Measurement of social diversity

Shlomo Weber

Center for Study of Diversity and Social Interactions, New Economic School

HCEO-NES Summer School on Socioeconomic Inequality

August 30, 2017, Moscow

The presentation is based, in part, on V. Ginsburgh and S. Weber "Economics of Language". The plan:

Diversity
- Measurement

Group Identification
- Group boundaries
- Group association

Indices

Fractionalization indices
- Dichotomous indices
- Nondichotomous indices
- Distances

Polarization indices
- Alienation and identification

Disenfranchisement indices
- Native languages
- All spoken languages

Applications

Social interactions

At the entrance to Higashi Honganji Temple in Kyoto, the ancient capital of Japan, a visitor is greeted by two sentences written on the wall:

"Living together in diversity. Learning to accept our differences."

"Before we can change direction, we have to question many of the assumptions underlying our current philosophy.

Assumptions like bigger is better; you can't stop progress; no speed is too fast; globalization is good.

Then we have replaced it with some different assumptions: small is beautiful; roots and traditions are worth preserving; variety is the spice of life; the only work worth doing is a meaningful work; biodiversity is the necessary pre-condition for human survival." Robert Bateman, Canadian artist.

There is a wide range of diversity facets:

religious

historical

economic

ideological

geographical

linguistic

genetic

and many others.

Is diversity good or bad?

"An angel is more valuable than a stone. It does not follow, however, that two angels are more valuable than the one angel and one stone."
Thomas Aquinas, *Summa Contra Gentiles, III.*

Saxenian (1996, 1999) argued that the remarkable success of Silicon Valley in 80s and 90s was due to a diverse cultural and professional background of scientists, entrepreneurs and managers who came from Europe, Japan, China, India, Israel, and other places. At that time 40% of businesses there had a foreign-born co-owner.

Florida (2002), Florida and Gates (2001) examined the importance of diversity to high-tech growth. They ranked 50 US cities in terms of diversity (number of artists, foreign-born, homosexuals) and showed the success of more diverse ones (from San Francisco to Buffalo).

In the same vain, Ottaviano and Peri (2003) claimed that a higher level of linguistic diversity in the US cities leads to higher earnings and standards of living.

On the other hand,

Ethnic, religious and linguistic conflicts conflicts everywhere (Asia, Africa, Latin America, Europe).

Tragedy of Africa (Easterly and Levine (1997).

Example: Production of cocoa in Ghana (Kwame Nkrumah).

Papua New Guinea - 857 active languages. And this is not that exceptional. In most of the African countries have more than 100 active languages. Also India.

European Union. Difficulties of finding a common way. (24 official languages).

What is the optimal degree of diversity?

Maybe, some diversity is good, but too much is “much too much”?

How many spoons do we put in our coffee or tea?

How salty is our meal?

How much colour would we want to see in painting we like?

Is it possible that the current structure of the European Union yileds an excessive degree of diversity?

Why the study of diversity is important?

Various policies, natural (and unnatural) disasters impact different population groups and regions in a non-uniform way.

Even successful economic policies make some population segments and regions worse off, and a correcting mechanism is needed to deal with some effects of globalization.

The situation where some parts of the population do not participate in the process of creation and advancement, it simply does not make an economic sense. In the environment of increasing globalization and firece competition, a country needs every population group to be competitive and successful. We need a larger cake to divide it between citizens.

When dealing with diversity, it is important to point out that it is a very dynamic concept.

What really matters is not the absolute levels of diversity but their evolution over time, and even more important, a *subjective view of diversity.*

• Measurement of diversity

In principle, there are two main elements in examining societal diversity.

One is the set of attributes (native language, age, education, gender, etc.) that exist in the society or the number of spieces in of some kind in the nature.

The type of measurement based on the number of spieces or attributes, or, more generally, on the ranking of the sets of attributes is often used in order to determine the degree of societal diversity. The approach is common in studies of biodiversity in biology, zoology, and botanics, as well as in the social choice literature. Bossert et al. (2003) point out: Does the newly created, left-wing party increase the diversity of political opinions available to the voters in a country? ... Would the extinction of giant pandas reduce the diversity?"

In social sciences we are mostly interested in the sizes of population groups that correspond to various attributes. That is, a distribution of the entire population across the groups population is inddispensable for our analysis.

Group identification

Two facets:

• Group boundaries – how does one define a partition of the country or countries into separate groups?

