Measuring Linguistic Diversity: A Case Study I

Amir Barghi
Department of Mathematics and Statistics, Saint Michael's College
Colchester, VT 05439, USA
abarghi@smcvt.edu

Abstract

In this paper, we look at different measures of linguistic diversity in thirty one countries, mainly in Asia, based on richness, Shannon, and Greenberg entropic indices and transforming them to the associated effective numbers. Moreover, we examine unweighted and weighted alpha, gamma, beta diversities (effective numbers) using Hill numbers. We then look at MacArthur's homogeneity and relative homogeneity. Finally, we put these countries into five regional groups and compute Sørensen-Dice and Jaccard indices in each regional group and between pairs of regional groups.

1 Introduction

In 1956, Greenberg [1] defined linguistic diversity as the probability that two randomly chosen individuals (with replacement) from a population have different first languages. In other words, if there are n languages in a population and p_i denotes the probability that an individual chosen at random from this population has the ith language as their first language, for $i \in \{1, \ldots, n\}$, then Greenberg's linguistic diversity index (LDI) is calculated as

$$H_2 = 1 - \sum_{i=1}^{n} p_i^2.$$

In ecology, H_2 is known as Gini-Simpson index (named after Italian Statistician, Corrado Gini[1], and after Edward Hugh Simpson [2]), and as discussed by Jost [3], it is measure of entropy (or uncertainty) and not diversity. According to Jost, a diversity index should measure "the effective number of elements in a system." In ecology, this is the number of equiprobable species in a population having the same entropic index. Jost shows that, in order to use the Gini-Simpson index (or equivalently, LDI) as a measure of diversity, one needs to consider the following transformation of H_2:

$$D_2 = \frac{1}{1 - H_2} = \frac{1}{\sum_{i=1}^{n} p_i^2}.$$

This is also known as the effective number of a population which was introduced by MacArthur in 1965 [4]. In addition to Gini-Simpson index, there are other measures of entropy (Richness, Shannon, HCDT, and Rényi) and their associated effective numbers that we will see later in this paper.

The notion of diversity using infromation theory, at least in ecology, has not been without controversy. Due to lack of consensus on its definition, Hurlbert suggested that this approach to

[1] https://www.britannica.com/biography/Corrado-Gini

measuring biodiversity to be abandoned in 1971 [5]. In response, in 1974, Peet defended the practice and provided guidelines for when and how entropic measures to be used to measure biodiversity [6]. More recently, Jost [3] clarified a common misconception among biologists and ecologists who use entropic indices as measures of diversity instead of transforming them into effective numbers, even though effective numbers were introduced by MacArthur in 1965 [4] and its properties were discussed by Patil and Taillie in 1982 in their comprehensive summary [7]. For a more recent treatment of the material covered by Patil and Taillie, see Ginebra and Puig [8].

The main focus of this work is to calculate effective numbers in thirty one Central, Southern, and Western Asian countries using different entropic indices. We also look at alpha, beta, and gamma diversities in these countries. Similar to linguistic effective numbers, alpha, beta, gamma diversities are computed based on alpha, beta, and gamma entropic indices, denoted by H_α, H_β, and H_γ, respectively. These diversities were briefly introduced and developed by Whitakker [9, 10]. In the context of this study, an alpha entropy H_α is the average of entropies within this region, i.e., these thirty one countries, a beta entropy H_β is the entropy between the countries, and a gamma entropy H_γ is the pooled entropy within this region, i.e., all thirty one countries combined. Both H_α and H_γ are transformed into measures of diversity and β diversity is calculated based on the other two diversities and it is the relative entropy between the countries or communities. Finally, we divide these countries into six groups, each with six to seven countries, based on geographical location and proximity. Based on these five groups, we explore the Sørensen-Dice and Jaccard indices to compare the commonality of languages and language families between pairs of regions.

2 Data

The thirty one Central, Southern, and Western Asian countries are the United Arab Emirates (AE), Afghanistan (AF), Armenia (AM), Azerbaijan (AZ), Bangladesh (BD), Bahrain (BH), Bhutan (BT), Cyprus (CY), Georgia (GE), India (IN), Iran (IR), Iraq (IQ), Isreal (IL), Jordan (JO), Kazakhstan (KZ), Kuwait (KW), Kyrgyzstan (KG), Lebanon (LB), Sri Lanka (LK), Nepal (NP), Oman (OM), Pakistan (PK), Palestine (PS), Qatar (QA), Saudi Arabia (SA), Syria (SY), Tajikistan (TJ), Turkmenistan (TM), Turkey (TR), Uzbekistan (UZ), and Yemen (YE). We divide these countries into five groups based on geographical location and proximity (see Figure 1):

- The Arabian Peninsula: AE, BH, KW, OM, QA, SA, YE;
- Central Asia: AF, KG, KZ, TM, TJ, UZ.
- Eastern Mediterranean: CY, IL, JO, LB, PS, SY;
- Southern Asia: BD, BT, IN, LK, NP, PK;
- Western Asia: AM, AZ, GE, IQ, IR, TR;

For our analysis, we use the following data sets:

- From Ethnologue Global Dataset (22nd Edition) [11], we use:
  - Table_of_Countries.tab
  - Table_of_LICs.tab
- From Glottolog 4.6 [12], we use:
  - languoid.csv

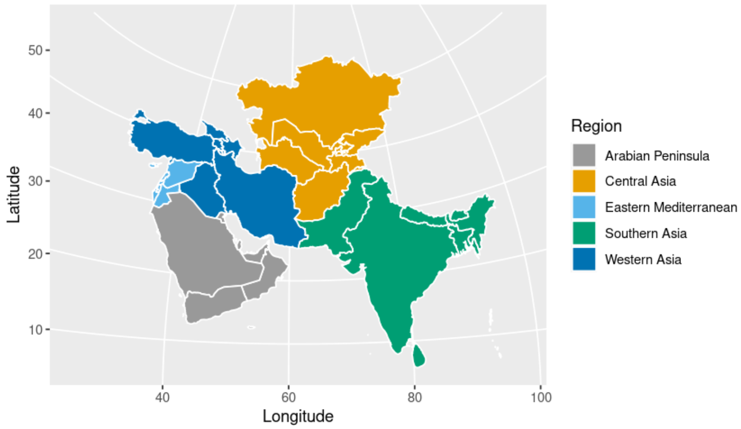


Figure 1: Thirty One Central, Southern, and Western Asian Divided into Five Country Groups Based on Geographical Location and Proximity

3 Preliminaries

In this section, we discuss different entropic indices and how to transform them into diversity measures, i.e., effective numbers. Recall that, in the context of this study, the effective number is the number of equiprobable languages in a population having the same entropic index. Throughout this paper, we will use diversity and effective number interchangeably. We also discuss alpha, beta, and gamma diversities in more detail in this section. Throughout this section, we assume that there are n languages in a population or a country, and p_i denotes the probability that an individual chosen at random from this population has the ith language as their first language, for $i \in \{1, \ldots, n\}$. We denote an entropy and a diversity by H_q and D_q, respectively, where q denotes the order of the entropy and of the diversity.

