The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications

by Pedro Conceição and Pedro Ferreira

Pedroc@uts.cc.utexas.edu
LBJ School of Public Affairs
The University of Texas at Austin
Austin, Texas 78713

pmf@rpcp.mit.edu
Internet and Telecoms Convergence Consortium
Massachusetts Institute of Technology
E40-218, One Amherst Street
Cambridge, MA 02139-4307

UTIP Working Paper Number 14

February 29, 2000

Abstract

Growing interest in inequality has generated an outpouring of scholarly research and has brought many discussions on the subject into the public realm. Surprisingly, most of these studies and discussions rely on a narrow set of indicators to measure inequality. Most of the time a single summary measure of inequality is considered: the Gini coefficient. This is surprising not only because there are many ways to measure inequality, but mostly because the Gini coefficient has only limited success in its ability to generate the amount and type of data required to analyze the complex patterns and dynamics of inequality within and across countries. Often, in defense of the use of the Gini coefficient, it is argued that this popular indicator has a readily intuitive interpretation. While from a formal point of view most measures of inequality are closely interrelated, at an intuitive level this interrelationship is rarely highlighted. This paper suggests an intuitive interpretation for the Theil index, a measure of inequality with unique properties that makes it a powerful instrument to produce data and to analyze patterns and dynamics of inequality. Since the potential of the Theil index to generate rich data sets has been analyzed elsewhere (Conceição and Galbraith, 1998), here we will focus on the intuitive interpretation of the Theil index and on its potential for analytical work. The discussion will be accompanied throughout with empirical applications, and concludes with the description of a simple software application that can be used to compute the Theil index at different levels of aggregation of the individuals that compose the distribution.

1- INTUITIONS: MEASURING THE WORLD DISTRIBUTION OF INCOME

> [The Theil index can be interpreted] as the expected information content of the indirect message which transforms the population shares as prior probabilities into the income shares as posterior probabilities.
>
> Henri Theil (1967:125-126)

> But the fact remains that [the Theil index] is an arbitrary formula, and the average of the logarithms of the reciprocals of income shares weighted by income is not a measure that is exactly overflowing with intuitive sense.
>
> Amartya Sen (1997:36)

A measure of economic inequality provides, ideally, a number summarizing the dispersion of the distribution of income among individuals[1]. Such a measure is an indication of the level of inequality of a society. Building on this intuition, most discussions of inequality indicators depart from an individual-level analysis. When the distribution of income is equal, each person has the same share of the overall available income, and the measure of inequality assumes its absolute minimum. Deviations from this equal distribution of

[1] We will discuss only *objective* measures of inequality, in the sense proposed by Sen (1997). The alternatives to the objective measures are what Sen calls normative measures of inequality, which have imbedded some notion of social welfare. Normative measures of inequality include, in some sense, an ethical evaluation of some kind, while objective measures, in themselves, are "ethically" neutral. Objective measures of inequality employ statistical and other types of formulae that account for the relative variation of income among individuals or groups of people.

income, when one or more individuals have a higher share than others, are captured by an increase in the level of the inequality measure.

This type of individual-level discussion of inequality provides a good intuitive framework for understanding some measures of inequality. For example, drawing from well-known statistical formulae, the variance can be used as a measure of inequality. Indeed, the variance (the sum of the squared differences between the income of each individual and the mean) is a common statistical measure of dispersion in a distribution. If all individuals have the same share of income, then each must have the mean income, and the variance is zero. If some individuals have a share of income that is different from the mean this is captured by the variance, and the larger the deviation from the mean the larger the impact in the increase of the level of the variance[2].

This individual-level discussion is not helpful, though, to acquire an intuitive understanding of other measures, such as the Gini coefficient, for example. The easiest intuitive interpretation of the Gini coefficient invokes the Lorenz curve, as we will explore below. Rarely one sees the Gini coefficient being motivated from an individual-level type of discussion, although this is entirely possible to do. Similarly, departing from an individual-level type of analysis does not provide the best intuition to interpret the Theil index. Theil's (1967) elegant first presentation of this measure of inequality was based on statistical information theory. Theil's original presentation of his inequality indicator is not intuitively appealing, as the quotes above suggest. Still, most of the times Theil's original discussion is replicated when the measure is introduced (as in Sen, 1997: 34-36). In other cases, no intuitive motivation is given, and it is simply mentioned that the Theil measure is based on information theory (as in Alison, 1978: 867).

