INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

A Generalized Index of Fractionalization

Walter Bossert, Conchita D'Ambrosio and Eliana Laferrara

Working Paper n. 313

July 2006

IGIER – Università Bocconi, Via Salasco 5, 20136 Milano –Italy
http://www.igier.uni-bocconi.it

A Generalized Index of Fractionalization*

Walter Bossert
Département de Sciences Economiques and CIREQ, Université de Montréal
walter.bossert@umontreal.ca

Conchita D'Ambrosio
Università di Milano-Bicocca and DIW Berlin
conchita.dambrosio@unibocconi.it

Eliana La Ferrara
Università Bocconi and IGIER
eliana.laferrara@unibocconi.it

This version July 2006

Abstract. The goal of this paper is to characterize a measure of diversity among individuals, which we call *generalized fractionalization index,* that uses information on similarities among individuals. We show that the generalized index is a natural extension of the widely used *ethno-linguistic fractionalization index* and is also simple to compute. The paper offers some empirical illustrations on how the new index can be operationalized and what difference it makes as compared to standard indices. These applications pertain to the pattern of diversity in the United States across states. *Journal of Economic Literature* Classification Nos.: C43, D63.

Keywords: Diversity, Similarity, Ethno-Linguistic Fractionalization.

*We are grateful to Itzhak Gilboa for extremely useful suggestions. We thank Vincent Buskens, Joan Esteban, Michele Pellizzari, Debraj Ray and seminar participants at CORE, Università di Milano, Università di Pavia, the 2005 Polarization and Conflict Workshop in Konstanz, the 2006 EURODIV Conference in Milano and the 2006 SCW Conference in Istanbul for helpful comments. Silvia Redaelli provided excellent research assistance. We also thank Università Bocconi for its hospitality during the preparation of this paper. Financial support from the Polarization and Conflict Project CIT-2-CT-2004-506084 funded by the European Commission-DG Research Sixth Framework Programme and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

1 Introduction

The traditional way of conceiving heterogeneity among individuals in Economics has been to think of income inequality, that is, individuals' differences in the command over economic resources. Many contributions have estimated the effects of inequality on all sorts of outcomes, and the literature on the measurement of inequality has proceeded on a parallel path, advancing to substantial degrees of sophistication. In recent times there has been a growing interest within Economics in the role that other types of heterogeneity, namely *ethnic or cultural diversity,* play in explaining socioeconomic outcomes. A number of empirical studies have found that ethnic diversity is associated with lower growth rates (Easterly and Levine, 1997), more corruption (Mauro, 1995), lower contributions to local public goods (Alesina, Baqir and Easterly, 1999), lower participation in groups and associations (Alesina and La Ferrara, 2000) and a higher propensity to form jurisdictions to sort into homogeneous groups (Alesina, Baqir and Hoxby, 2004). For an extensive review of these and other contributions on the relationship between ethnic diversity and economic performance, see Alesina and La Ferrara (2005). Yet the literature on the measurement of ethnic—and other forms of non-income related—heterogeneity has received considerably less attention.

The measure of ethnic diversity used almost universally in the empirical Economics literature is the so-called index of ethno-linguistic fractionalization (ELF), which is a decreasing transformation of the Herfindahl concentration index. In particular, if we consider a society composed of $K \geq 2$ different ethnic groups and let p_k indicate the share of group k in the total population, the resulting value of the ELF index is given by

$$1 - \sum_{k=1}^{K} p_k^2.$$

The popularity of this index in empirical applications can be attributed to two features. First, it is extremely simple to compute from micro as well as from aggregate data: all that is needed is the vector of shares of the various groups in the population. Second, ELF has a very intuitive interpretation: it measures the probability that two randomly drawn individuals from the overall population belong to different ethnic groups. On the other hand, the economic underpinnings for the use of this index seem underdeveloped. One of the few contributions that address this issue is Vigdor (2002), who proposes a behavioral interpretation of ELF in a model where individuals display differential altruism. He assumes that an individual's willingness to spend on local public goods depends partly on the benefits that other members of the community derive from the good, and that the weight of this altruistic component varies depending on how many members of the community share the same ethnicity of that individual.

The implicit contention is often that members of different ethnic groups may have different preferences, and this would generate conflicts of interests in economic decisions. Also, to the extent that skill complementarities among different types are important, it is unlikely that simple population shares will capture them. Presumably, people of different ethnicities will feel differently about each other depending on how similar they are. If this is the rationale for including ethnic diversity effects, then measuring fractionalization purely as a function of population shares seems a severe limitation. Similarity between individuals could depend, for example, on income, educational background, employment status, just to mention a few possible relevant attributes. If preferences might be induced by these other characteristics, then considering similarities between individuals will give a better understanding of the potential conflict in economic decisions. Providing a measure of fractionalization that accounts for the degree of similarity among agents seems therefore a useful task.

The goal of this paper is to characterize a *generalized ethno-linguistic fractionalization index* ($GELF$) that takes as primitive the individuals and uses information on their similarities to measure fractionalization. We show that the generalized index is a natural extension of ELF. The paper offers some empirical illustrations on how $GELF$ can be operationalized and what difference its application makes as compared to the standard ELF index. These applications pertain to the pattern of fractionalization in the United States across states.

Our paper is related to several strands of the literature. First, it naturally relates to the above-mentioned literature on ethnic diversity and its economic effects. While the bulk of this literature does not focus on the specific issue of measurement, a few contributions do. As the majority of applications have used language as a proxy for ethnicity, some authors have criticized the use of ELF on the grounds that linguistic diversity may not correspond to ethnic diversity. Among these, Alesina, Devleeschauwer, Easterly, Kurlat and Wacziarg (2003) have proposed a classification into groups that combines information on language with information on skin color. Note that this approach differs from ours because it defines ethnic categories on the basis of two criteria (language and skin color) and then applies the ELF formula to the resulting number of groups. Other authors, in particular Fearon (2003), have criticized standard applications of ELF on the grounds that they would fail to account for the *salience* of ethnic distinctions in different contexts. For example, the same two ethnic groups may be allies in one country and opponents in another, and using simply their shares in the population would fail to capture this. We share Fearon's concerns on this point, and indeed we hope that our index can be a first step towards incorporating issues of salience in the measurement of fractionalization, albeit in a simplistic way. In particular, if one thinks that differences in income, or education, or any other measurable characteristic may be the reason why ethnicity matters only in certain contexts, our $GELF$ index already 'weighs' ethnic categories by their salience. Turning to the notion of 'distance' among ethnic groups, relatively little has been done. Using a heuristic approach, Laitin (2000) and Fearon (2003) rely on measures of distance between languages to assess how different linguistic groups are across countries. Caselli and Coleman (2002) stress the importance of ethnic distance in a theoretical model and propose to measure it using surveys of anthropologists.

Second, the paper relates to the literature on ethnic polarization. Montalvo and Reynal-Querol (2005) proposed an index of ethnic polarization, RQ, as a more appropriate measure of conflict than ELF itself. RQ aims at capturing the distance of the distribution of the ethnic groups from the bipolar distribution, which represents the highest level of polarization. Montalvo and Reynal-Querol (2005) also show that this index is highly correlated with ELF at low levels, uncorrelated at intermediate levels and negatively correlated at high levels. Desmet, Ortuño-Ortín and Weber (2005) focus on ethno-linguistic conflict that arises between a dominant central group and peripheral minority groups. They propose an index of peripheral ethno-linguistic diversity, PD, which can capture both the notion of diversity and of polarization. The relationship between these indices and $GELF$ is discusses in depth in Section 4.

Third, the measurement of diversity has been formally analyzed in different contexts within the Economics literature. For example, Weitzman (1992) suggests an index that is primarily intended to measure biodiversity. Moreover, the measurement of diversity has become an increasingly important issue in the recent literature on the ranking of opportunity sets in terms of freedom of choice, where opportunity sets are interpreted as sets of options available to a decision maker. Examples for such studies include Weitzman (1998), Pattanaik and Xu (2000), Nehring and Puppe (2002) and Bossert, Pattanaik and Xu (2003). A fundamental difference between the above-mentioned contributions and the approach followed in this paper is the informational basis employed which results in a very different set of axioms that are suitable for a measure of diversity. Both Weitzman's (1992) seminal paper and the literature on incorporating notions of diversity in the context of measuring freedom of choice proceed by constructing a ranking of *sets* of objects (interpreted as sets of species in the case of biodiversity and as sets of available options in the context of freedom of choice), whereas we operate in an informationally richer environment: not only whether a group is present may influence the measure of fractionalization, but also the relative population shares of these groups along with the pairwise similarities among them.

The remainder of the paper is organized as follows. In Section 2, we introduce the formal framework used in the paper. Section 3 contains our main theoretical result, namely, an axiomatic characterization of $GELF$. The relationships between $GELF$ and alternative measures that appear in the literature are discussed in Section 4. Section 5 provides some empirical illustrations and Section 6 concludes.

2 Similarity, fractionalization and some examples

The characterization result we provide in the present contribution is very general: we do not impose any assumptions regarding the partition of the population into groups. We believe that a measure of fractionalization of a society should take as primitive the individual and consider attributes such as ethnicity like any other personal characteristic in determining the similarity between individuals. In much of our informal discussion, however, we refer to ethnic groups in order to be in line with the strand of the literature to which we aim at contributing. Similarly, the empirical application makes also use of these ethnic groups for comparison purposes with more standard indices. But the way we think of the problem to be modelled is without such a predefined partition. Our starting point is a society composed of individuals with personal characteristics, whatever they might be. Any two individuals may be perfectly identical according to the characteristics under consideration, completely dissimilar or similar to different degrees. For simplicity, we normalize the similarity values to be in the interval $[0, 1]$, assign the value one to perfect similarity and a value of zero to maximum dissimilarity. If the society is composed of n individuals, the comparison process will generate n^2 similarity values. These values are collected in a matrix, the *similarity matrix.* Each row i of this matrix contains the similarity values of individual i with respect to all members of society. Naturally, all entries on the main diagonal of such a matrix—the entries representing the similarity of each individual to itself—are equal to one: each individual is perfectly similar (identical) to itself. Furthermore, a similarity matrix is symmetric: the similarity between individuals i and j is equal to that between j and i. It could be argued that similarity need not be symmetric particularly when based on subjective indicators. Our index can be characterized on a larger domain where the notion of similarity is not necessarily symmetric; see the Appendix for details. In the empirical section of this paper we focus on objective characteristics of individuals, thus in what follows we assume symmetry.