3.1 Richness

Linguistic richness is the simplest measure of diversity and it counts the number of languages in a country. In mathematical terms, the entropic index is defined as

$$H_0 = \sum_{i=1}^{n} p_i^0,$$

and it is not difficult to see that its effective number is

$$D_0 = \sum_{i=1}^{n} p_i^0.$$

3.2 Shannon

Shannon entropy, the very first measure of entropy in the context of information theory, was introduced by Shannon in 1948 [13]. It is defined as

$$H_1 = \sum_{i=1}^{n} p_i \ln(p_i),$$

and it measures the expected value of uncertainty in a system. Hill [14] shows that the effective number of this entropy is calculated as

$$D_1 = \exp(H_1) = \exp\left(-\sum_{i=1}^{n} p_i \ln(p_i)\right).$$

3.3 Greenberg

We saw in Introduction that Greenberg's index of linguistic diversity and Gini-Simpson index are the same, and we discussed how to transform them into effective numbers. It should be pointed out that Lieberson [15] extends the work of Greenberg [1] to bilingual and multilingual populations and communities and defines a generalization of LDI; however, in this work we will not consider this generalization and will focus on L1 speakers, i.e., people who speak a language as their first language.

3.4 HCDT

HCDT entropy was introduced by Tsallis in statistical mechanics [16] and is defined as

$$H_q = \frac{(1 - \sum_{i=1}^{n} p_i^q)}{q-1},$$

for $q \neq 1$. Jost [3, Appendix 1, Proof 1] implicitly shows that the effective number associated with this entropy is

$$D_q = (1-(q-1)H_q)^{\frac{1}{1-q}} = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}},$$

which are known as a Hill number, introduced in ecology by Hill in 1973 [14]. Note that richness and Greenberg diversity indices are special case of D_q for $q = 0$ and $q = 2$, respectively. As it is pointed out by Keylock [17], when $0 < q < 1$, Hill numbers favor languages with smaller p_i's (more rare languages), while for $1 < q$, they favor languages with larger p_i's (more common languages). Also, as shown by Hill [14], the effective number of Shannon entropy, i.e., $\exp(H_1)$, is the limit of D_q as $q \to 1$.

3.5 Rényi

Introduced by Rényi in 1961 [18] as a generalization of Shannon's entropy, Rényi entropy is defined as

$$H_q = \frac{-\ln\left(\sum_{i=1}^{n} p_i^q\right)}{q-1},$$

for $q \neq 1$. In particular, Shannon entropy is the limit of H_q as $q \to 1$. Jost implicitly shows (see Appendix 1, Proof 1 in [3]) that the effective number for Rényi entropy is

$$D_q = \exp(H) = \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{1-q}}.$$

Notice that the effective numbers computed from HCDT entropy and Rényi entropy are the same, and in this paper, we would refer to them as Hill numbers.

3.6 Alpha, Beta, and Gamma Diversities

Given a group of countries or communities, an alpha entropy H_α is the average of entropies within this group. This is also known as the conditional entropy [3]. A gamma entropy H_γ is the pooled entropy within this group, i.e., all the countries or communities combined, and a beta entropy H_β is the relative entropy between the countries or communities [3, 19]. All three entropies are then transformed into effective numbers using the appropriate transformations, resulting in D_α, D_γ, and D_β, respectively. Jost [19] argues that

$$D_\alpha D_\beta = D_\gamma.$$

As pointed out by Jost in [19], alpha and beta diversities (or entropies) are independent of each other. Veech and Crist [20] take a stance against this position by arguing that they are not statistically independent. In response, Ricotta [21] and Jost [22] point out the fallacies in Veech and Crist's argument. In this paper, we assume that they are independent, and consequently, we compute D_α and D_γ first, and then find D_β as the ratio of the other two.

3.6.1 Weighted and Unweighted Alpha, Beta, Gamma Diversities

Let N be number of samples or populations (in our exploration, countries) and let S be the total number of species (in our case, languages). Let us denote the proportion of L1 speakers of the jth language in the ith country, where $1 \leq j \leq S$ and $1 \leq i \leq N$, by $p_{j,i}$. Suppose w_i is the weight of the ith country. In case of unweighted alpha, beta, and gamma diversities, we assume that $w_i = 1/N$. However, for weighted diversities, we let w_i to be the ratio between the ith country's population and the total population of the countries in this exploration. Jost [3] shows that the weighted alpha diversity, for $q \neq 1$, is

$$D^q_{w,\alpha} = \left(\frac{\sum_{i=1}^{N} w_i^q \sum_{j=1}^{S} p_{j,i}^q}{\sum_{i=1}^{N} w_i^q} \right)^{1/(1-q)}$$

and, for $q = 1$,

$$D^1_{w,\alpha} = \exp\left(-\sum_{i=1}^{N} w_i \sum_{j=1}^{S} p_{j,i} \ln(p_{j,i}) \right).$$

For unweighted alpha diversity, let $w_i = 1/N$ for all i and we have

$$D^q_{u,\alpha} = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{S} p_{j,i}^q}{N} \right)^{1/(1-q)}$$

and, for $q = 1$,

$$D^1_{u,\alpha} = \exp\left(-\frac{\sum_{i=1}^{N} \sum_{j=1}^{S} p_{j,i} \ln(p_{j,i})}{N} \right).$$

For $q \neq 1$, The weighted gamma diversity is

$$D^q_{w,\gamma} = \left(\frac{\sum_{j=1}^S \left(\sum_{i=1}^N w_i p_{j,i} \right)^q}{\sum_{i=1}^N w_i^q} \right)^{1/(1-q)}$$

and, for $q = 1$, it is

$$D^1_{w,\gamma} = \exp\left(-\sum_{j=1}^S \left(\sum_{i=1}^N w_i p_{j,i} \right) \ln \left(\sum_{i=1}^N w_i p_{j,i} \right) \right).$$

When $w_i = 1/N$ for all i, we have

$$D^q_{u,\gamma} = \left(\frac{\sum_{j=1}^S \left(\sum_{i=1}^N p_{j,i} \right)^q}{N} \right)^{1/(1-q)}$$

and, for $q = 1$,

$$D^1_{u,\gamma} = \exp\left(-\sum_{j=1}^S \left(\frac{\sum_{i=1}^N p_{j,i}}{N} \right) \ln \left(\frac{\sum_{i=1}^N p_{j,i}}{N} \right) \right).$$

Finally, for $q \geq 0$, weighted and unweighted beta diversities are

$$D^q_{w,\beta} = \frac{D^q_{w,\gamma}}{D^q_{w,\alpha}} \quad \text{and} \quad D^q_{u,\beta} = \frac{D^q_{u,\gamma}}{D^q_{u,\alpha}},$$

respectively.