• Group association – how do individuals identify themselves with a community to which they belong?

Group boundaries

A major challenge is the prevalence of multiple identities. People may speak several languages using them in communication across different cultural zones.

To construct an ethnolinguistic map, one can use the dominant linguistic identity. The first attempt of creating a comprehensive world *atlas* was undertaken by Soviet ethnographers in the Miklukho-Maklay Research Institute in Moscow. The result, called ELF (Ethno-Linguistic Fractionalization), was published in Atlas Narodov Mira in 1964. This remarkable dataset was picked by Western scholars, starting with Rustow (1967), Taylor and Hudson (1972) and for almost fifty years played the crucial role in analyzing the impact of linguistic diversity on growth, investment in public goods, quality of government services, corruption, etc.

Fearon (2003), Alesina et al. (2003), Alesina and Zhuravskaya (2008), Desmet et al. (2009), are among others who developed more advanced fractionalization datasets.

The identification of distinct languages or dialects may not be straightforward. Are Serbian and Croatian different languages? Should various dialects of Italian, German and Mandarin be treated as separate languages? While economists are not qualified to determine whether Serbian and Croatian are the same language or not, they can mitigate the impact of this determination by using the notion of linguistic proximity of, say, Serbian and Croatian. The notions of linguistic distanc will be discussed later.

Note that if one takes the different dialects of Italian to constitute different groups, then Italy appears to be very diverse. However, if one considers these different dialects to be only minor variations of Italian, then Italy turns to be quite homogeneous.

Desmet et al. (2015) use the linguistic tree (with more than six thousand existing languages), and study different group structure that depend on the level of aggregation on the tree.

Coarse linguistic divisions, obtained at high levels of aggregation, describe cleavages that have emerged thousands years ago, while lower levels of aggreagation generate finer partitions emerged through more recent changes. The research question at hand, whether it is a length or severity of conflicts (coarser partitions) examination of the growth, patterns of redistribution, provision of public goods (finer partitions), should endgenously determine a proper level of aggregation and a corresponding linguistic partition of the society.

Group association

While the issue of objective identification is extensively discussed in the economic literature, the question of self-identification requires more attention (Akerlof and Kranton (2000)).

Esteban and Ray (1994) extensively examine the abstract notions of identification with one's own group and alienation towards other groups. In their framework both notions solely depend on the size of the groups.

Aspachs-Bracons et al. (2008) studied the rise of Catalan identity after the introduction of the compulsory bilingual education in 1983.

Castaneda-Dower et al. (2017) conducted an empirical study of the rise of alienation levels of various groups based on the historical patterns of English acquisition in the pre-colonial period in the protracted civil war in Sri Lanka, where the linguistic divide played a crucial role and was claimed to be responsible for the outbreak of the war.

But still more empirical research is needed to study individual identity choices to associate them with one linguistic group or another. In the U.S. context, for instance, how strong is the association of individuals with African American Vernacular English (AAVE) and New York Latino English (NYLE)?

Obviously, the identification of individuals is driven by the fear of being rejected by their own community if they choose to speak Standard English instead of the vernacular language that the majority of the community speaks (Lewis, 2007). On the other hand, the societal "stigma' (Besley and Coates, 1992) may lessen ties with one's own community. It could be the case that the US population should be split into three groups: those who learn Standard English as their first language, those who learn a nonstandard dialect of English natively, and those who do not learn English as their mother tongue? (Baugh, 1999).

The bulding blocs, described in this section, namely the group identification and linguistic distances, are combined to define fractionalization indices, whose impact on growth. development, investment in public, quality of government policies, corruption and other ecomomic outcomes will be discussed later.

We describe here four groups of indices.

We start with dichotomous feactionalization indices that rely only on the size of distinct groups in the society.

We then proceed with combining group sizes with nondichotomous linguistic distances.

There are two extensions of fractionalozation indices.

One is polarization indices that incorporate the identification of individuals with their own group and alieantion toward the others.

Another extension is the departure from monolingual framework and accounting for language proficiency in nonnative languages, which allows for evaluating government linguistic policies through determination of disenfranchisement levels.

Fractionalization indices.

Dichotomous fractionalization indices

The most often used index defined for a multilingual society divided into distinct goups, each member of which speaks the same native language (we disregard the proficiency in other languages).

Let society with the total population of N individuals consist of K groups, $k = 1, \ldots, K$.