A plausible method of partitioning the individuals into groups is the following. Any two individuals i and j belong to the same group if the similarity between i and j is equal to one and, moreover, the similarities of i with respect to all other individuals k are the same as those of j. Using this process, a group partition emerges naturally from the similarity matrix without having to impose it in advance. This method has several advantages: (i) it releases the researcher of the choice of the one characteristic that determines fractionalization in the society of interest; (ii) it makes it possible to consider simultaneously multiple characteristics; (iii) it allows group formation across characteristics; (iv) it considers the intensity of similarities between groups.

We now define $GELF$ and use several examples to illustrate some important special cases, such as ELF. Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{R}$ the set of all real numbers. The set of all non-negative real numbers is $\mathbb{R}_+$ and the set of positive real numbers is $\mathbb{R}_{++}$. For $n \in \mathbb{N} \setminus \{1\}$, $\mathbb{R}^n$ is Euclidean n-space and Δ^n is the n-dimensional unit simplex. Furthermore, $\mathbf{0}^n$ is the vector consisting of n zeroes. A *similarity matrix of dimension* $n \in \mathbb{N} \setminus \{1\}$ is an $n \times n$ matrix $S = (s_{ij})_{i,j \in \{1,\ldots,n\}}$ such that:

(a) for all $i, j \in \{1, \ldots, n\}$, $s_{ij} \in [0, 1]$;
(b) for all $i \in \{1, \ldots, n\}$, $s_{ii} = 1$;
(c) for all $i, j \in \{1, \ldots, n\}$, $[s_{ij} = 1 \;\Rightarrow\; s_{ik} = s_{kj}$ for all $k \in \{1, \ldots, n\}]$.

The three restrictions on the elements of a similarity matrix have very intuitive interpretations. (a) is consistent with a normalization requiring that complete dissimilarity is assigned a value of zero and full similarity is represented by one. Clearly, this requires that each individual has a similarity value of one when assessing the similarity to itself, as stipulated in (b). Condition (c) requires that if two individuals are fully similar, it is not possible to distinguish between them as far as their similarity to others is concerned. Because $i = j$ is possible in (c), the conjunction of (b) and (c) implies that a similarity matrix is symmetric. Finally, (c) implies that full similarity is transitive in the sense that, if $s_{ij} = s_{ji} = s_{jk} = s_{kj} = 1$, then $s_{ik} = s_{ki} = 1$ for all $i, j, k \in \{1, \ldots, n\}$. Our characterization result remains valid if restriction (c) is dropped—that is, our index can be characterized on a larger domain where the notion of similarity is not necessarily symmetric, as may be the case if the similarity values are obtained from people's subjective views on the degree to which they differ from others. We state our main result with restriction (c) to emphasize that we do not need non-symmetric similarity matrices and, thus, our charcaterization is not dependent on an artificially large domain. See the Appendix for details.

Let $\mathcal{S}^n$ be the set of all n-dimensional similarity matrices, where $n \in \mathbb{N} \setminus \{1\}$. We use I^n to denote the $n \times n$ identity matrix and $\mathbf{1}^n$ to denote the $n \times n$ matrix all of whose entries are equal to one. Clearly, both of these matrices are in $\mathcal{S}^n$, and they represent extreme cases within this class. I^n can be thought of as having maximal diversity: any two individuals are completely dissimilar and, therefore, each individual is in a group by itself. $\mathbf{1}^n$, on the other hand, represents maximal concentration (and, thus, minimal diversity) because there is but a single group in the population all members of which are fully similar.

We let $\mathcal{S} = \cup_{n \in \mathbb{N} \setminus \{1\}} \mathcal{S}^n$, and a *diversity measure* is a function $D \colon \mathcal{S} \to \mathbb{R}_+$. The measure we suggest in this paper is what we call the *generalized ethno-linguistic fractionalization* ($GELF$) index G. It is defined by

$$G(S) = 1 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \tag{1}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$ (or any positive multiple; clearly, multiplying the index value by $\alpha \in \mathbb{R}_{++}$ leaves all diversity comparisons unchanged). $GELF$ is the expected dissimilarity between two individuals drawn at random.

As an example, suppose a three-dimensional similarity matrix is given by

$$S = \begin{pmatrix} 1 & 1/2 & 1/4 \\ 1/2 & 1 & 0 \\ 1/4 & 0 & 1 \end{pmatrix}.$$

The corresponding value of G is given by

$$G(S) = 1 - \frac{1}{9}\left[1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{2} + 1 + 0 + \frac{1}{4} + 0 + 1\right] = 1 - \frac{1}{2} = \frac{1}{2}.$$

Before providing a characterization of our new index, we illustrate that it is indeed a generalization of the commonly-employed ethno-linguistic fractionalization (ELF) index. The application of ELF is restricted to an environment where the only information available is the vector $p = (p_1, \ldots, p_K) \in \Delta^K$ of population shares for $K \in \mathbb{N}$ predefined groups. No partial similarity values are taken into consideration—individuals are either fully similar or completely dissimilar, that is, s_{ij} can assume the values one and zero only. Letting $\Delta = \cup_{K \in \mathbb{N}} \Delta^K$, the ELF index $E \colon \Delta \to \mathbb{R}_+$ is defined by letting

$$E(p) = 1 - \sum_{k=1}^{K} p_k^2$$

for all $K \in \mathbb{N}$ and for all $p \in \Delta^K$. Thus, ELF is one minus the well-known Herfindahl index of concentration.

In our setting, the ELF environment can be described by a subset $\mathcal{S}_{01} = \cup_{n \in \mathbb{N} \setminus \{1\}} \mathcal{S}_{01}^n$ of our class of similarity matrices where, for all $n \in \mathbb{N} \setminus \{1\}$, for all $S \in \mathcal{S}_{01}^n$ and for all $i, j \in \{1, \ldots, n\}$, $s_{ij} \in \{0, 1\}$. By properties (b) and (c), it follows that, within this subclass of matrices, the population $\{1, \ldots, n\}$ can be partitioned into $K \in \mathbb{N}$ non-empty and disjoint subgroups $N_1, \ldots, N_K$ with the property that, for all $i, j \in \{1, \ldots, n\}$,

$$s_{ij} = \begin{cases} 1 & \text{if there exists } k \in \{1, \ldots, K\} \text{ such that } i, j \in N_k; \\ 0 & \text{otherwise.} \end{cases}$$

Letting $n_k \in \mathbb{N}$ denote the cardinality of N_k for all $k \in \{1, \ldots, K\}$, it follows that $\sum_{k=1}^K n_k = n$ and $p_k = n_k/n$ for all $k \in \{1, \ldots, K\}$. For $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}_{01}^n$, we obtain

$$G(S) = 1 - \frac{1}{n^2} \sum_{k=1}^K n_k^2 = 1 - \sum_{k=1}^K p_k^2 = E(p).$$

For example, suppose that

$$S = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

that is, we are analyzing a society composed of three individuals. Two of them (individuals 1 and 2) are fully similar: the similarity values s_{12} and s_{21} are equal to one and, furthermore, they have the same degree of similarity—zero—with respect to the remaining member of society (individual 3). Because individual 3 is not completely similar to anyone else, it forms a group on its own. The corresponding value of G is given by

$$G(S) = 1 - \frac{1}{9}[1 + 1 + 0 + 1 + 1 + 0 + 0 + 0 + 1] = 1 - \frac{5}{9} = \frac{4}{9}.$$

Because $S \in \mathcal{S}_{01}^3$, we can alternatively calculate this diversity value using ELF. We have $K = 2$, $N_1 = \{1, 2\}$, $N_2 = \{3\}$, $p_1 = 2/3$ and $p_2 = 1/3$. Thus,

$$E(p) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \frac{5}{9} = \frac{4}{9} = G(S).$$

A second special case allows us to obtain population subgroups endogenously from similarity matrices even if similarity values can assume values other than zero and one. To do so, we define a partition of $\{1, \ldots, n\}$ into $K \in \mathbb{N}$ non-empty and disjoint subgroups $N_1, \ldots, N_K$. By properties (b) and (c), these subgroups are such that, for all

$k \in \{1, \ldots, K\}$, for all $i, j \in N_k$ and for all $h \in \{1, \ldots, n\}$, $s_{ij} = s_{ji} = 1$ and $s_{ih} = s_{hi} = s_{hj} = s_{jh}$. Thus, for all $k, \ell \in \{1, \ldots, K\}$, we can unambiguously define $\overline{s}_{k\ell} = s_{ij}$ for some $i \in N_k$ and some $j \in N_\ell$. Again using $n_k \in \mathbb{N}$ to denote the cardinality of N_k for all $k \in \{1, \ldots, K\}$, it follows that $\sum_{k=1}^{K} n_k = n$ and $p_k = n_k/n$ for all $k \in \{1, \ldots, K\}$. For $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}^n$, we obtain

$$G(S) = 1 - \frac{1}{n^2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} n_k n_\ell \overline{s}_{k\ell} = 1 - \sum_{k=1}^{K} \sum_{\ell=1}^{K} p_k p_\ell \overline{s}_{k\ell}. \tag{2}$$

Clearly, the ELF index E is obtained for the case where all off-diagonal entries of S are equal to zero.