3.7 MacArthur's Homogeneity Measure and Relative Homogeneity

Related to alpha, beta, gamma diversities are the notions of similarity and relative homogeneity. Let us assume that we are studying N countries or communities. Since D_β is relative diversity between them, MacArthur's measure of homogeneity [4] can be calculated as $1/D_\beta$ [19]. According to Jost [19], this "answers the question 'What proportion of total diversity is found within the average community or sample?'" As Jost points out, the range for $1/D_\beta$ is the interval $[1/N, 1]$, where $1/N$ represents heterogeneity and 1 homogeneity among the countries or communities. Consequently, Jost suggests

$$R = \frac{1/D_\beta - 1/N}{1 - 1/N}$$

as a measure of relative homogeneity. Based on equation (22) in Jost's paper [19], using weighted beta diversity, we can calculate the relative homogeneity using

$$R = \frac{1/D^1_{w,\beta} - 1/D^1_w}{1 - 1/D^1_w},$$

where

$$D^1_w = \exp\left(-\sum_{i=1}^N w_i \log(w_i) \right).$$

3.8 Sørensen-Dice and Jaccard Indices

Not computed using entropic indices, Sørensen-Dice and Jaccard indices are measures of similarity between two samples or populations. We are including them in the study since we use MacArthur's homogeneity measure and relative homogeneity, which are computed using entropic indices. Sørensen-Dice index, introduced independently by Sørensen in 1948 [23] and Dice in 1945 [24], is computed as follows. Suppose X and Y are two samples or populations. Then, the Sørensen-Dice index for X and Y is

$$\frac{2|X \cap Y|}{|X| + |Y|} = \frac{2|X \cap Y|}{|X \cup Y| + |X \cap Y|}.$$

Jaccard index, introduced in 1912 [25], computes similarity between X and Y as

$$\frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}.$$

4 Results

For analysis, I used the open source statistical and programming software R. And to run the code, I used Jupyter notebooks, using the R package `IRkernel`, to not only run the R code but also to create an interactive narrative and lecture slides. The R packages that I used for this project are as follows: `tidyverse`, `latex2exp`, `maps`, `gganimate`, and `gifski`. Please note that the tables and figures referred to in this section are in 6 (the Appendix).

4.1 Three Initial Measures of Linguistic Diversity

As we see in Table 1, based on richness, Shannon, and Greenberg effective numbers, India is the most linguistically diverse country among these thirty one countries. Each sub-table in Table 1 is ordered in a descending order based on the values of each of these effective numbers. If we take the ranks (row numbers) in each of these sub-tables, calculate the average of these ranks, and order them in an ascending order, we get Table 2. Based on Table 2, India, the United Arab Emirates, Israel, Afghanistan, and Iran are the most linguistically diverse countries in Southern Asia, the Arabian Peninsula, Eastern Mediterranean, Western Asia, and Central Asia, respectively. The maps in Figure 2 demonstrate richness, Shannon, and Greenberg effective numbers on logarithmic scale in these countries using color. The maps in Subfigure 2a, 2c, and 2e use different overall ranges. When we use the same range for all three maps, we get the maps in Subfigure 2b, 2d, and 2f.

4.2 Linguistic Diversity Based on Hill Numbers

We now consider the effective numbers for each country based on region for $0 \leq q \leq 5$ in Figure 3. We have included the effective numbers for each country for $0 \leq q \leq 5$ in Figure 4 for overall comparison. Based on Subfigure 3b, we see similar values of richness among the countries in Central Asia, with Kazakhstan having the highest value of richness. However, as q increases, Afghanistan demonstrates higher values of linguistic diversity by a large margin. The rest of the countries in this region show a similar behavior as q increases. By comparing Subfigure 3a and Subfigure 3b, even though Bhutan in Southern Asia has a lower value of richness diversity compared to Afghanistan, we see a similar behavior as q increases among these two countries. In Western Asia, based on Subfigure 3c, even though Iran has the highest value of richness, Iraq has a similar Shannon effective number to that of Iran's and higher Hill numbers, for $1 < q \leq 5$, relative to those of other countries in this region. We see a similar behavior among Azerbaijan, Georgia, and Turkey as q increases. In Eastern Mediterranean, Israel has the highest values of richness, Shannon, and Greenberg effective numbers by a large margin. The rest of the countries in this region show a similar behavior as q increases. In the Arabian Peninsula, Oman has the highest value of richness, the United Arab Emirates the highest value of Shannon, and Qatar the highest value of Greenberg effective numbers.

4.3 Alpha, Beta, and Gamma Diversities

In A Tidyverse Approach to Alpha, Beta and Gamma Diversities[2], I have checked the validity of my code for the current paper by comparing the results that I get here by applying my code to an ecological data set and those that I get from using functions from the R package `vegetarian`. I should point out that `vegetarian` is no longer available on the CRAN (The Comprehensive R Archive Network) repository. Since `vegetarian` is no longer maintained on CRAN, I am working on rewriting the code that uses another ecology R package called `vegan`, which is an alternative to `vegetarian`.

In Table 3, we have the weighted and unweighted gamma, alpha, and beta diversities. Moreover, the table contains the values of MacArthur's Homogeneity and Relative Homogeneity using richness, Shannon, and Greenberg effective unweighted beta diversities. In Figure 5, we have the plots for unweighted and weighted gamma, alpha, and beta diversities. As we see in Subfigure 5e and 5f, it not appropriate to use unweighted gamma, alpha, and beta diversities because of the great population disparities between these countries.

4.4 Sørensen-Dice and Jaccard Indices

As we see in Subfigure 6a and 6b, using both Sørensen-Dice and Jaccard indices, Central Asia has the highest level of language similarity and Southern Asia has the highest level of language dissimilarity on average. Moreover, as we see in Subfigure 6c and 6d, using both indices, Western Asia and Southern Asia have the highest level of similarity and dissimilarity on average, respectively, based on language families. The density functions when using both indices and using language level and family levels are given in Figure 8.