The population of k-th group is given by N_k and $\Sigma_{k=1}^{K} N_k = N$. Let $n_k = \frac{N_k}{N}$ be the fraction of the k-th group population in the entire society.

We define the index, referred to as A-index, as the the probability that two individuals, randomly picked from the entire society, belong to two different group.

Obviously, if a society is monolingual, all the members of the socrety belong to the same group and the value of the A-index is zero.

But if the society consists of a large number of small groups, the probablity that two randomly chosen individuals belong to two different groups is quite high.

Formally, the A-index can be presented as

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

The index was introduced by Gini (1912) as the *mutuality index*. It was later rediscovered by Simpson (1949) and Greenberg (1956), who called it the *monolingual nonweighted index*. It is also a reversed Hirschmann-Herfindahl index often applied for estimating the degree of industrial competiteveness.

Note that in calculating the value of A-index we utilize the dichotomous distances. Individuals either belong to same group or do not. In former case their linguistic distance is zero, and in the latter it is one. In doing so, we ignore the challenge of linguistic proximity and simply set 1 for any non-zero linguistic distance.

Another important nondichotomous index is the Shannon (or Shannon-Wiener) (1948) entropy:

$$E = -\sum_{k=1}^{K} n_k \log n_k.$$

The entropy is actually much more often used in biology, statistics and information science, but not in social sciences where the usage of the A-index is more prevalent. Both indices have similar mathematical properties and they were unified through the common axiomatical formulation in Davydov and Weber (2016) (see also Hill (1973) and Simovici and Jaroszewiz (2002)), who offered a general form

$$A^{\alpha} = 1 - \sum_{k=1}^{K} n_k^{\alpha},$$

where α is a positive parameter different from one. Obviously, the value of A-index coincides with A^{α} for $\alpha = 2$. It is also quite easy to verify that A^{α} approaches the entropy E when α tends to one.

The value of the parameter α in the above formulation can be interpreted as the degree societal sensitive towards diversity. Societies, regions, cities or counties may differ with the respect to its own value of fractionalization. Some could be threatened by diversity while others may welcome it, thus, exhibiting the attitudes that may have profound economic and politucal outcomes. Ottaviano and Peri (2006) show that Los Angeles, New York and San Francisco have a substantially higher degree of linguistic diversity than, say, midwestern cities Cincinnati and Indianoplis. Much less obvious are differences in "perceived diversity" that indicates how people feel about the impact of globalization, channeled through employment prospects and the presence of immigrants in their own communities. In other words, different societies choose a different α, and the identification of that parameter should be an important topic in the research in this field.

Nondichotomous fractionalization indices

The reliance on A-index may produce unexpected results. Desmet et al. (2009) compared the value of that index in two European countries, Andorra and Belgium.

In a small southeuropean principality of Andiorra, with the population of less than 100,000 peole, roughly a half of its residents have Catalan as the native tongue, whereas the native tongue of the other half is Spanish.

In Belgium the split is about 60 and 40 percent between the Dutch-speaking and the French-speaking populations.

A simple algebra shows that the value of the A index is $1-0.5^2-0.5^2 = 0.5$ for Andorra and $1-0.6^2-0.4^2 = 0.48$. In other words, Andorra is more linguistically diverse than Belgium!

The reason for this bizarre conclusion that A-index does not take into account the proximity between languages. Catalan and Spanish are similar Romance languages, whereas Dutch and French, being members of two distinct language families, Germanic and Romance, are quite distant from each other. The incorporation of linguistic proximity would (and does) make Belgium more linguistic diverse than Andorra.

• Linguistic distances
Language trees

Lexicostatical distances:

The distance matrix is based on cognate date collected by Isidore Dyen at Yale University in the 1960s:

200 basic meanings (chosen by Swadesh (1952))

95 Indo-European speech varieties (languages and dialects)

For each meaning - there is a cognate class of different speech varieties that have an unbroken history of descent from common ancestral word.

For every two varieties, we calculate the number of "cognate" and "non-cognate" meanings. If for example, we have 80 cognate and 120 non-cognate for a pair of languages, the Dyen distance is $\frac{120}{200} = 0.6$.