To provide a numerical illustration of this case, let

$$S = \begin{pmatrix} 1 & 1 & 1/2 \\ 1 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix},$$

that is, we consider another society of three individuals. Again, two of them (individuals 1 and 2) are fully similar: the similarity values s_{12} and s_{21} are equal to one and, furthermore, they have the same degree of similarity with respect to the remaining member of society (individual 3). This time, however, the similarity between the members of the first group and the remaining individual is equal to $1/2$ rather than zero. Individual 3 is not completely similar to anyone, thus is in a group by itself. The corresponding index value is

$$G(S) = 1 - \frac{1}{9}\left[1 + 1 + \frac{1}{2} + 1 + 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1\right] = 1 - \frac{7}{9} = \frac{2}{9}.$$

According to the method outlined above, we can alternatively partition the population $\{1, 2, 3\}$ into two groups $N_1 = \{1, 2\}$ and $N_2 = \{3\}$. The population shares of these groups are $p_1 = 2/3$ and $p_2 = 1/3$. We obtain the intergroup similarity values $\overline{s}_{11} = \overline{s}_{22} = s_{11} = s_{22} = s_{12} = s_{21} = 1$ and $\overline{s}_{12} = \overline{s}_{21} = s_{i3} = s_{3i} = 1/2$ for $i \in \{1, 2\}$, which leads to the index value

$$G(S) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}\right] = 1 - \frac{7}{9} = \frac{2}{9}.$$

3 A characterization of $GELF$

We now turn to a characterization of $GELF$. Our first axiom is a straightforward normalization property. It requires that the value of D at $\mathbf{1}^n$ is equal to zero and the value of D at I^n is positive for all $n \in \mathbb{N} \setminus \{1\}$. Given that the matrix $\mathbf{1}^n$ is associated with minimal diversity, it is a very plausible restriction to require that D assumes its minimal value for these matrices. Note that this minimal value is the same across population sizes. This is plausible because, no matter what the population size n might be, there is but a single group of perfectly similar individuals and, thus, there is no diversity at all. In contrast, it would be much less natural to require that the value of D at I^n be identical for all population sizes n. It is quite plausible to argue that having more distinct groups each of which consists of a single individual leads to more fractionalization than a situation where there are fewer groups containing one individual each. Thus, we obtain the following axiom.

Normalization. For all $n \in \mathbb{N} \setminus \{1\}$,

$$D(\mathbf{1}^n) = 0 \quad \text{and} \quad D(I^n) > 0.$$

Our second axiom is very uncontroversial as well. It requires that individuals are treated impartially, paying no attention to their identities. For $n \in \mathbb{N} \setminus \{1\}$, let Π^n be the set of permutations of $\{1, \ldots, n\}$, that is, the set of bijections $\pi \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. For $n \in \mathbb{N} \setminus \{1\}$, $S \in \mathcal{S}^n$ and $\pi \in \Pi^n$, S_π is obtained from S by permuting the rows and columns of S according to π. Anonymity requires that D is invariant with respect to permutations.

Anonymity. For all $n \in \mathbb{N} \setminus \{1\}$, for all $S \in \mathcal{S}^n$ and for all $\pi \in \Pi^n$,

$$D(S_\pi) = D(S).$$

Many social index numbers have an additive structure. Additivity entails a separability property: the contribution of any variable to the overall index value can be examined in isolation, without having to know the values of the other variables. Thus, additivity properties are often linked to independence conditions of various forms. The additivity property we use is standard except that we have to respect the restrictions imposed by the definition of $\mathcal{S}^n$. In particular, we cannot simply add two similarity matrices S and T of dimension n because, according to ordinary matrix addition, all entries on the diagonal of the sum $S + T$ will be equal to two rather than one and, therefore, $S + T$ is not an element of $\mathcal{S}^n$. For that reason, we define the following operation $\oplus$ on the sets $\mathcal{S}^n$ by letting, for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S, T \in \mathcal{S}^n$, $S \oplus T = (s_{ij} \oplus t_{ij})_{i,j \in \{1,\ldots,n\}}$ with

$$s_{ij} \oplus t_{ij} = \begin{cases} 1 & \text{if } i = j; \\ s_{ij} + t_{ij} & \text{if } i \neq j. \end{cases}$$

The standard additivity axiom has to be modified in another respect. Because the diagonal is unchanged when moving from S and T to $S \oplus T$, it would be questionable to require the value of D at $S \oplus T$ to be given by the sum of $D(S)$ and $D(T)$ because, in doing so, we would double-count the diagonal elements in S and in T. Therefore, this sum has to be corrected by the value of D at I^n, and we obtain the following axiom.

Additivity. For all $n \in \mathbb{N} \setminus \{1\}$ and for all $S, T \in \mathcal{S}^n$ such that $(S \oplus T) \in \mathcal{S}^n$,

$$D(S \oplus T) = D(S) + D(T) - D(I^n).$$

With the partial exception of the normalization condition (which implies that our diversity measure assumes the same value for the matrix $\mathbf{1}^n$ for all population sizes n), the first three axioms apply to diversity comparisons involving fixed population sizes only. Our last axiom imposes restrictions on comparisons across population sizes. We consider specific replications and require the index to be invariant with respect to these replications. The scope of the axiom is limited to what we consider clear-cut cases and, therefore, represents a rather mild variable-population requirement. In particular, consider the n-dimensional identity matrix I^n. As argued before, this matrix represents an extreme degree of diversity: each individual is in a group by itself and shares no similarities with anyone else. Now consider a population of size nm where there are m copies of each individual $i \in \{1, \ldots, n\}$ such that, within any group of m copies, all similarity values are equal to one and all other similarity values are equal to zero. Thus, this particular replication has the effect that, instead of n groups of size one that do not have any similarity to other groups, now we have n groups each of which consists of m identical individuals and, again, all other similarity values are equal to zero. As before, the population is divided into n homogeneous groups of equal size. Adopting a relative notion of diversity, it would seem natural to require that diversity has not changed as a consequence of this replication. To provide a precise formulation of the resulting axiom, we use the following notation. For $n, m \in \mathbb{N} \setminus \{1\}$, we define the matrix $R^n_m = (r_{ij})_{i,j \in \{1,\ldots,nm\}} \in \mathcal{S}^{nm}$ by

$$r_{ij} = \begin{cases} 1 & \text{if } \exists h \in \{1, \ldots, n\} \text{ such that } i, j \in \{(h-1)m + 1, \ldots, hm\}; \\ 0 & \text{otherwise.} \end{cases}$$

Now we can define our replication invariance axiom.

Replication invariance. For all $n, m \in \mathbb{N} \setminus \{1\}$,

$$D(R^n_m) = D(I^n).$$

These four axioms characterize $GELF$.

Theorem 1 *A diversity measure $D\colon \mathcal{S} \to \mathbb{R}_+$ satisfies normalization, anonymity, additivity and replication invariance if and only if D is a positive multiple of G.*

Proof. That any positive multiple of G satisfies the axioms is straightforward to verify. Conversely, suppose D is a diversity measure satisfying normalization, anonymity, additivity and replication invariance. Let $n \in \mathbb{N} \setminus \{1\}$, and define the set $\mathcal{X}^n \subseteq \mathbb{R}^{n(n-1)/2}$ by

$$\begin{aligned} \mathcal{X}^n \quad = \quad \{x = (x_{ij})_{\substack{i \in \{1,\ldots,n-1\} \\ j \in \{i+1,\ldots,n\}}} \mid \exists S \in \mathcal{S}^n \text{ such that } s_{ij} = x_{ij} \text{ for all } i \in \{1,\ldots,n-1\} \\ \text{and for all } j \in \{i+1,\ldots,n\}\}. \end{aligned}$$

Define the function $F^n\colon \mathcal{X}^n \to \mathbb{R}$ by letting, for all $x \in \mathcal{X}^n$,

$$F^n(x) = D(S) - D(I^n) \tag{3}$$

where $S \in \mathcal{S}^n$ is such that $s_{ij} = x_{ij}$ for all $i \in \{1,\ldots,n-1\}$ and for all $j \in \{i+1,\ldots,n\}$. This function is well-defined because $\mathcal{S}^n$ contains symmetric matrices with ones on the main diagonal only. Because D is bounded below by zero, it follows that F^n is bounded below by $-D(I^n)$. Furthermore, the additivity of D implies that F^n satisfies Cauchy's basic functional equation

$$F^n(x + y) = F^n(x) + F^n(y) \tag{4}$$

for all $x, y \in \mathcal{X}^n$ such that $(x + y) \in \mathcal{X}^n$; see Aczél (1966, Section 2.1). We have to address a slight complexity in solving this equation because the domain $\mathcal{X}^n$ of F^n is not a Cartesian product, which is why we provide a few further details rather than invoking the corresponding standard result immediately.

Fix $i \in \{1,\ldots,n-1\}$ and $j \in \{i+1,\ldots,n\}$, and define the function $f^n_{ij}\colon [0,1] \to \mathbb{R}$ by

$$f^n_{ij}(x_{ij}) = F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$$

for all $x_{ij} \in [0,1]$, where the vector $(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$ is such that the component corresponding to ij is given by x_{ij} and all other entries (if any) are equal to zero. Note that this vector is indeed an element of $\mathcal{X}^n$ and, therefore, f_{ij}^n is well-defined. The function f_{ij}^n is bounded below because F^n is and, as an immediate consequence of (4), it satisfies the Cauchy equation

$$f_{ij}^n(x_{ij}+y_{ij}) = f_{ij}^n(x_{ij}) + f_{ij}^n(y_{ij}) \tag{5}$$

for all $x_{ij}, y_{ij} \in [0,1]$ such that $(x_{ij}+y_{ij}) \in [0,1]$. Because the domain of f_{ij}^n is an interval containing the origin and f_{ij}^n is bounded below, the only solutions to (5) are linear functions; see Aczél (1966, Section 2.1). Thus, there exists $c_{ij}^n \in \mathbb{R}$ such that

$$F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1}) = f_{ij}^n(x_{ij}) = c_{ij}^n x_{ij} \tag{6}$$

for all $x_{ij} \in [0,1]$.