Examining pairs of regions, based on both Sørensen-Dice and Jaccard indices, we see the highest level of average similarity between Central Asia and Western Asia and the highest level of average dissimilarity between Southern Asia and Eastern Mediterranean, as we see in Figure 7.

5 Future Directions

One possibility is to use Rao's quadratic index along with linguistic family trees to define similarity functions to measure similarity in each country and all the thirty one countries combined. Rao's quadratic index of biodiversity that measures similarity (and dissimilarity) in a population [26] is defined as

$$H = \sum_{i,j}^{n} d_{i,j}\, p_i p_j,$$

where $d_{i,j}$ is a measure of similarity (or dissimilarity). In the context of this paper, it basically measures how linguistically similar (or dissimilar) two individuals, who are randomly selected (with

[2] https://github.com/abarghi/Diversity_Using_Vegetarian_and_Tidyverse

replacement) from a population, are from each other. In particular, one uses linguistic family trees in each country, and find Rao's index of similarity using the following similarity function:

$$d_{i,j} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j \text{ and lanuage } i \text{ and } j \text{ belong to two different family trees;} \\ \frac{1}{d+1}, & \text{if } i \neq j \text{ and there is a path of length } d \text{ between language } i \text{ and } j. \end{cases}$$

Note that when $d_{i,i} = 1$ and $d_{i,j} = 0$ for $i \neq j$, we have Greenberg entropic index as a special case of Rao's entropy. The rationale behind this similarity function is that if two individual have the same first languages, their linguistic similarity is equal to one. However, if they have different first languages and these languages belong to different family trees, besides some common vocabulary, these languages have a similarity of zero. Lastly, if both languages belong to the same family tree, there is a level of similarity and this is measure by the inverse of length of the path connecting the two on the family tree.

Another future direction is to study different measures of linguistic diversity as time-series, similar to what Harmon and Loh have done using LDI [27]. It would be interesting to see how this approach can be applied using Rao's similarity measures, or other tree-based [28, 29] or distance-based methods [30] for regional linguistic studies, to make predictions about the status of more vulnerable languages in the future. Regarding tree-based methods, it should be pointed out that Rao's quadratic entropy is the basis for some of the approaches for computing phylogenetic diversities in ecology [31, 32, 33] and one possible direction is to use this available literature and apply it in linguistics. Moreover, in computing Rao's quadratic diversities, one can incorporate lexical or other linguistic similarities into the similarity (distance) functions.

6 Acknowledgments

I would like to thank the Office of the Dean of the Faculty at Saint Michael's College for the Merit-based Course Reduction Award in Fall 2020 that allowed me to continue my work on this paper.

References

[1] J. H. Greenberg. The measurement of linguistic diversity. *Language*, 32(1):109–115, 1956.

[2] E. H Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.

[3] L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[4] R. H MacArthur. Patterns of species diversity. *Biological Reviews*, 40(4):510–533, 1965.

[5] St. H Hurlbert. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4):577–586, 1971.

[6] R. K. Peet. The measurement of species diversity. *Annual Review of Ecology and Systematics*, pages 285–307, 1974.

[7] G. P. Patil and C. Taillie. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–561, 1982.

[8] J. Ginebra and X. Puig. On the measure and the estimation of evenness and diversity. *Computational Statistics & Data Analysis*, 54(9):2187–2201, 2010.

[9] R. H. Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological Monographs*, 30(3):279–338, 1960.

[10] R. H. Whittaker. Evolution and measurement of species diversity. *Taxon*, 21(2-3):213–251, 1972.

[11] D. M. Eberhard, G. F. Simons, and C. D. Fennig. Ethnologue: Languages of the world, 2019.

[12] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 4.6., 2022.

[13] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[14] M. O. Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[15] S. Lieberson. An extension of greenberg's linguistic diversity measures. *Language*, 40(4):526–531, 1964.

[16] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1):479–487, 1988.

[17] C. J. Keylock. Simpson diversity and the shannon–wiener index as special cases of a generalized entropy. *Oikos*, 109(1):203–207, 2005.

[18] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. University of California Press, Berkeley, California, USA, 1961.

[19] L. Jost. Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10):2427–2439, 2007.

[20] J. A. Veech and T. O. Crist. Diversity partitioning without statistical independence of alpha and beta. *Ecology*, 91(7):1964–1969, 2010.

[21] C. Ricotta. On beta diversity decomposition: trouble shared is not trouble halved. *Ecology*, 91(7):1981–1983, 2010.

[22] L. Jost. Independence of alpha and beta diversities. *Ecology*, 91(7):1969–1974, 2010.

[23] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Kongelige Danske Videnskabernes Selskab*, pages 1–34, 1948.

[24] L. R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[25] P. Jaccard. The distribution of the flora in the alpine zone. 1. *New Phytologist*, 11(2):37–50, 1912.

[26] C. R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

[27] D. Harmon and J. Loh. The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation & Conservation*, 4:97–151, 2010.

[28] M. R. Helmus, T. J. Bland, C. K. Williams, and A. R. Ives. Phylogenetic measures of biodiversity. *The American Naturalist*, 169(3):E68–E83, 2007.

[29] A. R. Ives and M. R. Helmus. Phylogenetic metrics of community similarity. *The American Naturalist*, 176(5):E128–E142, 2010.

[30] S. Champely and D. Chessel. Measuring biological diversity using euclidean metrics. *Environmental and Ecological Statistics*, 9(2):167–177, 2002.

[31] M. W. Cadotte, T. J. Davies, J. Regetz, S. W. Kembel, E. Clevand, and T. Oakley. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology Letters*, 13(1):96–105, 2010.

[32] A. Chao, C.-H. Chiu, and L. Jost. Phylogenetic diversity measures based on hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1558):3599–3609, 2010.

[33] C.-H. Chiu, L. Jost, and A. Chao. Phylogenetic beta diversity, similarity, and differentiation measures based on hill numbers. *Ecological Monographs*, 84(1):21–44, 2014.