Our objective in this section is to provide a new way to approach the derivation of the Theil index, which will simultaneously suggest a new intuitive interpretation and a more direct presentation of its many advantages vis-à-vis other inequality measures. To do so,

[2] The variance, it turns out, is not such a good measure of inequality, since it does not comply with other requirements commonly demanded from inequality measures.

instead of departing from an individual-level analysis we will start assuming that individuals are grouped. Thus we will be looking primarily at inequality between groups of individuals, and not at inequality between individuals. The criterion for grouping is irrelevant here. It could be one of a series of exogenous factors according to which we have an interest in grouping individuals for analytical purposes. Examples include geographic units, race, ethnicity, sex, education level, urban vs. rural population, or even income intervals. If we take geographic units, for instance, we could be doing so because we were interested in variations in the distribution of income across countries, or across states in the US.

Beyond the plausibility of being interested in having grouped-level data for analytical reasons, there is a more pragmatic rationale for this approach. At the outset of this section, we mentioned that, ideally, a measure of inequality would provide a number indicating the dispersion of income among individuals. In practice, however, this objective is virtually impossible to accomplish. Information on individual income for every single citizen of a country is simply not available, at least with high frequency. Sampling and household surveys are often used instead as the raw information taken to compute "comprehensive" inequality measures, but these are approximations, a fact rarely highlighted. Even the Lorenz curve is usually constructed by grouping individuals in income intervals.

In summary, we are arguing that analyzing inequality often requires grouping individuals and that, even when we are interested in inequality at the most fundamental level (between individuals), the reality of data collection almost always entails some level of aggregation, particularly if one is interested in frequent sampling. Thus, it will be important to differentiate in the forthcoming discussion three "types" of inequality: overall comprehensive inequality between individuals (total inequality – almost always unobservable), inequality between the groups (between-group inequality) and the residual or remaining inequality among individuals that is not accounted for by the between-group inequality.

We will accompany the discussion with an illustrative example. We will use the GDP and population data for 108 countries from the Heston and Summers (1991) Penn World Tables Mark 5.6[3]. Let us suppose we were interested in a measure of world inequality indicating the global variation in the distribution of income. As a first approximation, let us consider a simple division between rich and poor countries for 1970. We first rank the 108 countries according the their level of GDP per capita, and place the first half in the "rich countries" group, and the second half in the "poor countries" group.

An equal distribution of income *between* the two groups requires the comparison of the population share with the income share of each group. In fact, the condition to have equality between groups is slightly weaker than the one we would have if we were comparing two individuals. In the latter case, we would need a fifty-fifty distribution of income between two individuals to have equality. But since we are comparing groups, all we need to have is the population share of each group equal to that same group's income share; this share does not have to be 50% in the current case where we have only two groups. We should stress that we are considering only the inequality *between* groups, not total inequality[4].

Figure 1 shows the population and the income shares of each of the two country groups. The richest 54 countries (richest half) have about 36% of the world's population, while the countries included in the poorest group account for the remaining 64%. However, the rich countries have 82% of the world's income. In other words, in 1970 there was a large inequality in the distribution of income between these two groups. The "fair share" of

[3] GDP is in 1985 PPP expressed in dollars for all countries, which allows us to aggregate income across countries.

[4] In fact, since our criterion to distinguish countries was per capita GDP, to have total equality among individuals in the world we would have to have a 50% distribution for each group, but it is easy to see that this is not always the case. Consider, for example, what would happen if we had divided the countries between those that are in the African continent, and the non-Africa countries: equality *does not* require that we have 50% of the income and population in Africa, only that its share of income be the same as its share of population. A further example below will make this point clearer.

income for the rich countries – that is, the income share in an equal world – should be 36% (equal to the population share), but it was in fact more than two times as large.

The representation in Figure 1 provides a graphic illustration of the inequality in 1970 between the two groups of countries. To summarize textually the inequality expressed in Figure 1 we can say that 36% of the world's population lived in 1970 in countries that had 82% of the world's available income.