Let $S \in \mathcal{S}^n$. By additivity, the definition of F^n and (6),

$$F^n\left((s_{ij})_{\substack{i\in\{1,\ldots,n-1\}\\ j\in\{i+1,\ldots,n\}}}\right) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} F^n(s_{ij}; \mathbf{0}^{n(n-1)/2-1}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f_{ij}^n(s_{ij}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij}$$

and, defining $d^n = D(I^n)$ and substituting into (3), we obtain

$$D(S) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij} + d^n. \tag{7}$$

Now fix $i,k \in \{1,\ldots,n-1\}$, $j \in \{i+1,\ldots,n\}$ and $\ell \in \{k+1,\ldots,n\}$, and let $S \in \mathcal{S}^n$ be such that $s_{ij} = s_{ji} = 1$ and all other off-diagonal entries of S are equal to zero. Let the bijection $\pi \in \Pi^n$ be such that $\pi(i) = k$, $\pi(j) = \ell$, $\pi(k) = i$, $\pi(\ell) = j$ and $\pi(h) = h$ for all $h \in \{1,\ldots,n\} \setminus \{i,j,k,\ell\}$. By (7), we obtain

$$D(S) = c_{ij}^n + d^n \quad \text{and} \quad D(S_\pi) = c_{k\ell}^n + d^n,$$

and anonymity implies $c_{ij}^n = c_{k\ell}^n$. Therefore, there exists $c^n \in \mathbb{R}$ such that $c_{ij}^n = c^n$ for all $i \in \{1,\ldots,n-1\}$ and for all $j \in \{i+1,\ldots,n\}$, and substituting into (7) yields

$$D(S) = c^n \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij} + d^n$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Normalization requires

$$D(\mathbf{1}^n) = c^n \frac{n(n-1)}{2} + d^n = 0$$

and, therefore, $d^n = -c^n n(n-1)/2$ for all $n \in \mathbb{N} \setminus \{1\}$. Using normalization again, we obtain

$$D(I^n) = -c^n \frac{n(n-1)}{2} > 0$$

which implies $c^n < 0$ for all $n \in \mathbb{N} \setminus \{1\}$. Thus,

$$D(S) = c^n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} - c^n \frac{n(n-1)}{2} \tag{8}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Let n be an even integer greater than or equal to four. By replication invariance and (8),

$$D(R^2_{n/2}) = c^n \frac{n}{2}\left(\frac{n}{2} - 1\right) - c^n \frac{n(n-1)}{2} = -c^2 = D(I^2).$$

Solving, we obtain

$$c^n = 4\frac{c^2}{n^2}. \tag{9}$$

Now let n be an odd integer greater than or equal to three. Thus, $q = 2n$ is even, and the above argument implies

$$c^q = 4\frac{c^2}{q^2} = \frac{c^2}{n^2}. \tag{10}$$

Furthermore, replication invariance requires

$$D(R^n_2) = D(R^{q/2}_2) = c^q \frac{q}{2} - c^q \frac{q(q-1)}{2} = -c^n \frac{n(n-1)}{2} = D(I^n).$$

Solving for c^n and using the equality $q = 2n$, it follows that $c^n = 4c^q$ and, combined with (10), we obtain (9) for all odd $n \in \mathbb{N} \setminus \{1\}$ as well.

Substituting into (8), simplifying and defining $\beta = -2c^2 > 0$, it follows that, for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$,

$$\begin{aligned}
D(S) &= 4\frac{c^2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} - 2\frac{c^2}{n^2} n(n-1) \\
&= 2\frac{c^2}{n^2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij} - 2c^2 + 2\frac{c^2}{n} \\
&= -2c^2 \left[1 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij} - \frac{1}{n} \right] \\
&= -2c^2 \left[1 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \right] \\
&= \beta G(S). \blacksquare
\end{aligned}$$

4 Alternative and related approaches

In this section we discuss the differences between $GELF$ and related indices proposed in various literatures. We start briefly with the Linguistics and Statistics literature and compare $GELF$ with Greenberg's (1956) index and with the quadratic entropy index (QE); we continue with the Economics literature with the indices of ethnic polarization (RQ) and peripheral diversity (PD).

What is known in the Economics literature as ELF is, in the Statistics literature, the *Gini-Simpson* index, introduced first by Gini (1912) and then by Simpson (1949) as a measure of diversity of the multinomial distribution. The same index has been proposed by Greenberg (1956) termed as the "A index". In the same article, Greenberg suggested a way to incorporate the degree of *resemblance* among K languages. Letting $\tau_{k\ell} \geq 0$ be the resemblance between language k and ℓ, the proposed index B is given by:

$$B = 1 - \sum_{k=1}^{K} \sum_{\ell=1}^{K} p_k p_\ell \tau_{k\ell}.$$

In an independent contribution, Rao (1982) suggested the same generalization of ELF, the *quadratic entropy* index (QE), in order to take into account different distance values, $d_{k\ell} \geq 0$, of different pairs of categories, k and ℓ. Rao (1984) and Rao and Nayak (1985) provide various axiomatizations of the measure. QE is an index that, rewritten in the setting of our paper, considers distances other than zero and one between individuals belonging to different groups, that is,

$$QE = \sum_{k=1}^{K} \sum_{\ell=1}^{K} p_k p_\ell d_{k\ell}.$$

Letting $d_{k\ell} = 1 - \overline{s}_{k\ell}$, $GELF$ is equal to QE, and hence B, when the population is partitioned exogenously (ex-ante) into groups on the basis of a characteristic, usually ethnicity.

$GELF$ is the expected distance between two individuals drawn at random. ELF can be interpreted as one minus a weighted sum of population shares p_k, where the weights are these shares themselves. $GELF$, on the other hand, is its natural generalization: when the population is partitioned exogenously, $GELF$ as well can be written as one minus a weighted sum of the population shares. However, the weight assigned to p_k is now not merely p_k itself but a considerably more refined expression that takes account of the similarities of the group members to the individuals in other groups. In calculating $GELF$, each individual counts in two capacities. Through its membership in its *own* group, an individual contributes to the population share of the group. In addition, there is a secondary contribution via the similarities to individuals of *other* groups.

Clearly, when the distance values are differences in income, QE is twice the well-known absolute Gini coefficient. The latter, when normalized by mean income, is one of the most popular indices of income inequality.

In Economics, the index of ethnic polarization RQ (see Montalvo and Reynal-Querol, 2005) shares a structure similar to that of ELF and of $GELF$. It is defined by

$$RQ = 1 - \sum_{k=1}^{K} \left(\frac{1/2 - p_k}{1/2} \right)^2 p_k.$$

As is the case for ELF, RQ employs a weighted sum of population shares. The weights employed in RQ capture the deviation of each group from the maximum polarization share $1/2$ as a proportion of $1/2$. Analogously to ELF, underlying that formula is the implicit assumption that any two groups are either completely similar or completely dissimilar and, thus, the weights depend on population shares only.

The index of peripheral diversity PD (see Desmet, Ortuño-Ortín and Weber, 2005) is a specification of the original Esteban and Ray (1994) polarization index. It is derived from the alienation-identification framework proposed by Esteban and Ray (1994), applied to distances between languages spoken rather than to income distances as in Esteban and Ray (1994). Desmet, Ortuño-Ortín and Weber (2005) distinguish between the effective alienation felt by the dominant group and that of the minorities. In particular, expressed in the setting of our paper, the index is defined by

$$PD = \sum_{k=1}^{K} \left[p_k^{1+\alpha} \left(1 - \overline{s}_{0k}\right) + p_k p_0^{1+\alpha} \left(1 - \overline{s}_{0k}\right) \right],$$

where $\alpha \in \mathbb{R}$ is a parameter indicating the importance given to the identification component, 0 is the dominant group and the other K are minority groups. When $\alpha < 0$, PD is an index of peripheral diversity; when $\alpha > 0$, PD is an index of peripheral polarization. The structure of this index is different from that of those previously discussed. As is the case for $GELF$, it does incorporate a notion of dissimilarity between groups, given by the complement to one of the similarity value. On the other hand, as opposed to the previous indices, the identification component plays a crucial role enhancing (when $\alpha > 0$) or diminishing (when $\alpha < 0$) the alienation produced by distances between groups. An additional difference to the other indices discussed in this section is the distinction between the dominant groups and the minorities.

5 An empirical illustration

In this section we provide an application of *GELF* to the pattern of diversity in the United States across states. Our goal is to compare the extent of diversity across states taking into account different dimensions of similarity among individuals, in particular: racial identity, household income, education and employment status of the head of the household.

5.1 Data and methodology

The data set used is the 5 percent IPUMS from the 1990 Census. We use individual level information on the following characteristics of household heads:

(a) RACE. Each individual is attributed to one of five racial groups, that is, (i) White; (ii) Black; (iii) American Indian, Eskimo or Aleutian; (iv) Asian or Pacific Islander; and (v) Other.[1]

(b) INCOME. Total household income.

(c) EDUCATION. The years of education of the individual.

(d) EMPLOYMENT. Each individual is attributed to one of four categories, namely, (i) Civilian employed or armed forces, at work; (ii) Civilian employed or armed forces, with a job but not at work; (iii) Unemployed; and (iv) Not in labor force.