A Tables

Country Code	Richness
IN	423
NP	123
PK	79
IR	64
BD	43
KZ	43
IL	42
TR	41
AF	36
AZ	34
UZ	33
TM	30
OM	28
KG	27
GE	26
AE	25
BT	25
TJ	25
SA	22
SY	22
IQ	21
QA	17
CY	14
LB	14
YE	14
BH	13
AM	11
LK	10
JO	9
PS	7
KW	5

Country Code	Shannon
IN	23.218441
AE	11.172717
BT	9.562095
NP	9.468600
QA	8.897871
IL	7.772443
PK	7.677709
OM	7.122931
AF	6.846279
IQ	5.980274
IR	5.786115
BH	4.446231
YE	3.635219
SA	3.277666
KZ	3.146846
UZ	3.086781
GE	2.921120
TM	2.878044
JO	2.865541
KG	2.737652
CY	2.551236
SY	2.488467
KW	2.466614
BD	2.365351
TR	2.202198
LK	2.086926
AZ	1.829342
TJ	1.829291
LB	1.795033
PS	1.714478
AM	1.164742

Country Code	Greenberg
IN	9.969579
QA	6.336755
AE	5.835562
BT	5.768101
AF	4.912150
IQ	4.250173
PK	4.152624
NP	4.078858
IL	3.911446
OM	3.689144
YE	3.022755
BH	2.755357
IR	2.632791
KZ	2.056855
JO	1.990195
UZ	1.865696
KG	1.846733
TM	1.842836
KW	1.821256
CY	1.793944
SA	1.752183
GE	1.742736
LK	1.621985
SY	1.589308
BD	1.584692
TR	1.435122
PS	1.424603
TJ	1.382107
LB	1.302633
AZ	1.262504
AM	1.051925

Table 1: Richness, Shannon, and Greenberg Effective Numbers in Descending Order Among These Thirty One Countries

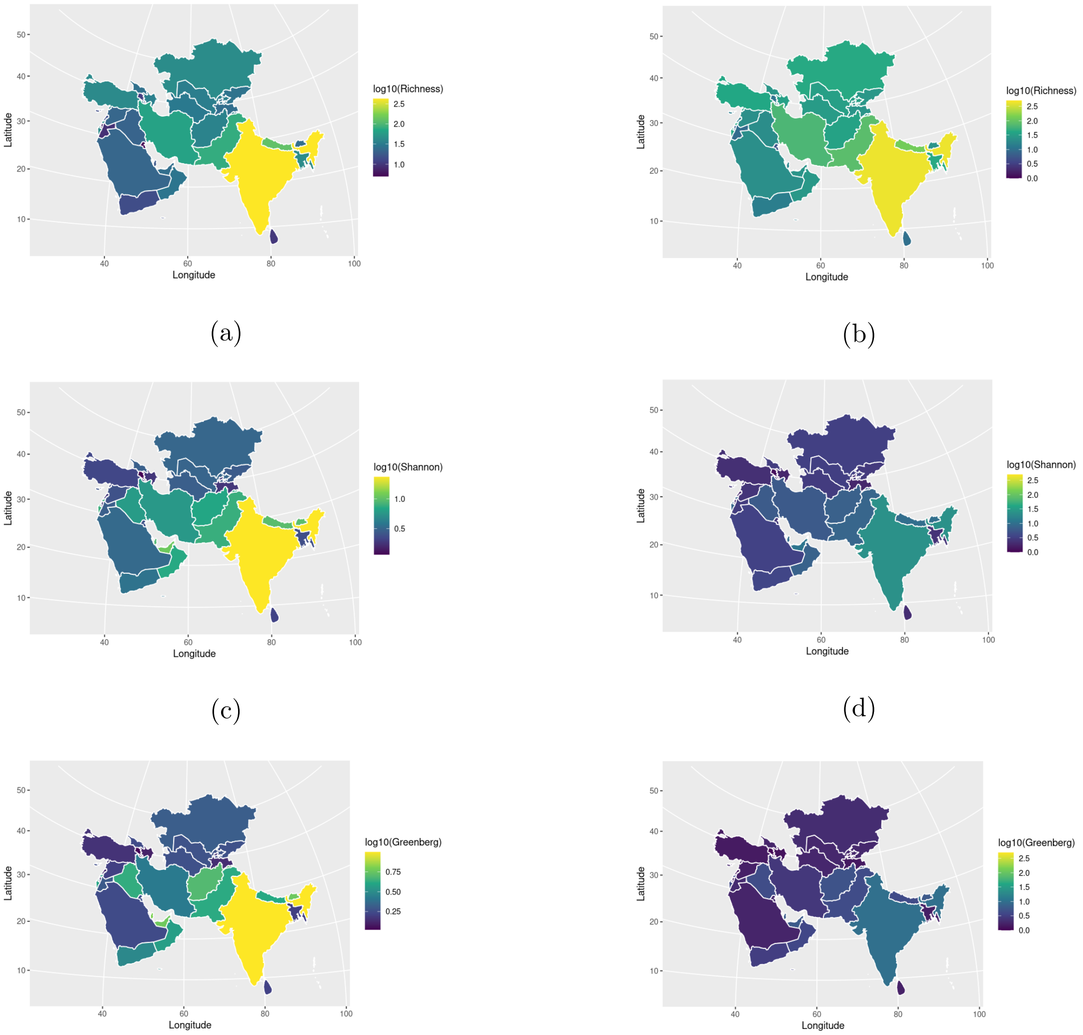
Country Code	Average Overall Rank
IN	1.000000
NP	4.666667
PK	5.666667
AE	7.000000
IL	7.333333
AF	7.666667
BT	8.000000
IR	9.333333
QA	9.666667
OM	10.333333
KZ	11.666667
IQ	12.333333
UZ	14.333333
TM	16.000000
YE	16.333333
BH	16.666667
KG	17.000000
BD	18.000000
GE	18.000000
SA	18.000000
TR	19.666667
JO	21.000000
CY	21.333333
SY	22.000000
AZ	22.333333
KW	24.333333
TJ	24.666667
LK	25.666667
LB	27.333333
PS	29.000000
AM	29.666667

Table 2: Average Overall Rank in Ascending Order Among These Thirty One Countries

Type	Method	Richness	Shannon	Greenberg
Gamma	Unweighted	790	59.5264	37.8207
Gamma	Weighted	790	48.4439	19.3729
Alpha	Unweighted	42.7742	3.85418	2.1383
Alpha	Weighted	42.7742	11.5658	8.5082
Beta	Unweighted	18.4690	15.4447	17.6870
Beta	Weighted	18.4691	4.1885	2.2767
MacArthur's Homogeneity	Unweighted	0.0542	0.0648	0.0565
MacArthur's Homogeneity	Weighted	–	–	–
Relative Homogeneity	Unweighted	0.0226	0.0336	0.0251
Relative Homogeneity	Weighted	–	0.0665	–

Table 3: Unweighted and Weighted Gamma, Alpha, and Beta Diversities, Along with MacArthur's Homogeneity and Relative Homogeneity, Using Richness, Shannon, and Greenberg

B Figures



(a)

(b)