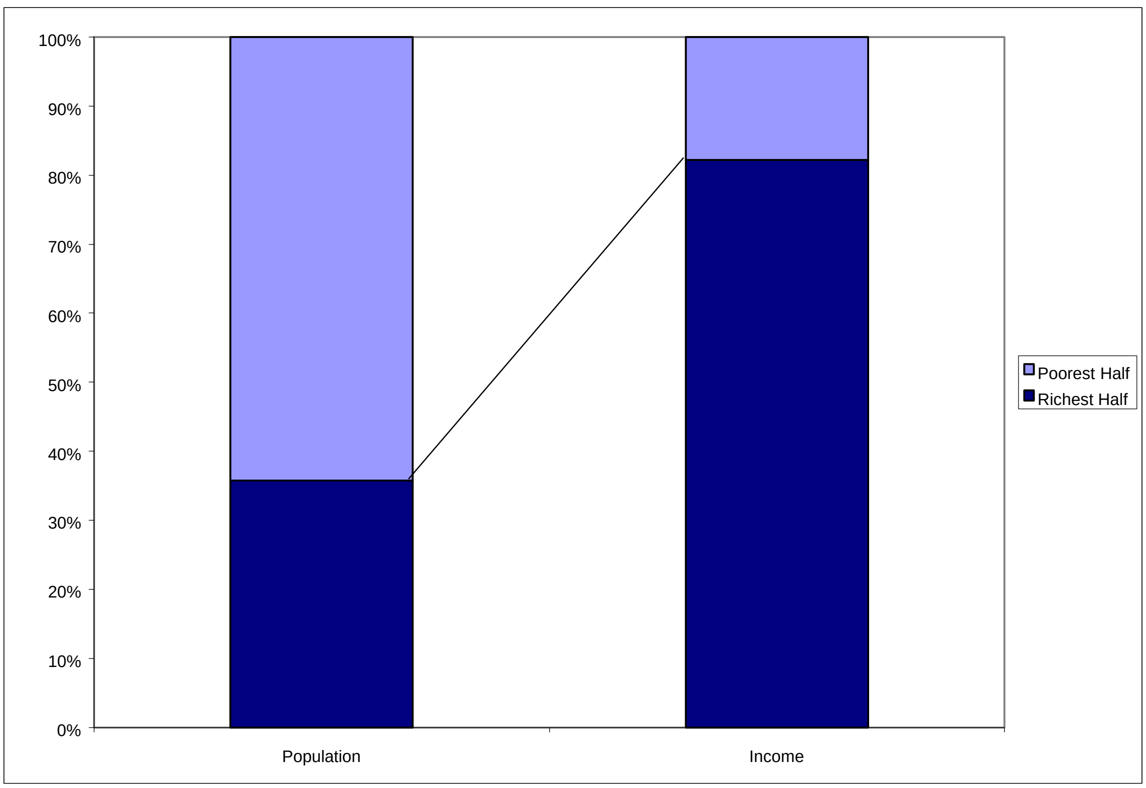

Figure 1- World Inequality: Population and Income Shares of the 54 Richest and 54 Poorest Countries in the World in 1970.

But neither the graphic representation nor the textual description, compelling as they may be, provide us with a measure of inequality. To clarify what we are looking for, some symbolic representation helps. As we said above, total inequality (inequality among all individuals in the world: I_{World}) is composed of the inequality between the groups we are considering (I'_{World}) plus the remaining inequality that is not accounted for by the between group inequality:

[1] $$I_{\text{World}} = I'_{\text{World}} + I_{\text{remaining}}$$

We should note that the remaining inequality is certainly very large, and I'_{World} provides only a lower bound. For now we will concentrate on looking for a measure for I'_{World}. As we go along, we will discuss how we can go about determining $I_{remaining}$.

Intuitively, a measure for I' should give us an indication of the discrepancy between the population share and the income share of each group. Let us call the income shares w_{rich} and w_{poor} and the population shares n_{rich} and n_{poor}; the values are shown in Table 1.

Table 1- Income and population shares for the richest and poorest countries in the world in 1970.

