution even by a few arcseconds makes the data unmanageably large. * Regarding statistical testing of the tables: -> We’ve added sd to the tables but we’re not sure adding statistical testing would make much sense. We could add some credible intervals to the plot, but that would only clutter and, we believe, send the wrong message on what we’re trying to do with the paper. Proper, exhaustive comparisons of how these distances would require a whole different paper with a lot more tests. * Regarding randomly chosen phoneme inventories: -> The reviewer is correct. If one is interested in exploring some concrete linguistic question, then randomly choosing inventories might not be the best approach. However, we are only using these datasets as illustration of what can be done with our distances, and do not really think it is worth it to try to find the best dataset choices within phoible in this paper. * Regarding cutoff points: -> That’s precisely what we chose, phonemes which are neither too frequently present nor too frequently absent. * Regarding polygons and points: -> We agree. Languages are not points. Dealing with polygons is part of ongoing research by the first author, and it is still unclear how distances should be measured in such cases (from the centroid? on average? what about overlaps? etc.). But, however one wants to resolve that question, the technical question of how to then measure distances will be related to the methods we suggest here. * REgarding benchmarks and "controlling for": -> Well, technically, “controlling for” and “estimating” are the same thing when building a model. So when comparing model performance, if a distance produces better accuracy in a model like ours, it means it will ‘control for’ contact better, in the sense that it will capture a larger

amount of variance due to spatial correlations. -> In other words. Imagine you're interested in feature F for some group of languages. You have two options, geodesic distances and topographic distances. If you test the distance metric on its own, topographic distances achieve an accuracy of 0.8 while geodesic distances an accuracy of 0.6. This means that topographic distances capture more information about the spatial distribution and spatial relations in your data. -> What we present *is* formal model comparison, we are just using acc as a metric. We could also use something like ELPD instead, which is often preferred for these types of models. The downside of ELPD is that it's hard to interpret and not sound given the issues with walking distances and model sampling. This is the mean reason we think accuracy is a better choice in our case.

Competing Interests: No competing interests were disclosed.