Dyen Matrix of distances between the EU25 + RU, UKR languages

	IT	FR	SP	PT	GE	DU	SW	DA	EN	LI	LA	SV	CZ	SL	PL
IT	0	0,20	0,21	0,23	0,73	0,74	0,74	0,74	0,75	0,76	0,78	0,76	0,75	0,75	0,76
FR	0,20	0	0,27	0,29	0,76	0,76	0,76	0,76	0,76	0,78	0,79	0,78	0,77	0,76	0,78
SP	0,21	0,27	0	0,13	0,75	0,74	0,75	0,75	0,76	0,77	0,79	0,77	0,76	0,76	0,77
PT	0,23	0,29	0,13	0	0,75	0,75	0,74	0,75	0,76	0,78	0,80	0,78	0,76	0,76	0,77
GE	0,73	0,76	0,75	0,75	0	0,16	0,30	0,29	0,42	0,78	0,80	0,73	0,74	0,74	0,75
DU	0,74	0,76	0,74	0,75	0,16	0	0,31	0,34	0,39	0,79	0,80	0,75	0,76	0,75	0,77
SW	0,74	0,76	0,75	0,74	0,30	0,31	0	0,13	0,41	0,78	0,79	0,75	0,75	0,74	0,76
DA	0,74	0,76	0,75	0,75	0,29	0,34	0,13	0	0,41	0,78	0,80	0,73	0,75	0,73	0,75
EN	0,75	0,76	0,76	0,76	0,42	0,39	0,41	0,41	0	0,78	0,80	0,75	0,76	0,75	0,76
LI	0,76	0,78	0,77	0,78	0,78	0,79	0,78	0,78	0,78	0	0,39	0,66	0,62	0,60	0,64
LA	0,78	0,79	0,79	0,80	0,80	0,80	0,79	0,80	0,80	0,39	0	0,68	0,67	0,64	0,67
SV	0,76	0,78	0,77	0,78	0,73	0,75	0,75	0,73	0,75	0,66	0,68	0	0,34	0,31	0,37
CZ	0,75	0,77	0,76	0,76	0,74	0,76	0,75	0,75	0,76	0,62	0,67	0,34	0	0,09	0,23
SL	0,75	0,76	0,75	0,76	0,74	0,75	0,74	0,73	0,75	0,60	0,64	0,31	0,09	0	0,22
PL	0,76	0,78	0,77	0,78	0,75	0,77	0,76	0,75	0,76	0,64	0,67	0,37	0,23	0,22	0
GR	0,82	0,84	0,83	0,83	0,81	0,81	0,81	0,82	0,84	0,83	0,85	0,82	0,84	0,83	0,84
RU	0,76	0,77	0,77	0,77	0,76	0,78	0,75	0,74	0,76	0,62	0,64	0,39	0,26	0,26	0,27
UKR	0,77	0,78	0,78	0,78	0,76	0,79	0,76	0,76	0,78	0,63	0,64	0,36	0,24	0,19	0,20

IT - Italian; FR - French; SP - Spanish; PT - Portugal; GE - German; DU - Dutch; SW - Swedish; DA - Danish; EN - English; LI - Lithuanian; LA - Latvian; SV - Slovenian; CZ - Czech; SL - Slovak, PL - Polish; GR - Greek.

Greenberg (1956) indroduced a monolingual weighted index, which, in simple words, accounts for an average linguistic distance between randomly chosen individuals within the society, In addition to the notation of the previous subsection, let d_{ki} denote the linguistic distance bertween two gtoups $i, k, = 1. \ldots, K$. The distance could be derived via any of the methods described earlier in thuis section. Then we have index B whose formal presentation is given by

$$B = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k n_i d_{ki}.$$

It is quite easy to see that B-index is a generalization of the a-index, dischotomous distances are replaced by an arbitrary distnace metric. Indeed, if we impose a dichotomy on A, i.e., assume that $d_{ki} = 0$ if $i = k$ and $d_{ki} = 1$ if $i \neq k$, then B turns into A:

$$B = \sum_{k=1}^{K} n_k(1 - n_k) = 1 - \sum_{k=1}^{K} n_k^2.$$

Bossert et al. (2011) offer an axiomatic foundation of a variant of the B-index. They, however, rely on primitives of individuals, rather than ethnic groups, where individuals are not pre-assigned to exogenously determined groups within a society.

It worth to pointing out that the emprical analysis relying on the B index requires a more extensive dataset than simply using the index A. However, the effort could be worth it, as Desmet et al. (2009) in their cross-country analysis of redistribution patterns, clearly indicate a much stronger explanatory power of index B. The importance of incorporation of linguistic distances in the definition of societal indices is also supported by Dower et al. (2017) in the context of their analysis of the Sri Lanka conflict.