Drawing on the above information, we construct *GELF* in several ways. The first, and most general, is an implementation of formula (1) that takes into account all four dimensions at the same time without imposing an exogenous partition into groups. In particular, starting from the variables (a)–(d), we rely on principal component analysis[2] to extract for each individual i a synthetic measure x_i that we employ to compute pair-wise distances among all individuals living in the same state, i.e., $|x_i - x_j|$. To generate similarity values s_{ij} that are bounded between 0 and 1, we normalize this distance by the difference between the maximum and the minimum value of the x_i's in the entire US

[1] The last category includes any other race except the four mentioned. The 1990 Census does not identify Hispanic as a separate racial category. However, Alesina, Baqir and Easterly (1999), who construct *ELF* from the same five categories, have verified that the category Hispanic (obtained from a different source) has a correlation of more than 0.9 with the category Other in the Census data.

[2] We have experimented with the standard principal component method as well as with an application that employs a polychoric correlation matrix to take into account the fact that some of our variables are categorical. The estimates reported below rely on the latter method; results obtained using the standard method are available from the authors.

sample, and we subtract the resulting value from 1. Once we have the full set of similarity values $\{s_{ij}\}_{i,j\in\{1,\ldots,n\}}$, computation of (1) is straightforward.

Our second set of results is obtained by assuming that individuals can be aggregated into exogenously defined groups—specifically, the five racial groups described under (a)—and measuring the similarity among these groups along the remaining dimensions. The choice of race as the exogenously given category is purely instrumental to compare our results to the widely used ELF index that relies exclusively on racial shares to assess the extent of diversity. Obviously, depending on the specific application, the grouping could be done on the cleavage that is most relevant for the phenomenon under study. The idea underlying this second set of results is to propose a way to compute $GELF$ that is less data intensive and see whether the qualitative pattern of results differs from that obtained using the full similarity matrix. This second set of results, in turn, is obtained under two alternative methods. The first requires the availability of the entire distribution of individual characteristics, and can be used when individual survey data is available. The second relies only on aggregate data on *mean* characteristics by group. In what follows we briefly describe the two methods.

5.1.1 $GELF$ and similarity of distributions

Once the population is exogenously partitioned into racial groups, we can assess the 'distance' among these groups by comparing the distributions of individual characteristics such as income, education, employment. Consider for example income. We first estimate non-parametrically the distributions of household income by race of the head of the household, $\widehat{f^k}(y)$, for group k. The estimation method applied in the paper is derived from a generalization of the kernel density estimator to take into account the sample weights attached to each observation in each group, namely, from the *adaptive* or *variable* kernel. After estimating the densities of household income by race, we measure the overlap among them, implying that two racial groups whose income distributions perfectly overlap are considered perfectly similar. The measure of overlap of distributions applied is the Kolmogorov measure of variation distance:

$$Kov_{k\ell} = \frac{1}{2}\int \left|\widehat{f^k}(y) - \widehat{f^\ell}(y)\right| dy.$$

$Kov_{k\ell}$ is a measure of the lack of overlap between groups k and l. It ranges between 0 and 1, taking value zero if $\widehat{f^k}(y) = \widehat{f^\ell}(y)$ for all $y \in \mathbb{R}$ and one if $\widehat{f^k}(y)$ and $\widehat{f^\ell}(y)$ do not overlap at all.[3] The resulting measure of similarity between any two groups k and ℓ, that we employ to implement formula (2) for grouped $GELF$, is

$$\overline{s}_{k\ell} = 1 - Kov_{k\ell}.$$

This method is also applied on the distribution of the synthetic measure x_i obtained for each individual in each group k by principal component analysis. In this case we estimate $\widehat{f^k}(x)$, the distribution of the synthetic measure by race, compute the Kolmogorov measure of variation distance and the measure of similarity as described above.

5.1.2 $GELF$ and similarity of means

As an alternative to the distance among distributions, we compute a crude measure of similarity based on the *expected value* of the distribution of the characteristic analyzed. This is to illustrate the performance of $GELF$ in case of grouped data or poor availability of information in the data set.

We can measure similarity with respect to continuous or to categorical variables. For continuous variables, such as household income or education, we indicate by λ^k the sample mean of the distribution for group k, by λ_{Max} the maximum mean value among all groups in all states, and by λ_{Min} the minimum. Then we can compute $\overline{s}_{k\ell}$ for each state as

$$\overline{s}_{k\ell} = 1 - \left| \frac{\lambda^k - \lambda^\ell}{\lambda_{Max} - \lambda_{Min}} \right|. \tag{11}$$

Note that expression (11) is bounded between zero and one by construction.

For categorical variables like employment, we create a dummy variable that assumes the value one if the household head is employed, and zero if he is unemployed or not in the labor force.[4] Indicating by δ^k the sample means of this variable for group k (i.e., the share of the population assuming value one), similarity between any two groups k and ℓ is

$$\overline{s}_{k\ell} = 1 - \left| \delta^k - \delta^\ell \right|.$$

Again, sample weights are used in the computations for these variables.

[3] The distance is sensitive to changes in the distributions only when both take positive values, being insensitive to changes whenever one of them is zero. It will not change if the distributions move apart, provided that there is no overlap between them or that the overlapping part remains unchanged.

[4] We have also experimented with a different definition where one corresponds to households whose head is employed or not in the labor force, and zero to unemployed. The results were not significantly affected and are available from the authors.

5.2 Results

We discuss our results starting with computations based on the $GELF$ formula (1), which relies on the original similarity matrix without pre-assigning individuals to groups. We refer to this index as '$GELF$' with no further specifications. We then turn to approaches that pre-assign individuals to racial groups. In this case the distance among groups is computed on the basis of characteristics other than race (e.g., income) and we refer to the indices as '$GroupedGELF\_income$', etc.

[Insert Figure 1]

The main result of our analysis is summarized in Figure 1. On the horizontal axis we plot values of ELF for all states in the US in 1990. The vertical axis reports the corresponding value of $GELF$. While the two are positively correlated, their relationship is far from linear: the correlation coefficient is only .59. In particular, states like Hawaii, California and Nevada are much more heterogenaous if one only looks at racial shares than if all dimensions are considered jointly. This is because in these states the distribution of income, education and employment is relatively more similar among races than in other states. At the opposite end we have states like Alaska, Kentucky, Rhode Island, Massachussets and in general New England, where diversity measured in terms of racial shares is relatively low, but different races differ in the distrubution of the remaining characteristics to such an extent that they are actually more diverse when the full similarity $GELF$ is employed.

[Insert Table 1]

Table 1 provides the counterpart to the graphical analysis, as it reports the full set of states listed in decreasing order of ethno-linguistic fractionalization, the corresponding values of ELF, $GELF$ and the difference in ranks between ELF and $GELF$ for each state. We prefer to rely on a comparison of ranks because the absolute values of the two indices are not comparable. In particular, in the last column of table 1 we report the difference $ELFrank - GELFrank$, so that negative values indicate that a given state is less fractionalized according to $GELF$ than according to ELF, while positive values indicate the opposite. The magnitude of the difference gives a rough approximation of how big a difference it makes for a particular state to use one index over the other, in terms of relative rankings.

We next turn to an examination of what happens when race is isolated to define relevant subgroups and distance is computed on the remaining components. In particular, we implement formula (2) with the slight modification that individuals are *exogenously* (not endogenously) grouped into five categories—in this case racial groups—and distances among groups are measured as the difference in a synthetic measure of income, education and employment.[5] The results are displayed in Table 2.

[Insert Table 2]

States in Table 2 are listed in decreasing order of $GELF$, and two additional indices (with the corresponding ranks) are reported. The first index, which we denote as $GroupedGELF\_d$ employs the Kolmogorov distance among distributions of the synthetic index to compute similarity values that are the used in formula (2). The second index, denoted simply as *GroupedGELF*, is simpler in that only the *average* value of the synthetic index for each racial group is used when computing distances (differences). While the use of means or of the entire distribution yield very similar results, the comparison with $GELF$ suggests that for some states the exogenous definition of racial categories does make a difference: these are the same states for which the difference between ELF and $GELF$ in Figure 1 was more pronounced. In this sense, and not surprisingly, the $GroupedGELF$ index calculated according to (2) is more similar to ELF than the $GELF$ index (1) calculated on the full similarity matrix.

[Insert Table 3 and Figure 2]

Finally, in Table 3 we try to disentagle the contribution of each individual dimension to overall diversity by implementing a version of (2) where distance among racial groups is measured solely in terms of differences in average income (*GroupedGELF_ income*), differences in average years of education (*GroupedGELF_ edu*), or difference in the share of people employed (*GroupedGELF_ empl*). For each index, we report the value and the rank, and states are still listed in decreasing order of the full similarity $GELF$. The results are quite informative and are more easily visualized through Figure 2. Panel A of the figure plots the original values of ELF on the horizontal axis against *GroupedGELF_ income* on the vertical one. The two measures are closely correlated with two extreme outliers: Hawaii is much less fractionalized when we use *GroupedGELF_ income* than when we use ELF, while the opposite occurs for the District of Columbia. The intuition is similar to that provided when commenting on Figure 1, i.e., in states like Hawaii or California

[5] As before, this synthetic index is the first principal component extracted from our income, education and employment variables, where we use a polychoric correlation matrix to take into account the fact that employment is a categorical variable.

average income levels are relatively more similar among races than they are in DC or in Connecticut, for example. A similar picture is offered in Panel B with respect to years of education. Interestingly, however, when we look at employment levels (Panel C), the relationship between the two indices becomes hump-shaped. The maximum value of diversity according to *GroupedGELF_ empl* corresponds to *intermediate* levels of ethnic fractionalization; on the other hand, very low or very high levels of ELF translate into middle range values of diversity when both race and similarity in employment status are taken into account. A possible interpretation of this result is that sizeable differences in employment status (e.g., high unemployment levels for minorities) may be politically difficult to sustain in states where a relatively high fraction of the population is non-white. On the other hand, the same does not hold for income, as if income differences were more easily acceptable compared to the universal right of access to employment.