Income shares		Population shares	
w_{rich}	0.82	n_{rich}	0.36
w_{poor}	0.18	n_{poor}	0.64

If we are interested in getting to a measure of inequality, how can we summarize the discrepancy between w_{rich} and n_{rich} in a single number? One easy way is to compute the absolute value of the difference:

$$I'_1 = |w_{rich} - n_{rich}| = |.82 - .36| = .46$$

So one possible measure of inequality, which we will call I'_1 – our first inequality measure – could be defined in this way: take the income and population shares of the group with highest income share; subtract the population share from the income share and take the absolute value; the resulting number, a measure of the discrepancy between the shares of people and of income in this group, is an indicator of inequality. Note that the higher the discrepancy between population and income shares, the higher is our measure of inequality. And also if we have $w_{rich} = n_{rich}$ then our measure is zero. Since our measure can never be negative, when we have perfect equality between groups I'_1 attains its minimum: zero.

However, taking only one group ignores valuable information on the distribution of income between other groups. In our current example, this is not such a big problem, since we have only two groups, but it could be if we had more groups. By taking only the highest income share group, our measure of inequality would be ignoring the distribution of income between the remaining groups. Following the same logic as above, an easy way to include all groups is to define a measure of inequality, I'_2, which sums the absolute values of the differences between income and population shares for every group. In our case, for 1970:

$$I'_2 = |w_{rich} - n_{rich}| + |w_{poor} - n_{poor}| = |.82 - .36| + |.18 - .64| = .46 + .46 = .92$$

Again, I'_2 is always positive, and it is zero (minimum value) when population and income shares are the same for each and every group.

So far we should have the intuition for what we are looking for: a measure of inequality that highlights the fact that some groups have a higher (lower) share of income than their "fair share" of income, given their population shares. If we manage to build a measure of inequality that is always positive, then when we have perfect equality this measure should be zero. Continuing with our search, we should first note that despite including all the groups, the second measure of inequality, I'_2, does not add much to the first, I'_1. In fact, it is easy to see that with two groups I'_2 merely doubles the value we get from I'_1. What we need to do is make sure that our measure of inequality "understands" that the richest half and the poorest half are different groups, so that our measure of inequality does not merely duplicates what one gets when only one group is considered. Note the symmetry between the differences of the shares of each group: the difference is the same, in absolute terms.

One way to achieve the goal of differentiating the groups (in a way, of breaking the symmetry between the groups) is to multiply each difference by the share of income of the group it refers to[5]. By doing so, we take a first step in incorporating the fact that the groups considered are different and that one of the differences comes precisely from their shares of income. The measure where each difference is weighted by the income share is:

$$I' = w_{poor} \text{ x } | w_{poor} - n_{poor} | + w_{rich} \text{ x } | w_{rich} - n_{rich} |$$

but this is simply equal to $|w_{rich} - n_{rich}|$, which is I'_1 [6]. An alternative would be to consider the differences without taking the absolute value, but this would make the measure negative for some range of the differences between the shares.

It is important to recall that our objective is to “produce” an inequality measure that translates the discrepancies, for each group, of the income and population shares into a number. Our first attempt was based on **differences** between the income and population shares of each group. Another option would have been to use the **ratio** between the income and the population shares of each group. To use the ratio of the shares is slightly less intuitive. In particular, when the income share for a group is equal to that group’s population share – so that the group has its “fair share of income” – the ratio is one. This

---

[5] Instead, we could have multiplied the difference between the shares of income and population for each group by its share of population. The consequences of following this option will be explored later.

[6] Since $w_{poor} + w_{rich} = 1$ and $n_{poor} + n_{rich} = 1$. Again, if we were considering more than three groups this result would not be valid, but the point remains that the structure of this inequality measure does not fundamentally change the nature of the first measure.

means that incorporating ratios of shares into a measure of inequality must be performed in such a way that the contribution to the inequality measure when the shares are equal is zero (and not one). Obviously, the easiest way to do this is to subtract the number one from the ratio. If we take the absolute value of the sum of the differences between the ratio of the shares and one (to guarantee that the measure remains positive), we obtain a third measure of inequality:

$$I'_3 = |(1 - w_{rich} / n_{rich}) + (1 - w_{poor} / n_{poor})| = .58$$

Clearly, I'_3 is not that different from I'_2, since it can be also expressed as

$$I'_3 = |(1 / n_{rich}) (n_{rich} - w_{rich}) + (1 / n_{poor}) (n_{poor} - w_{poor})|$$

Which is, again, a weighted summation of the differences of the shares, where now the weights are the inverse of the population shares.