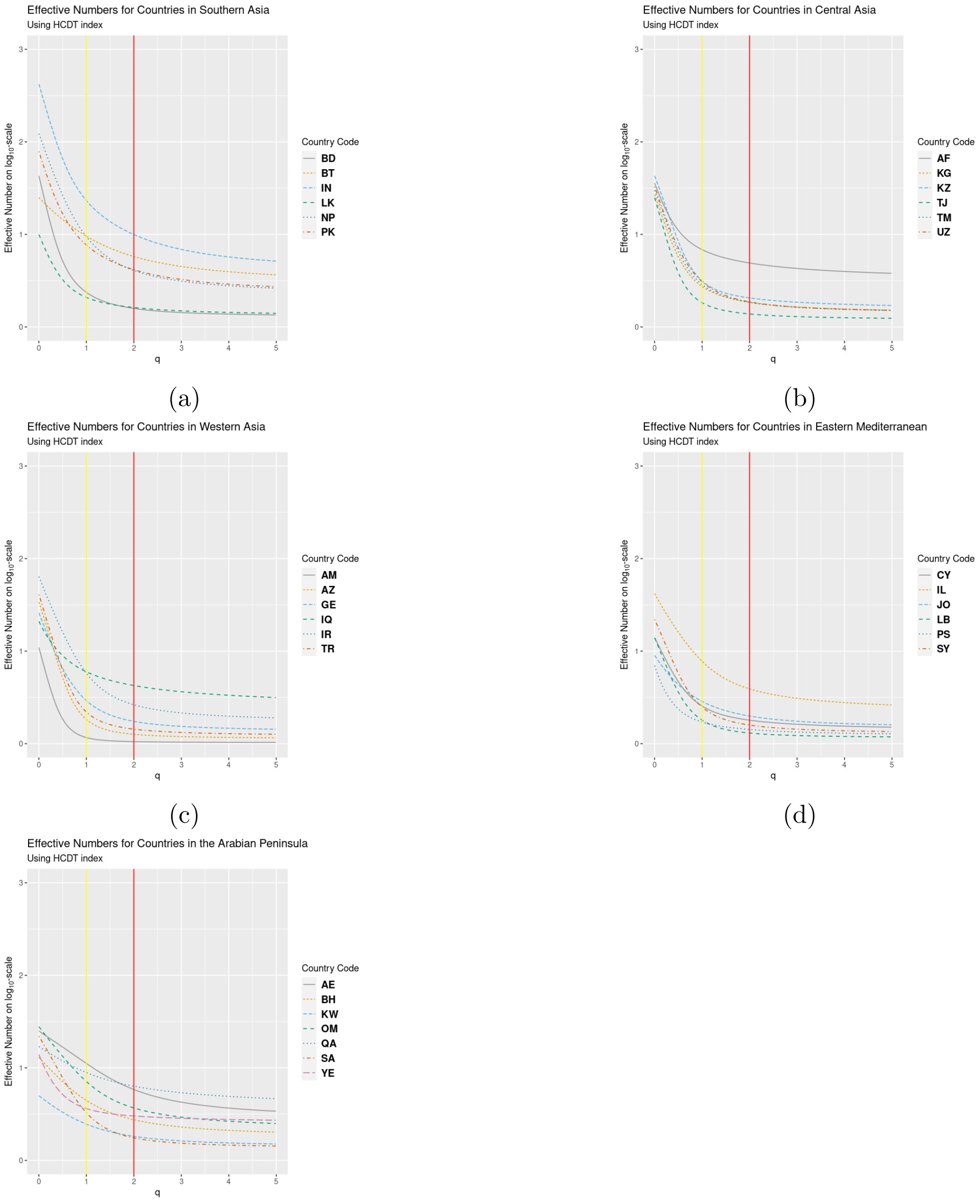
(c)

(d)

(e)

(f)

Figure 2: Richness, Shannon, and Greenberg Effective Numbers on Logarithmic Scale



(a)

(b)

(c)

(d)

(e)

Figure 3: Hill Numbers on Logarithmic Scale, for $0 \leq q \leq 5$, with Shannon and Greenberg Effective Numbers Represented by the Yellow and Red Vertical Lines, Respectively, Based on Region