While only illustrative, the above analysis highlights some of the potential benefits that may derive from the use of fractionalization indices that do not simply rely on population shares, but also try to incorporate information on other dimensions along which individuals may differ.

6 Concluding remarks

The main purpose of this paper is to provide a theoretical foundation and an empirical application of a new measure of ethnic or cultural diversity. Unlike the most commonly used ELF index, our generalized version $GELF$ makes use of a broader informational base. Instead of limiting the relevant variables to the population shares of predefined groups, we start out with a notion of similarity among individuals and calculate our index value accordingly. It is possible to derive a partition into group endogenously, and the standard ELF index emerges as a special case when no partial similarity is allowed.

The index characterized in the paper is based on information on similarities among individuals. The concept of similarity itself has not been the subject of our investigation; we assumed throughout that it is known how to measure the degree to which any two individuals are similar. In the application to the US we choose as dimensions of similarities across groups ethnicity, household income, education and employment status of the head of the household since we believe that these are important aspects of the US economy that could influence the behavior of individuals. This need not necessarily be the case for other countries. For example, in less developed countries, it might be more important to consider the amount of natural resources, the quality of the land or a combination of characteristics. Allowing any possible concept of similarity has the advantage of leaving the researcher free to pick the most appropriate in the context analyzed. In addition, and most importantly, our index allows to incorporate a multidimensional concept of similarity, as opposed to the single dimension.

The application of our new index is not limited to studies involving ethno-linguistic fractionalization. The generalized index that we propose could be applied to various areas in Economics. It is an index of diversity, and the difference between one and the index value can be interpreted as an index of concentration. One of the most widely used concentration indices is the Herfindahl index, which is obtained by subtracting ELF from one. The Herfindahl index has widespread applications in various areas including academic research as well as antitrust regulation. For example, since 1992 the US Department of Justice has used the Herfindahl index as a measure of market concentration to enforce antitrust regulation. According to the DOJ Horizontal Merger Guidelines of 1992, markets with an index of 0.18 or more should be considered 'concentrated'. A natural alternative concentration index based on similarity information can be defined by subtracting $GELF$ from one.

Appendix

In this appendix, we illustrate that our characterization result is unchanged if the set of similarity matrices $\mathcal{S}^n$ consists of all $n \times n$ matrices S satisfying conditions (a) and (b) of Section 2, but not necessarily (c). This is achieved by some straightforard modifications of the definitions used in the proof of Theorem 1.

That any positive multiple of G satisfies the axioms on the larger domain as well is, again, straightforward to verify. Conversely, suppose D is a diversity measure defined on the larger domain satisfying normalization, anonymity, additivity and replication invariance. Let $n \in \mathbb{N} \setminus \{1\}$, and define the set $\mathcal{X}^n \subseteq \mathbb{R}^{n(n-1)/2}$ by

$$\begin{aligned}\mathcal{X}^n \;=\; \{x = (x_{ij})_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,n\}\setminus\{i\}}} \mid \exists S \in \mathcal{S}^n \text{ such that } s_{ij} = x_{ij} \text{ for all } i \in \{1,\ldots,n\} \\ \text{and for all } j \in \{1,\ldots,n\} \setminus \{i\}\}.\end{aligned}$$

Define the function $F^n \colon \mathcal{X}^n \to \mathbb{R}$ by letting, for all $x \in \mathcal{X}^n$,

$$F^n(x) = D(S) - D(I^n) \tag{12}$$

where $S \in \mathcal{S}^n$ is such that $s_{ij} = x_{ij}$ for all $i \in \{1,\ldots,n\}$ and for all $j \in \{1,\ldots,n\} \setminus \{i\}$. Because D is bounded below by zero, it follows that F^n is bounded below by $-D(I^n)$.

Furthermore, the additivity of D implies that F^n satisfies Cauchy's basic functional equation

$$F^n(x+y) = F^n(x) + F^n(y) \tag{13}$$

for all $x, y \in \mathcal{X}^n$ such that $(x+y) \in \mathcal{X}^n$; see Aczél (1966, Section 2.1).

Fix $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n\} \setminus \{i\}$, and define the function $f_{ij}^n \colon [0,1] \to \mathbb{R}$ by

$$f_{ij}^n(x_{ij}) = F^n(x_{ij}; \mathbf{0}^{n(n-1)-1})$$

for all $x_{ij} \in [0,1]$, where the vector $(x_{ij}; \mathbf{0}^{n(n-1)-1})$ is such that the component corresponding to ij is given by x_{ij} and all other entries (if any) are equal to zero. The function f_{ij}^n is bounded below because F^n is and, as an immediate consequence of (13), it satisfies the Cauchy equation

$$f_{ij}^n(x_{ij} + y_{ij}) = f_{ij}^n(x_{ij}) + f_{ij}^n(y_{ij}) \tag{14}$$

for all $x_{ij}, y_{ij} \in [0,1]$ such that $(x_{ij} + y_{ij}) \in [0,1]$. Because the domain of f_{ij}^n is an interval containing the origin and f_{ij}^n is bounded below, the only solutions to (14) are linear functions; see Aczél (1966, Section 2.1). Thus, there exists $c_{ij}^n \in \mathbb{R}$ such that

$$F^n(x_{ij}; \mathbf{0}^{n(n-1)-1}) = f_{ij}^n(x_{ij}) = c_{ij}^n x_{ij} \tag{15}$$

for all $x_{ij} \in [0,1]$.

Let $S \in \mathcal{S}^n$. By additivity, the definition of F^n and (15),

$$F^n\left((s_{ij})_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,n\}\setminus\{i\}}}\right) = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} F^n(s_{ij}; \mathbf{0}^{n(n-1)-1}) = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} f_{ij}^n(s_{ij}) = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} c_{ij}^n s_{ij}$$

and, defining $d^n = D(I^n)$ and substituting into (12), we obtain

$$D(S) = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} c_{ij}^n s_{ij} + d^n. \tag{16}$$

Now fix $i, k \in \{1, \ldots, n\}$, $j \in \{1, \ldots, n\} \setminus \{i\}$ and $\ell \in \{1, \ldots, n\} \setminus \{k\}$, and let $S \in \mathcal{S}^n$ be such that $s_{ij} = 1$ and all other off-diagonal entries of S are equal to zero. Let the bijection $\pi \in \Pi^n$ be such that $\pi(i) = k$, $\pi(j) = \ell$, $\pi(k) = i$, $\pi(\ell) = j$ and $\pi(h) = h$ for all $h \in \{1, \ldots, n\} \setminus \{i, j, k, \ell\}$. By (16), we obtain

$$D(S) = c_{ij}^n + d^n \quad \text{and} \quad D(S_\pi) = c_{k\ell}^n + d^n,$$

and anonymity implies $c^n_{ij} = c^n_{k\ell}$. Therefore, there exists $c^n \in \mathbb{R}$ such that $c^n_{ij} = c^n$ for all $i \in \{1, \ldots, n\}$ and for all $j \in \{1, \ldots, n\} \setminus \{i\}$, and substituting into (16) yields

$$D(S) = c^n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} + d^n$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Normalization requires

$$D(\mathbf{1}^n) = c^n n(n-1) + d^n = 0$$

and, therefore, $d^n = -c^n n(n-1)$ for all $n \in \mathbb{N} \setminus \{1\}$. Using normalization again, we obtain

$$D(I^n) = -c^n n(n-1) > 0$$

which implies $c^n < 0$ for all $n \in \mathbb{N} \setminus \{1\}$. Thus,

$$D(S) = c^n \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij} - c^n n(n-1) \tag{17}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Let n be an even integer greater than or equal to four. By replication invariance and (17),

$$D(R^2_{n/2}) = c^n n \left(\frac{n}{2} - 1 \right) - c^n n(n-1) = -c^2 = D(I^2).$$

Solving, we obtain

$$c^n = 2 \frac{c^2}{n^2}. \tag{18}$$

Now let n be an odd integer greater than or equal to three. Thus, $q = 2n$ is even, and the above argument implies

$$c^q = 2 \frac{c^2}{q^2} = \frac{c^2}{2n^2}. \tag{19}$$

Furthermore, replication invariance requires

$$D(R^n_2) = D(R^{q/2}_2) = c^q q - c^q q(q-1) = -c^n n(n-1) = D(I^n).$$

Solving for c^n and using the equality $q = 2n$, it follows that $c^n = 4c^q$ and, combined with (19), we obtain (18) for all odd $n \in \mathbb{N} \setminus \{1\}$ as well.

Substituting into (17), simplifying and defining $\beta = -2c^2 > 0$, it follows that, for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$,

$$\begin{aligned} D(S) &= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} s_{ij} - c^n n(n-1) \\ &= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} - 2\frac{c^2}{n^2}n - 2\frac{c^2}{n^2}n(n-1) \\ &= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\right] \\ &= \beta G(S). \ \blacksquare \end{aligned}$$

References

[1] Aczél, János (1966), *Lectures on Functional Equations and Their Applications*, Academic Press, New York.

[2] Alesina, Alberto, Reza Baqir and William Easterly (1999), "Public Goods and Ethnic Divisions", *Quarterly Journal of Economics*, 114, 1243–1284.

[3] Alesina, Alberto, Reza Baqir and Caroline Hoxby (2004), "Political Jurisdictions in Heterogeneous Communities", *Journal of Political Economy*, 112, 348–396.

[4] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), "Fractionalization", *Journal of Economic Growth*, 8, 155–194.

[5] Alesina, Alberto and Eliana La Ferrara (2000), "Participation in Heterogeneous Communities", *Quarterly Journal of Economics*, 115, 847–904.