In many applications of diversity indices, especially, with regard to provision of public good, redistribution, and intensity of conflicts within a country, one has to take into account the special status of various regions, Of particular importnace is center-periphery relations or even tension between the centeral and prerepheral regions. To highlight this point, Desmet et al. (2009, 2017) have assigned a special role to one of the regions, say, region 1, called the "center" and denoted by c. The other $K-1$ regions are assumed to be "peripheral". In calculating a variant of B-index, only the distances between the center and perphery are accounted for. whereas the bilateral links between any two perpheral regions are disregarded. That is,

$$CP = n_c \sum_{k=2}^{K} n_k d_{kc}.$$

Polarization indices

The notion of polarization, and the indices it generates, adds an additional element to the build-up that led to the introduction of A- and B-index indices. To recall, both of those indices rely on the notion of pre-existing partition into distinct linguistic groups. The polarization approach adds an important facet of individual self-identification for members of the society. The self-identification comes through in two ways. One, is the strength of identification with others in one's own group, another is alienation toward the others. Esteban and Ray (1994) defined a notion of *social effective antagonism* that combines both identification and alienation, that depend only on the size of the groups. Esteban and Ray examine income pollarization when the groups are identified by their income levels and the distances are income differentials between the groups. The fuctional form of their index bult on axiomatic foundations, is close to that of the index B:

$$P = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^{1+\alpha} n_i d_{ki},$$

where α is a positive parameter ranging between 1 and 1.6.

For $\alpha = 0$, the index P is a Gini coefficient of income inequality.

Consequently, Geng (2012) has imposed additional axioms to sprink the range of α's to a single point, $\alpha = 1$.

The Esteban and Ray approach can be adjusted to incorporate linguistic distanced (Montalvo and Reynal-Querol (2005), Desmet et al. (2017) and Dower et al. (2017)). For $\alpha = 1$, Reynal-Querol (2002) offered a dichotomous version of this index that assumes that d_{ki} is equal to zero if $k = i$, and d_{ki} is equal to one $k \neq i$. The Reynal-Querol index then obtains the following functional form

$$RQ = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^2 n_i = \sum_{k=1}^{K} n_k^2 (1 - n_k).$$

It is worth pointing out the intuitive difference between two dichotomous indices A and RQ. To recall, index A is determined by the probability that two randomly chosen individuals belong to two different groups. Thus, the value of A is the sum of the terms $n_k(1 - n_k)$, each identifying the probabily that one individual belong to group k, while the other does not.

The value of RQ is determined by the probability that among three randomly chosen individuals two belong to one group, while the third belongs to other. Thus, RQ is represented by the sum of the terms $n_k^2(1 - n_k)$, that is, two individuals belong to group k, while the third does not.

Moreover, in the Esteban and Ray model the identification and alienation depend only on the sizes of relevant groups. One may assume that some additional factors, such as linguistic proximity or historical path could play an important part in determining the degree of identification and alienation. In their sttudy of the protracted war in Sri Lanka, Dower et al. (2017) introduce an ethnolinguistic polarization measure that takes into account the impact of historical factors on inter-group relations driven by different patterns of English language acquisition in the colonial era. They used the Reynal-Querol variant of the Esteban-Ray index (both in dichotomous and nondichotomous forms) by comparing its value across all dostricts j:

$$D^j = \sum_{k=1}^{K} \sum_{i=1}^{K} (n_k^j)^2 n_i^j d_{ki},$$

where n_k^j and n_k^j denote the fraction of linguistic groups k and i, respectively in district j, whereas d_{ki} is the linguistic distance between groups k and i. However. that linguistic takes into account changes groups' English proficiency in the precolonial period. By examining the protracted war in Sri Lanka and applying that measure to a dataset on victims of terrorist attacks by district and war period, they find that increasing the share of English speakers resulting from colonial times in each district would result in increasing the number of war victims.

Disenfranchisement indices

The notion of linguistic disenfranchisement has been formally introduced in Ginsburgh and Weber (2005) to provide a rigorous framework to examine linguistic discrimination and denial of linguistic rights. Language matters. Two observations: "like religion, language does not lend itself easily to compromise."(Laponce (1992).)

"Language may be the most explosive issue universally and over time. This mainly because language alone, unlike all other concerns associated with nationalism and ethnocentrism ... is so closely tied to the individual self. Fear of being deprived of communicating skills seem to rise political passion to a fever pitch." (Bretton (1976).)