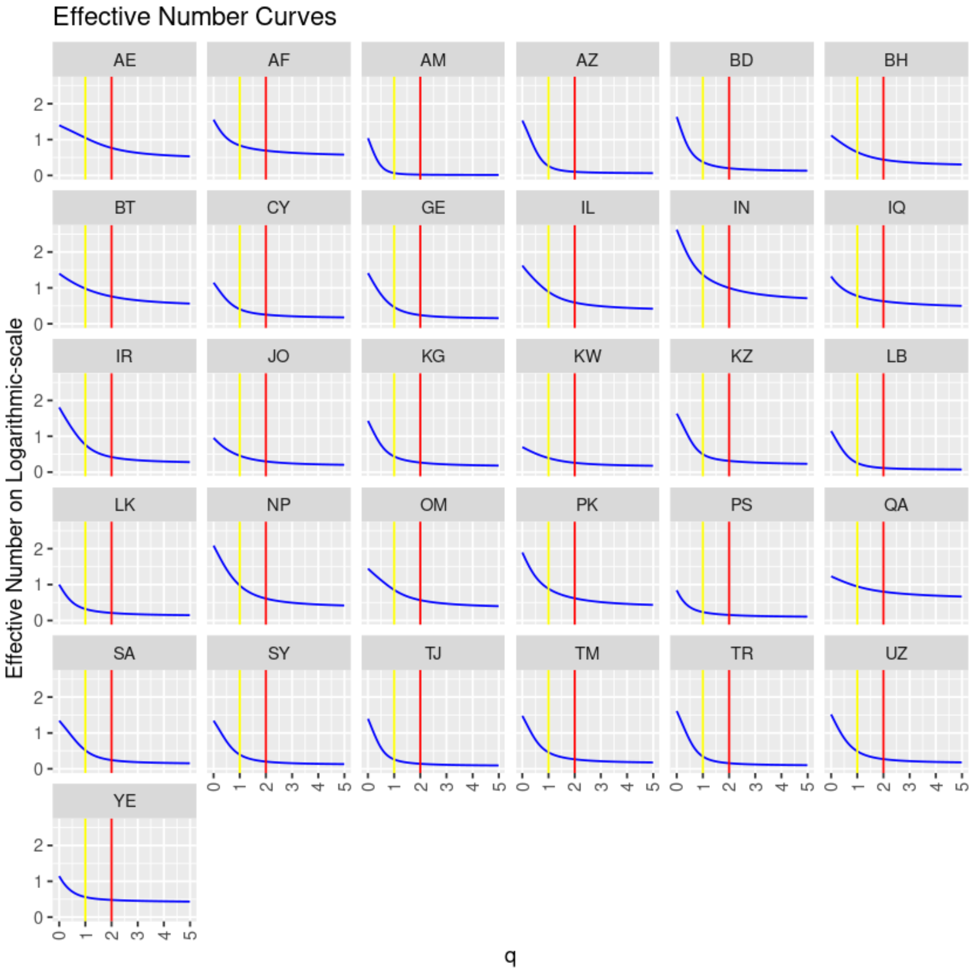
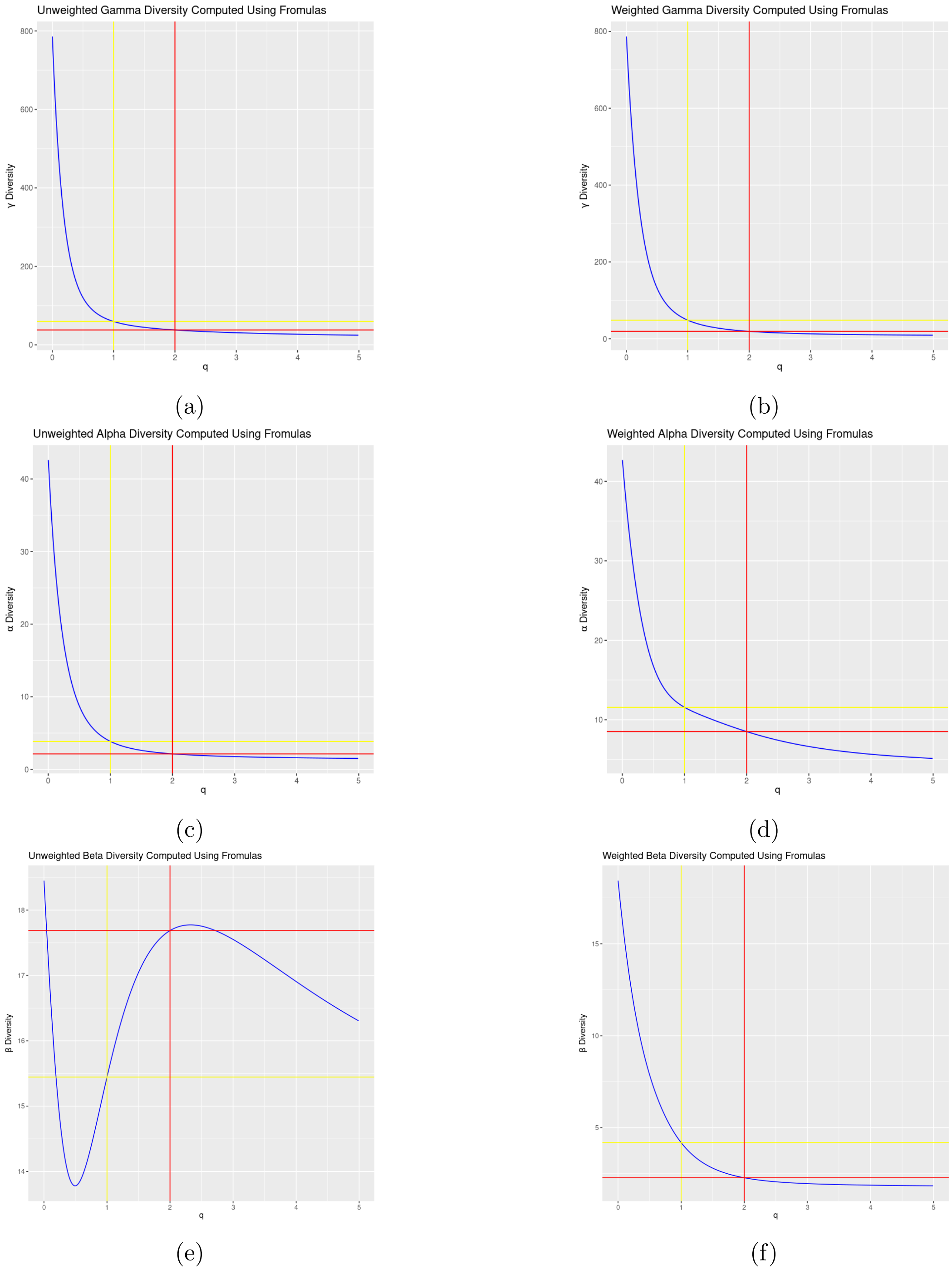
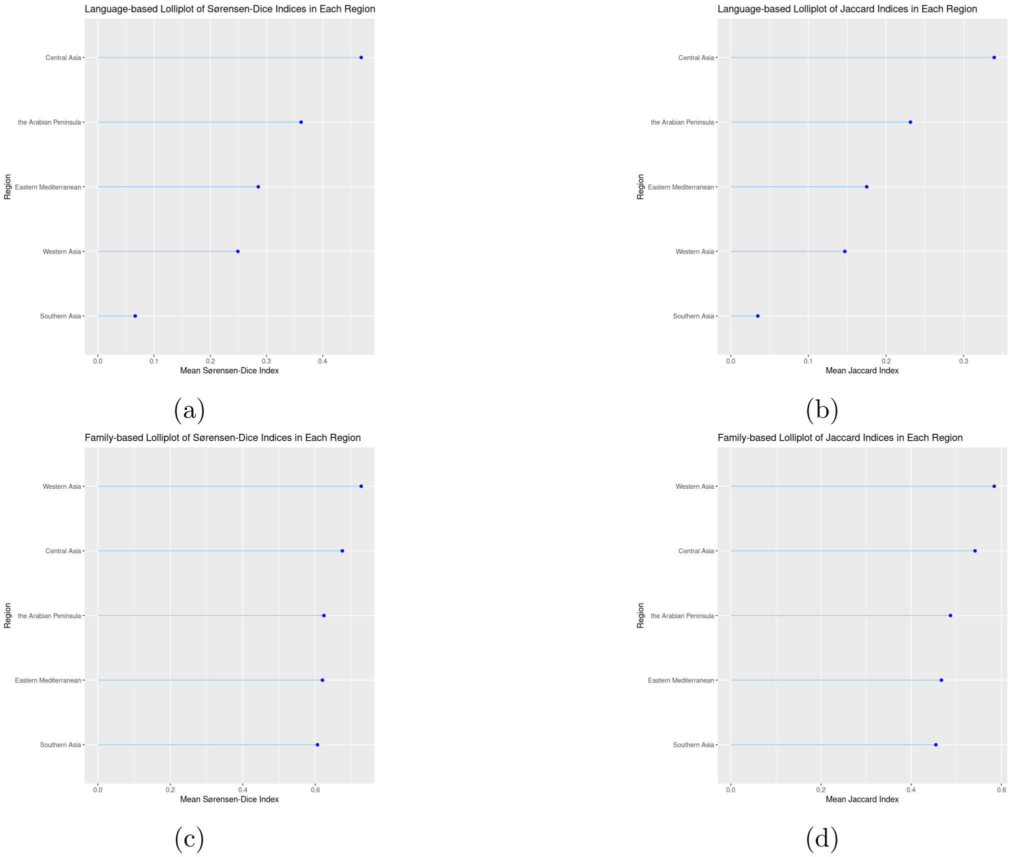


