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

Matías Guzmán Naranjo¹, Gerhard Jäger²

¹Linguistics, Albert-Ludwigs-Universität Freiburg, Freiburg, Baden-Württemberg, 79085, Germany
²Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen, Tübingen, Baden-Württemberg, 72074, Germany

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 distances. 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.

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

This article is included in the Linguistic Diversity collection.
5.1 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; Murawaki & Yamauchi, 2018; 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 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 bias 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 for typologists and linguists in general to study contact. 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 5.2 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 5.4 describes a case study on modelling potential contact with the different datasets. Finally, Section 5.7 concludes the paper.

5.2 Distance metrics

5.2.1 Euclidean distance

The simplest type of distance metric we will discuss here are Euclidean distances. This is the distance of a straight line between two points in space. Given the points \( (x_1, y_1) \) and \( (x_2, y_2) \), their distance is given by the formula:

\[
d_e(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}
\]

This type of distance has been used explicitly (Guzmán Naranjo & Becker, 2021), 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 centres 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.

5.2.2 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 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 & Summer, 2021) in R.\(^4\)

5.2.3 Topographic distance

The topographic distance (or how the wolf runs) is the shortest distance between two points, but we include the elevation

---

\(^1\)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 rather convoluted, and we will not discuss it here.

\(^4\)A middle way between Euclidean and geodesic distances is to use a UTM projection for the data. Using a UTM projection ensures that calculating Euclidean distances on the coordinates produces near correct distances. We will not explore this approach in this paper.
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 northern most and southern most 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, 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 the incidence matrix as the transitions between any two points in the map. The transition between two points is given by \( \sqrt{(h^2 + v^2)} \) where \( h \) is the horizontal distance between two points, and \( v \) the elevation difference. With the incidence matrix for the whole DEM, we can then 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).

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 77 languages in the Americas, so it does not really represent the type of data we are trying to build in this study.

### 5.2.4 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.

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. For the routing we use the OSRM 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 mainland by ferry, 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.

---

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

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

7While OSM contains ferry information for many places, it is unclear how much information is missing.
5.2.5 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.

5.3 Evaluation: materials and methods
5.3.1 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 dataset 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 equal. Second, we chose three areas for which there are both interesting topographic features, and at least some note in the literature about contact effects.

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.

5.3.2 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. 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.

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.

---

Figure 1. Three Hindu-Kush languages.
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).

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 predictor.\(^\text{10}\) 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:

\[
Y \sim \text{Bernoulli}(\text{inv-logit}(\mu))
\]

\[
\mu = \alpha + \eta
\]

\[
\alpha \sim \text{Normal}(0, 1)
\]

\[
\eta \sim \text{MultiNormal}(0, \Sigma_{GP})
\]

\[
\Sigma_{GP} = K(x|\lambda, \delta, D)
\]

\[
\lambda \sim \text{Inverse-Gamma}(3, 5)
\]

\[
\delta \sim \text{Normal}(0, 3)
\]

\[(9)\]

\[(10)\]

Where \( \alpha \) is the model intercept, \( \lambda \) is the length-scale of the GP, \( \delta \) 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 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.

5.4 Case study: The Hindu-Kush
5.4.1 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 to syntax. Since the values of many of these

---

\(^{10}\)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). The consequence of this is that the estimation of the parameters in the models is biased. 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. European languages.
feature were identical for all or almost all languages, we removed features with fewer than 10 or more than 49 positive values. This left us with a total of 48 features. 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), 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).

5.4.2 Results

Figure 3 shows the balanced accuracy of each model for each grammatical feature. This metric is similar to accuracy, but it takes into account the imbalances in the feature values. A balanced accuracy of 0.5 means that the model is performing at random chance, independently of how skewed the distribution of the feature values is.

Figure 3. Balanced accuracy by feature and distance metric for the Hindu-Kush.
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 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 effects of a spatial model are predictions of a grid of points on the area in question (from 69 to 77.5 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 there being 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.

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

<table>
<thead>
<tr>
<th>distance</th>
<th>mean accuracy</th>
<th>n. times best accuracy</th>
<th>n. times best acc &gt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>topographic</td>
<td>0.66</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>euclidean</td>
<td>0.62</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>geodesic</td>
<td>0.67</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>walking</td>
<td>0.68</td>
<td>17</td>
<td>9</td>
</tr>
</tbody>
</table>

13A (computationally very costly) alternative would be to mix topographic and walking distances, faulting 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 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.

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 distance build 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.

5.5 Case study: European phoneme inventories

5.5.1 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, 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.

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

Table 2

---


15An alternative database one could look at is Nikolaev (2018), but we chose Phoibl because it seems to be more popular among typologists.
5.6 Case study: South American features

5.6.1 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.

5.6.2 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.

5.7 Concluding remarks
We have presented an overview of four distance metrics for typological research, two of which had not been computed before in a large scale. We show that it is in fact possible to compute topographic and walking distances for the worlds languages.

The results in terms of predictive performance are less clear, however. For the Hindu-Kush dataset, the walking distances showed a clear advantage over the 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 topographic and Euclidean distances performed considerably better than the walking distance and geodesic distance. Given the fact that the mapping

---

16Unlike the other datasets in this paper, 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 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.

<table>
<thead>
<tr>
<th>distance</th>
<th>mean accuracy</th>
<th>n. times best accuracy</th>
<th>n.times best acc &gt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>topographic</td>
<td>0.54</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>euclidean</td>
<td>0.54</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>geodesic</td>
<td>0.57</td>
<td>34</td>
<td>7</td>
</tr>
<tr>
<td>walking</td>
<td>0.53</td>
<td>18</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
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 features, we saw upwards of 10% difference between the best and worse distance metric (e.g. Zero Copula for Predicative 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.

### 6 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.

### 7 Author contributions

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

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

### 8 Acknowledgments

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

### References

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 seem 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 fastest 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*h*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 testing 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

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.