If we want to devise a measure of inequality based on the ratio of the shares that yields zero when the group shares are equal, a stronger transformation, and less intuitive a priori, is to apply a logarithmic transformation before each ratio. We will see how this transformation provides a measure of inequality with many interesting properties, but for now we need only to stress that applying the logarithm (a monotonic transformation) before each ratio does, indeed, give zero when the shares are equal and, furthermore (if weighted by the income shares) is always positive. Thus, our fourth proposed measure of inequality is:

[2] $$T' = w_{rich}.[\log (w_{rich} / n_{rich})] + w_{poor}.[\log (w_{poor} / n_{poor})] = .46$$

This expression is equivalent to

$$T' = w_{rich}.[\log (w_{rich}) - \log (n_{rich})] + w_{poor}.[\log (w_{poor}) - \log (n_{poor})]$$

showing that T' can also be understood as the summation of the (weighted) difference of the logarithms of the shares, instead of the direct difference of the shares as in all the other previous measures considered here. The weighting of the difference by the income shares of each group guarantees that T' is always positive, so that we do not have to force the usage of absolute terms (which provides for an easier algebraic manipulation of the measure). The reason why this guarantees that the measure is always positive will become apparent later.

The formula for T', thus, is similar (conceptually) to the one we used to define the other measures of inequality. The intuitive principle is the same: to highlight the discrepancy between income and population shares. There are two important distinctions between T' and I'_2. First, in T' we subtract the logarithms of the shares, and not the shares directly. Therefore, the symmetry that existed between the groups in I'_2 is now gone. The logarithm staggers the shares, and clearly separates the difference between shares in the poor from the difference in the rich group. The second difference is that instead of using absolute values, we multiply the difference in the logarithms of the shares by the share of income in each group. Failing to do this would result in negative values.

This measure we just defined, T', is, indeed, the Theil index. It is not in the form in which it is normally written, but we will get there. For now, let us look at the behavior of T', compare it with the behavior of the other measures, and show that it has the basic properties we have been demanding. Figure 2 plots the evolution of T', of I'_1 and of I'_3 (I'_2, as we saw, is just two times I'_1). Each curve corresponds to a hypothetical distribution of income supposing that the distribution of population remains constant, and as presented in Table 1. The income share of the “rich” half of world countries is then changed continuously from 0 to 1 (which entails a simultaneous change of the “poor” group from 1 to 0). This varying income share is the “independent variable” represented in the horizontal axis. Therefore, each line shows how each inequality measure responds to changes in the shares of income.