. In Skutnabb-Kangas and Phillipson (1989) formulation, it amounts to introduction of "ideologies and structures which are used to legitimate, effectuate, and reproduce unequal division of power and resources (both material and non-material) between groups which are defined on the basis of language". The denial of linguistic rights was and is very common phenomena (see Ginsburgh and Weber (2011)). To a different degree, it covers all continents. Probably, the most severe example is Africa, where as Skutnabb-Kangasand Phillipson (1989) point out, "The majority of Africans are governed in a language that they do not understand".

Sri-Lanka.

Let us turn to William Shakespeare's play King Richard the Second. The King and his uncle John of Gaunt try to convince Henry Bolingbroke, Gaunts son and Kings first cousin, and Thomas Mowbray, the Duke of Norfolk, to stop quarreling. The King was unable to calm the warriors down and delivers punishments to both Thomas and Henry. Mowbray is banished from England forever and Bolingbroke for ten years. The interesting part is Thomas reaction to his banishment. He does not lament about the loss of land or status but rather talks about the inability to speak his native language in exile:

A heavy sentence, my most sovereign liege,
and all unlookd for from your Highness mouth,
A dearer merit, not so deep a maim
As to be cast forth in common air,
Have I deserved at your Highness hands.
The language I have learnd these forty years,
My native English, now I must forego;
And now my tongues use is to me no more
Than an unstringed viol or a harp,
Or like a cunning instrument casd up,
Or, being open, put into his hands
That knows no touch to tune the harmony:
Within my mouth you engaold my tongue,
Doubly portcullisd with my teeth and lips;
And dull, unfeeling, barren ignorance
Is made my gaoler to attend me.

Finally, the complaint of despair and hopelessness, recognizing Thomas inability to master another language (at the time, he was only forty years old):

I am too old to fawn upon a nurse,
Too far in years to be a pupil now.
What is thy sentence, then but speechless death,
Which robs my tongue from breathing native breath?

To provide a formal response in our framework, consider the set L of all language spoken in the country. It may coincide with all K native languages we consider, but may also include some other world languages that are pracrtically not spoken as a native tongue, like English in Russia or China. The additional dadta we need is the language proficiency. That is, through surveys or otherwise, we assume to know the linguistic repertoire of every individual in the society, including her native language.

Connsider a subset T which we consider as a possible candidate to form the set of official languages to be used for official documentation, television and educational purposes. For each individual m we define a degree of her disenfranchisement, based on the languages she speaks and the way the disenfranchisement is mesured. We then aggregate the disenfranchisement levels across the entire society.

For each m $k(m)$ denotes her native tongue and $L(m)$ the set of all languages she speaks. There four ways to measure the disenfranchisement:

• DN - dichotomous distances and native languages. That is, the disenfranchisement for m iis 0 if her native language $k(m)$ is included in T. It is one, if $k(m)$ is not included in T. The index DN is given by

$$DN = \sum_{k:k\notin t} n_k.$$

.

• NN - nondichotomous distances and native languages. The disenfranchisement for m is again zero if her native language $k(m)$ is included in T. If it is not, the disenfranchisement is equal to the minimal distance between $k(m)$ and the languages in T. Denote that distance by $d^{nat}(m,T)$

$$d^{nat}(m,T) = \min_{l\in T} d(m,l)$$

and the index DN is given by

$$NN = \sum_{1}^{N} d^{nat}(m,T).$$

.

• DA - dichotomous distances and all spoken languages. That is, the disenfranchisement for m is 0 if m speaks at least one language in T and 1, otherwise. The index DA is given by

$$DA = \sum_{m:L(m)\cap T\neq\emptyset}$$

• NA - nondichotomous distances and all spoken languages. That is, the disenfranchisement for m iis again 0 if m speaks at least one language in T. If not, the disenfranchisement is equal to the minimal distance between the set of languages she speaks and the languages in T. Denote that distance by $d^{all}(m,T)$.

$$d^{all}(m,T) = \min_{q\in k(m),l\in T} d(q,l)$$

Then

$$NA = \sum_{1}^{N} d^{all}(m,T).$$

Penrose Law

To mitigate the excessive contribution to the aggregate index by large groups (say, Germany vs Malta in the EU), the Penrose Law adds an an additional factor that relates to population sizes. For example, the adjusted NN index would be:

$$NN = \sum_{1}^{N} \frac{d^{nat}(m, T)}{\sqrt{p_m}},$$

where p_m is the population size of the group to which m belongs.