[6] Alesina, Alberto and Eliana La Ferrara (2005), "Ethnic Diversity and Economic Performance", *Journal of Economic Literature*, 43, 762–800.

[7] Bossert, Walter, Prasanta K. Pattanaik and Yongsheng Xu (2003), "Similarity of Options and the Measurement of Diversity", *Journal of Theoretical Politics*, 15, 405–421.

[8] Caselli, Francesco and Wilbur J. Coleman (2002), "On the Theory of Ethnic Conflict", unpublished manuscript, Harvard University.

[9] Desmet, Klaus, Ignacio Ortuño-Ortín and Shlomo Weber (2005), "Peripheral Diversity and Redistribution", CEPR, Discussion Paper No.5112.

[10] Easterly, William and Ross Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, 111, 1203–1250.

[11] Esteban, Joan-Maria and Debraj Ray (1994), "On the Measurement of Polarization", *Econometrica*, 62, 819–851.

[12] Fearon, James D. (2003), "Ethnic and Cultural Diversity by Country", *Journal of Economic Growth*, 8, 195–222.

[13] Gini, Corrado (1912), "Variabilità e Mutabilità" Studi Economico-Giuridici della Facoltà di Giurisprudenza dell'Università di Cagliari, a.III, parte II.

[14] Greenberg, Joseph H. (1956) "The Measurement of Linguistic Diversity", *Language*, 32, 109–115.

[15] Laitin, David (2000), "What is a Language Community?", *American Journal of Political Science*, 44, 142–154.

[16] Mauro, Paolo (1995), "Corruption and Growth", *Quarterly Journal of Economics*, 110, 681–712.

[17] Montalvo, Jose G. and Marta Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict, and Civil Wars", *American Economic Review*, 95, 796–816.

[18] Nehring, Klaus and Clemens Puppe (2002), "A Theory of Diversity", *Econometrica*, 70, 1155–1198.

[19] Pattanaik, Prasanta K. and Yongsheng Xu (2000), "On Diversity and Freedom of Choice", *Mathematical Social Sciences*, 40, 123–130.

[20] Rao, Radhakrishna C. (1982), "Diversity: Its Measurement, Decomposition, Apportionment and Analysis", *Sankhyā*, 44, A, 1–22.

[21] Rao, Radhakrishna C. (1984), "Convexity Properties of Entropy Functions and Analysis of Diversity", in *Inequalities in Statistics and Probability*, Y.L. Tong Ed., IMS Lecture Notes, 5, 68–77.

[22] Rao, Radhakrishna C. and Tapan K. Nayak (1985), "Cross Entropy, Dissimilarity Measures, and Characterizations of Quadratic Entropy", *IEEE Transactions on Information Theory*, IT-31, 5, 589–593.

[23] Simpson, Edward H. (1949), "Measurement of Diversity", *Nature*, 163, 688.

[24] Vigdor, Jacob L. (2002), "Interpreting Ethnic Fragmentation Effects", *Economics Letters*, 75, 271–76.

[25] Weitzman, Martin (1992), "On Diversity", *Quarterly Journal of Economics*, 107, 363–405.

[26] Weitzman, Martin (1998), "The Noah's Ark Problem", *Econometrica*, 66, 1279–1298.

Figure 1: *GELF* and *ELF* in the US States.

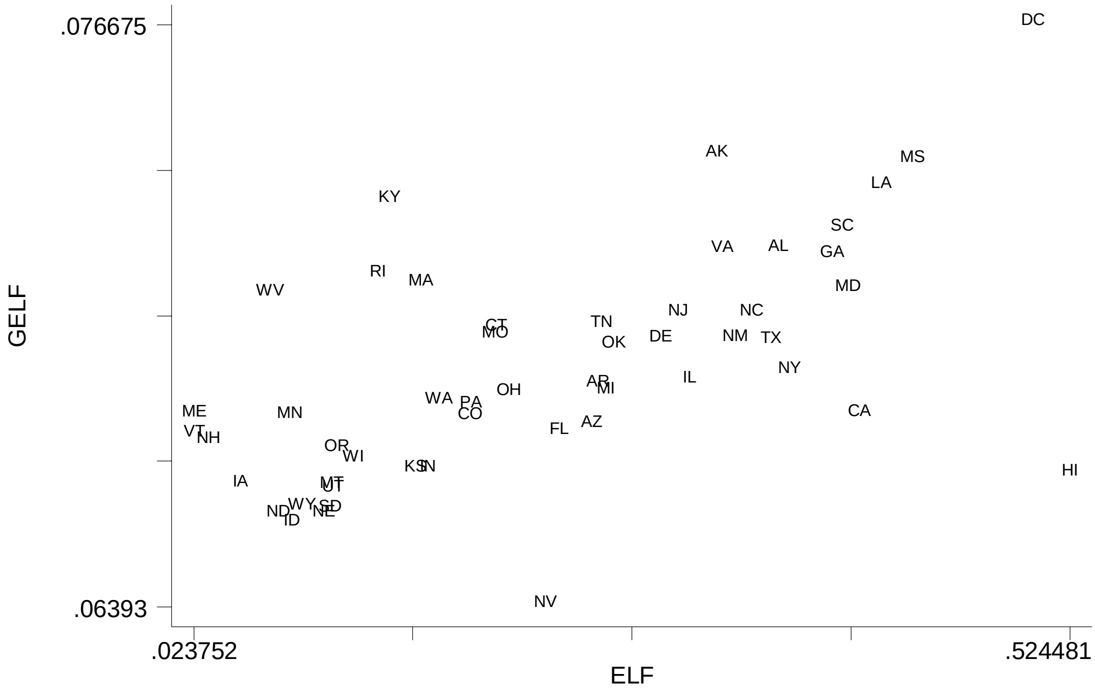
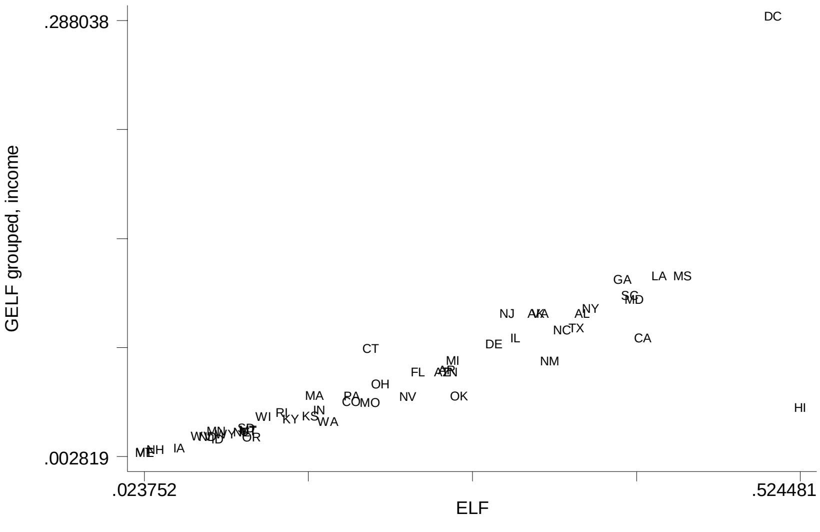
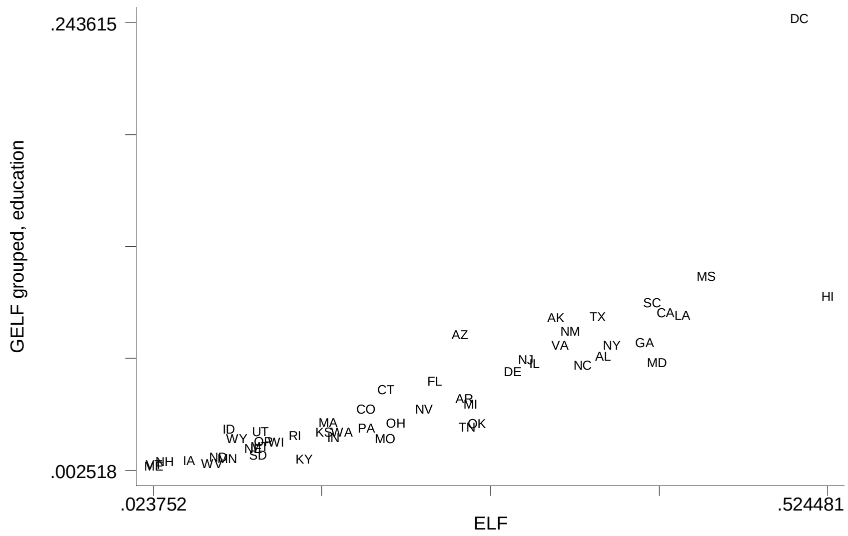


Figure 2: *GroupedGELF* (income, education, employment) and *ELF* in the US States.

Panel A



Panel B



Panel C

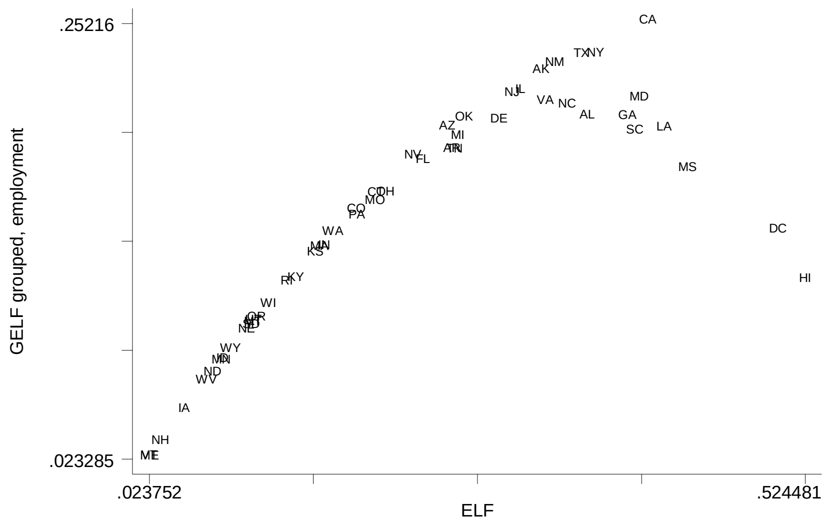


Table 1: *GELF* and *ELF* in the US States.