Figure 4: Hill Numbers on Logarithmic Scale, for $0 \leq q \leq 5$, with Shannon and Greenberg Effective Numbers Represented by the Yellow and Red Vertical Lines, Respectively



(a) (b) (c) (d) (e) (f)

Figure 5: Unweighted and Weighted Gamma, Alpha, and Beta Diversities Using Hill Numbers for $0 \leq q \leq 5$



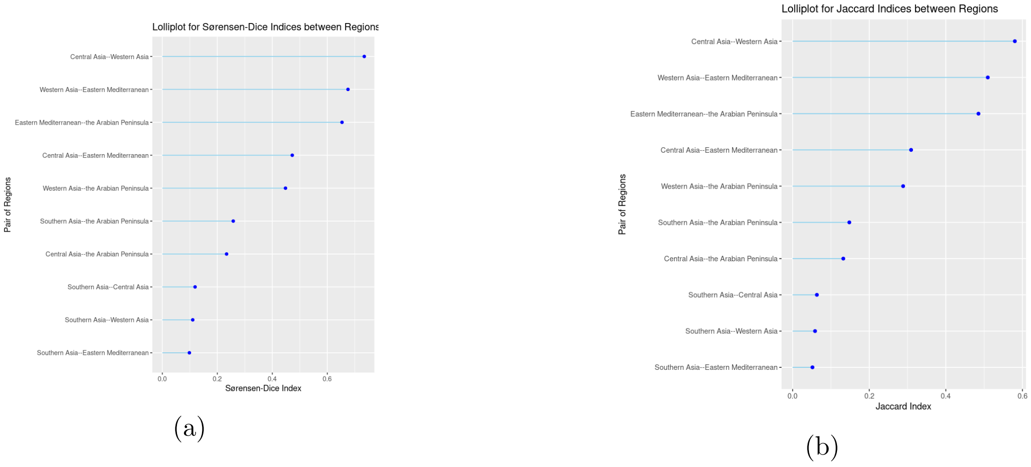
(a)

(b)

(c)

(d)

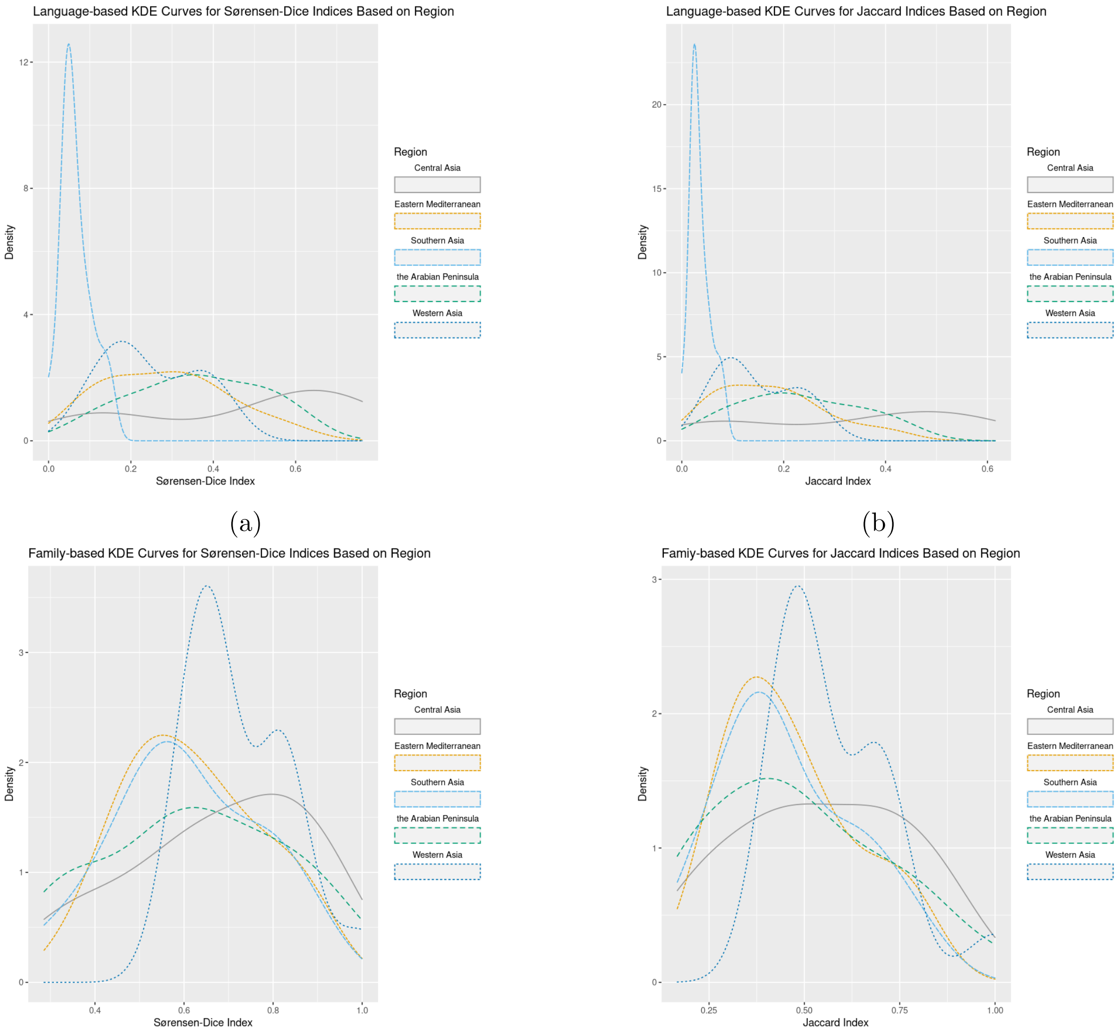
Figure 6: Mean Sørensen-Dice and Jaccard Indices within Regions



(a)

(b)

Figure 7: Mean Sørensen-Dice and Jaccard Indices between Pairs of Regions



(a)

(b)

(c)

(d)

Figure 8: Sørensen-Dice and Jaccard Indices within Regions