Applications

Most of the empirical research in the field has utilzed the A-index. Nevertheless, as Desmet et al. (2009) and Bossert et al. (2011) indicate, in several applications B index has a stronger explanatory power than A-index.

We mainly focus here on the impact linguistic (and ethnolinguistic) diversity on various economic and political outcomes: growth, quality of the governmental institutions, and poltical stability.

Easterly and Levine (1997) demonstrate that etholinguistic diversity account for much of Africa's growth tragedy. They compare growth rates in East Asia for the period of 1969-1990 and argue that about 40 percent of the annual growth differential between two regions (abot 3.5 percent) can be attributed to affects of fractionalization.

Those conclusions are supported by Alesina et al. (2003). who showed that linguistic diversity, has a significantly negative effect on income growth in a panel of countries. Collier (2001) indicates, however, as well as Alesina and La Ferrara (2005), that the adverse effect of diversity on growth is mitigated in democratic societies, while it remains strong under dictatoships (Lian and O'Neal, (1997).

Campos et al. (2011) also demonstrate an insignificant

impact of diversity on economic growth in twenty six former communist countries of Eastern and Central Europe, Central Aton and Mongolia in the period of 1989-2007, during and after the collapse of the Soviet Union. However, Campos et al. show by focusing on dynamic aspects of fractionalization and shifts in the value of index A, treating diversity as an enogenous variable restres a strong link between diversity and growth.

Mauro (1995) argues that a high degree of ethnic fragmentation is correlated with institutional inefficincy, poltical instability and corruption. The reason is that, in a fragmented society various groups fiercely compete for their piece of the cake, thus, raising increasing the amount of lobbying and the scope ofinefficiency and corruption.

La Porta et al. (1999) indicate that ethnolinguistic fractionalization, has a negative impact on various public services and goods, including literacy rates, infant mortality, education and infrastructure. By using a more extensive dataset on ethnic, linguistic and religious fractionalization and polarization, Alesina et al. (2003), in general, confirm the results of La Porta et al. Desmet et al. (2009) show that values of index B, that account for the distance between groups, is negatively associated with redistribution, measured by the share of transfers and subsidies in GDP.

Annett (2001) shows that ethnic fractionalization breeds

political instability, which as Alesina et al. (1996) in the panel of 113 countries point out, have a negative impact on growth. Annett also identifies an excessive government consumption that emerges in highly fractionalized societies, as another impedient to economic growth. Alesina et al. (1999) show more fragmented cities, metropolitan areas and urban counties in the United States, are associated with lower provision of education, roads and seeage. Thier results have been confirmed for other countries by Kujis (2005).

One has to point out that, while most of the cross-country results linking diversity and economic outcomes show the negative link between the two, the conclusions are somewhat different on the city and the firm level. Lazear (1999) and Prat (2002) argue that a successul team should exhibit a certain degree of cultural and linguistic diversity. Collier (2001) also argues that sometimes diverse societies perfform better than more homogeneous ones. In his analysis of 50 US cities, Florida (2002) shows metropolitan areas with a higher degree diversity with regard to education, culture and native language, exhibit a higher level of economic development. At the cross-city level Ottaviano and Peri (2006), who examined 160 cities in the U.S. over the 1970-1990 period, also point to a positive effect of cultural and linguistic diversity on the productivity of U.S. workers.

We would like to conclude this section by pointing out

that sometimes the choice of index to be used in the analysis could be endogenous. We offer several examples

• Fearon and Laitin (2003), Collier and Hoeffler (2004) indicate that A-index does not provide an obvious link for likelihood of civil conflicts. The analyssis here requires an alternative index. and, indeed, Montalvo and Reynald-Querol (2005) argue that the polarization index RQ does explain the incidene of civil wars. Moreover, Montalvo and Reynald-Querol (2002) show that a higher level of RQ points out to a longer civil conflicts.

• The fractionalization index A^{α}, defined earlier in this section, may indicate that different societies exhibit different levels of α. There are different attitudes towards immirgants and the importance of diversity for well-being varies across the communities. It is up to a researcher to identify a proper value of α for each community, while refuting the approach that considers all societies being equal.