State	ELF	ELF rank	GELF	GELF rank	Difference (ELF rank-GELF rank)
HI	0.5245	1	0.0668	42	-41
DC	0.5032	2	0.0767	1	1
MS	0.4344	3	0.0737	3	0
LA	0.4165	4	0.0731	4	0
CA	0.4042	5	0.0681	30	-25
MD	0.3975	6	0.0709	12	-6
SC	0.3940	7	0.0722	6	1
GA	0.3885	8	0.0716	9	-1
NY	0.3644	9	0.0690	23	-14
AL	0.3577	10	0.0717	7	3
TX	0.3534	11	0.0697	21	-10
NC	0.3425	12	0.0703	15	-3
NM	0.3332	13	0.0698	19	-6
VA	0.3259	14	0.0717	8	6
AK	0.3225	15	0.0738	2	13
IL	0.3069	16	0.0688	24	-8
NJ	0.3005	17	0.0703	14	3
DE	0.2904	18	0.0697	20	-2
OK	0.2640	19	0.0696	22	-3
MI	0.2591	20	0.0686	26	-6
TN	0.2566	21	0.0701	16	5
AR	0.2546	22	0.0688	25	-3
AZ	0.2509	23	0.0679	34	-11
FL	0.2324	24	0.0677	35	-11
NV	0.2248	25	0.0639	51	-26
OH	0.2037	26	0.0686	27	-1
CT	0.1967	27	0.0700	17	10
MO	0.1958	28	0.0698	18	10
PA	0.1821	29	0.0683	29	0
CO	0.1815	30	0.0680	33	-3
WA	0.1637	31	0.0684	28	3
IN	0.1574	32	0.0669	41	-9
MA	0.1535	33	0.0710	11	22
KS	0.1501	34	0.0669	40	-6
KY	0.1354	35	0.0728	5	30
RI	0.1290	36	0.0712	10	26
WI	0.1145	37	0.0671	39	-2
OR	0.1054	38	0.0673	38	0
UT	0.1033	39	0.0664	45	-6
MT	0.1027	40	0.0665	44	-4
SD	0.1015	41	0.0660	47	-6
NE	0.0980	42	0.0659	49	-7
WY	0.0856	43	0.0661	46	-3
ID	0.0797	44	0.0657	50	-6
MN	0.0788	45	0.0681	32	13
ND	0.0718	46	0.0659	48	-2
WV	0.0674	47	0.0708	13	34
IA	0.0503	48	0.0666	43	5
NH	0.0321	49	0.0675	37	12
VT	0.0240	50	0.0677	36	14
ME	0.0238	51	0.0681	31	20

Table 2: *GELF* and *GroupedGelf* (Kolmogorov and Average) in the US States.

State	*GELF*	*GELF rank*	*GroupedGELF_d*	*GroupedGELF_d rank*	*GroupedGELF*	*GroupedGELF rank*
HI	0.0668	42	0.0917	6	0.0588	13
DC	0.0767	1	0.2306	1	0.1864	1
MS	0.0737	3	0.1161	2	0.1061	2
LA	0.0731	4	0.1070	3	0.0951	3
CA	0.0681	30	0.0793	9	0.0586	14
MD	0.0709	12	0.0675	14	0.0564	16
SC	0.0722	6	0.0974	4	0.0879	4
GA	0.0716	9	0.0850	7	0.0758	6
NY	0.0690	23	0.0701	12	0.0617	10
AL	0.0717	7	0.0810	8	0.0727	7
TX	0.0697	21	0.0783	10	0.0646	8
NC	0.0703	15	0.0701	13	0.0613	12
NM	0.0698	19	0.0656	17	0.0614	11
VA	0.0717	8	0.0752	11	0.0637	9
AK	0.0738	2	0.0966	5	0.0875	5
IL	0.0688	24	0.0664	15	0.0576	15
NJ	0.0703	14	0.0661	16	0.0541	18
DE	0.0697	20	0.0545	19	0.0510	20
OK	0.0696	22	0.0365	26	0.0303	26
MI	0.0686	26	0.0558	18	0.0516	19
TN	0.0701	16	0.0432	24	0.0356	25
AR	0.0688	25	0.0543	20	0.0475	21
AZ	0.0679	34	0.0440	23	0.0558	17
FL	0.0677	35	0.0448	22	0.0388	23
NV	0.0639	51	0.0363	27	0.0285	29
OH	0.0686	27	0.0392	25	0.0364	24
CT	0.0700	17	0.0481	21	0.0407	22
MO	0.0698	18	0.0291	31	0.0252	31
PA	0.0683	29	0.0324	29	0.0294	27
CO	0.0680	33	0.0348	28	0.0293	28
WA	0.0684	28	0.0257	35	0.0184	37
IN	0.0669	41	0.0284	32	0.0241	32
MA	0.0710	11	0.0310	30	0.0256	30
KS	0.0669	40	0.0261	34	0.0216	33
KY	0.0728	5	0.0202	41	0.0146	42
RI	0.0712	10	0.0247	36	0.0197	36
WI	0.0671	39	0.0272	33	0.0213	34
OR	0.0673	38	0.0168	44	0.0125	45
UT	0.0664	45	0.0222	39	0.0180	39
MT	0.0665	44	0.0222	38	0.0184	38
SD	0.0660	47	0.0243	37	0.0201	35
NE	0.0659	49	0.0182	43	0.0148	41
WY	0.0661	46	0.0189	42	0.0157	40
ID	0.0657	50	0.0214	40	0.0145	43
MN	0.0681	32	0.0157	46	0.0114	46
ND	0.0659	48	0.0164	45	0.0128	44
WV	0.0708	13	0.0123	47	0.0097	47
IA	0.0666	43	0.0085	48	0.0058	48
NH	0.0675	37	0.0052	49	0.0040	49
VT	0.0677	36	0.0044	50	0.0020	51
ME	0.0681	31	0.0043	51	0.0030	50

Table 3: *GELF* and *GroupedGelf* (income, education, employment) in the US States.

State	GELF rank	GroupedGELF_income	rank	GroupedGELF_edu	rank	GroupedGELF_empl	rank
DC	1	0.2880	1	0.2436	1	0.1426	29
AK	2	0.0936	10	0.0825	8	0.2261	5
MS	3	0.1181	2	0.1048	2	0.1748	23
LA	4	0.1181	3	0.0836	6	0.1960	16
KY	5	0.0249	36	0.0063	46	0.1169	34
SC	6	0.1054	5	0.0905	4	0.1943	17
AL	7	0.0937	9	0.0618	14	0.2023	11
VA	8	0.0938	8	0.0675	12	0.2100	9
GA	9	0.1161	4	0.0690	11	0.2021	12
RI	10	0.0285	33	0.0190	35	0.1151	36
MA	11	0.0401	25	0.0259	26	0.1332	32
MD	12	0.1028	6	0.0581	16	0.2117	8
WV	13	0.0135	44	0.0039	49	0.0631	47
NJ	14	0.0936	11	0.0596	15	0.2140	7
NC	15	0.0827	13	0.0568	18	0.2081	10
TN	16	0.0555	22	0.0236	29	0.1845	20
CT	17	0.0709	17	0.0435	21	0.1615	25
MO	18	0.0353	30	0.0174	37	0.1575	26
NM	19	0.0625	19	0.0751	9	0.2300	4
DE	20	0.0735	16	0.0533	19	0.2004	14
TX	21	0.0840	12	0.0830	7	0.2346	3
OK	22	0.0396	26	0.0254	28	0.2013	13
NY	23	0.0967	7	0.0674	13	0.2349	2
IL	24	0.0778	15	0.0576	17	0.2156	6
AR	25	0.0564	20	0.0389	22	0.1849	19
MI	26	0.0629	18	0.0356	23	0.1917	18
OH	27	0.0475	24	0.0257	27	0.1617	24
WA	28	0.0230	37	0.0209	34	0.1411	30
PA	29	0.0396	27	0.0229	30	0.1499	28
CA	30	0.0779	14	0.0851	5	0.2522	1
ME	31	0.0028	51	0.0025	51	0.0233	51
MN	32	0.0169	40	0.0067	45	0.0737	45
CO	33	0.0361	29	0.0330	25	0.1531	27
AZ	34	0.0556	21	0.0733	10	0.1965	15
FL	35	0.0552	23	0.0480	20	0.1790	22
VT	36	0.0031	50	0.0034	50	0.0235	50
NH	37	0.0047	49	0.0051	48	0.0313	49
OR	38	0.0129	46	0.0156	39	0.0963	38
WI	39	0.0264	35	0.0153	40	0.1033	37
KS	40	0.0265	34	0.0209	33	0.1304	33
IN	41	0.0304	32	0.0178	36	0.1337	31
HI	42	0.0325	31	0.0940	3	0.1166	35
IA	43	0.0054	48	0.0056	47	0.0482	48
MT	44	0.0175	39	0.0131	41	0.0935	40
UT	45	0.0168	41	0.0211	32	0.0945	39
WY	46	0.0150	43	0.0171	38	0.0796	43
SD	47	0.0188	38	0.0084	43	0.0922	41
ND	48	0.0131	45	0.0073	44	0.0674	46
NE	49	0.0162	42	0.0119	42	0.0899	42
ID	50	0.0117	47	0.0224	31	0.0745	44
NV	51	0.0394	28	0.0330	24	0.1811	21