• Desmet et al. (2015) construct various linguistic partitions for examining various question. They argue that the data should determine which level of aggregation to select and show that with regard to civil conflict and redistribution, deeper cleavages, and, thus, coarser partitions are more significant. That is, the historical path of language development matters. In contrast, for economic growth and provision public goods, the diversity measure

based on more disaggregated classifications of linguistic groups, capturing finer distinctions between languages, are important correlates of growth and public goods provision both in terms of statistical significance and in terms of economic magnitude.

Disenfranchisement indices

EU

NN: GE $\rightarrow$ GE-IT $\rightarrow$ GE-IT-PL $\rightarrow$ GE-IT-PL-EN

Nondichotomous -native: If one language to be chosen, it is German, if two, German-Italian, if three then German-Ialian-Polish, if four, German-Italian-Polish-English.

NA: EN $\rightarrow$ EN-FR $\rightarrow$ EN-FR-GE $\rightarrow$ EN-FR-GE-PL

Nondichotomous - allspoken languages, If one language to be chosen, it is English, if two, English-French, if three then English-French-German, if four, English-French-German-Polish.

Social Interactions

Consider the following simple model with a finite set of players denoted by $N = \{1, 2, \ldots, n\}$.

X_i is a set of possible actions of player i (infinite or infnite) (a compact set in an Euclidean space).

The choices of individual strategies $x_i \in X_i$ for all $i \in N$ yield the n-dimensional strategies profile $bfx = (x_1, x_2, \ldots, x_n)$.

Profle $\mathbf{x}$ generates a partition $\pi(x)$ of N, where players i and j belong to the same group in $\pi(\mathbf{x})$ if their choices are identical, i.e., $x_i = x_j$.

We denote by $Si(\mathbf{x})$ the group in $\pi(x)$ which contains player i.

The society is partitioned into peer groups, where each player i has a peer group $P_i \subset N$ that may inflluence i's choices.

The payoff $U_i(\mathbf{x})$ of player$i$, is the sum of three terms:

$$U_i(\mathbf{x}) = V_i(x_i) + \sum_{j \in P_i} W_i^j(x_i, x_j, d_{ij}) + h(x_i, |S^i(\mathbf{x})|)$$

where d_ij is the distance between players i and j.

The frst term describes the intrinsic taste of player i for her chosen action x_i.

The other two terms, unlike the frst one, reflect complementary aspects of social interaction.

The second term represents bilateral social interactions of player i with her peer group.

The last term, which captures a *conformity* facet of social interaction and depends on the number of players who have chosen the same action x_i.

We impose the following assumptions on players' payoff functions:

Assumption A1 - Symmetry. $W_i^j(x_i, x_j, d_{ij}) = W_j^i(x_i, x_j, d$ for every $i, j \in N$, every $x_i \in X_i$ and every $x_j \in X_j$. That is, the impact of bilateral interaction between any two players i and j, represented by the values of W_i^j and W_j^i, is identical for both players. (It is 0 for players in different peer groups.)

Assumption A3 - Conformity. $h(x, S) \leq h(x, T)$ for all $x \in X$ and all groups $S, T \subset N$ with $|S| \leq |T|$.

Assumption A2 - Semi-continuity. The functions V_i, W_j^i and $h(\cdot, r)$ (the latter for all $r \leq n$) are upper semicontinuous on their corresponding domains.

Theorem: Under A1-A3, every game from the described class admits a Nash equilibrium in pure strategies.

If the strategies sets are finite, the continuity requirement A3 is vacuous and the conformity condition A2 is redundant:

Corollary: If the strategy sets are finite, then under A1, every game from the described class admits a Nash equilibrium in pure strategies.

- Theory (Blume and Durlauf (2001))
- To start a diet (Harris and Lopez-Varcarcel (2004))
- To quit smoking (Jones (1994))
- To become member of a club or of organization (Dixit (2003))
- To participate in criminal activities (Glaeser, Sacerdote and Sheinkman (1996))
- To enter an industrial alliance (R&D consor- tium) (Axelrod et al. (1995))
- To participate in a riot or a strike (Schelling (1978), Chwe (1999, 2000), Granovetter (1978))
- To chose a side in an international conflict (Altfeld and Bueno de Mesquita (1979), Axelrod and Bennett (1993), Galam(1996))
- To buy a house in a specific residential area (Schelling (1969, 1971))
- To display a national flag in one's window (Chwe (2006))
- International alliances (Axelrod and Bennett (1993)).