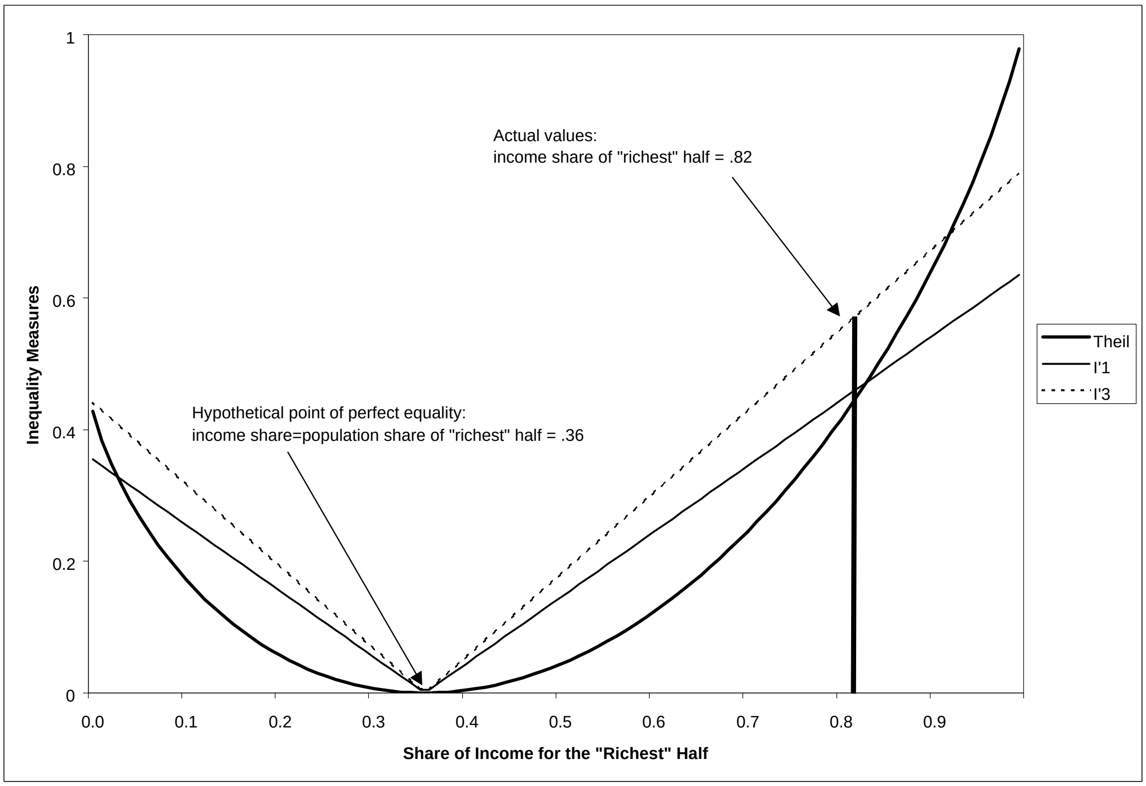

Figure 2- Simulation of the Evolution of Inequality Measures as the Shares of Income Change.

The actual values are indicated by a thick vertical line, at the point where the share of income in the rich countries is .82. Now suppose that the share of income of rich countries increases, which means that we are moving towards the right of the .82 point in the horizontal axis. We can see that all three measures increase, as the discrepancy between the population shares and income shares grows wider. Note that the behavior of I'_1 and I'_3 is linear, as was to be expected, since they represent the difference between the share of income and the population share. T', in contrast, grows much more than linearly. In fact, its slope is steeper the larger the income share of the rich group is. This behavior reflects a well-known property of the Theil index: the large sensitivity to income transfers from the poor to the rich. Note that as we move along the horizontal axis to the right such a transfer from poor to rich must occur, in order for the income share of the rich countries to increase. As the transfer from poor to rich grows, so increases the steepness of the Theil line. The linear measures are insensitive to this type of behavior.

Going towards the left of the point with the actual share of income, we see that all measures of inequality decrease, as the income share of the rich-country group decreases. Again, we see that the Theil index decreases more than linearly, reflecting its sensitivity to income transfers, now from the rich to the poor. When the income share equals the population share of .36, all measures of inequality are zero. And as we move towards the left of the point of hypothetical equality, inequality starts increasing, as the share of income in the poor countries actually now surpasses that of the rich countries (thus the distinction between “rich” and “poor” becomes arbitrary in the hypothetical arena, and the names are used merely as tags). In summary, T' has the desired properties of an inequality measure, plus a bonus: its sensitivity to income transfers from poor to rich.

It is important, at this stage, to revisit the fundamental idea behind the intuition for the Theil index. The idea is that the Theil index provides a measure of the discrepancies between the distribution of income and the distribution of population between groups. Essentially, the Theil index compares the income and population distribution structures by summing, across groups, the weighted logarithm of the ratio between each groups income and population shares. When this ratio is one for some group, then this group’s contribution to inequality is zero. When all the groups have a share of income equal to their population share, the overall Theil measure is zero. We also saw that, as an added benefit, the Theil index is sensitive to transfers of income from poor to rich, and that this sensitivity increases with the width of the difference between rich and poor.

We now move towards a deeper exploration of the Theil index. While the Theil index is always positive, the contributions of each group (the terms in the summation) can be negative. Indeed, for some groups, they have to be negative. This is easy to understand with our example. Since the ratio between income and population shares for the richest group of countries is larger than one (.82/.36 = 2.28), the logarithm of this ratio is larger than zero. However, the same ratio for the poorest group has to be lower than one (.18/.64 = .28) remains below one so as long as the income share is lower than the population share. Consequently, the logarithm is negative. The way in which the two groups' contributions to the Theil index work can be understood with the help of Figure 3, which is similar to Figure 2 in the way it was constructed.

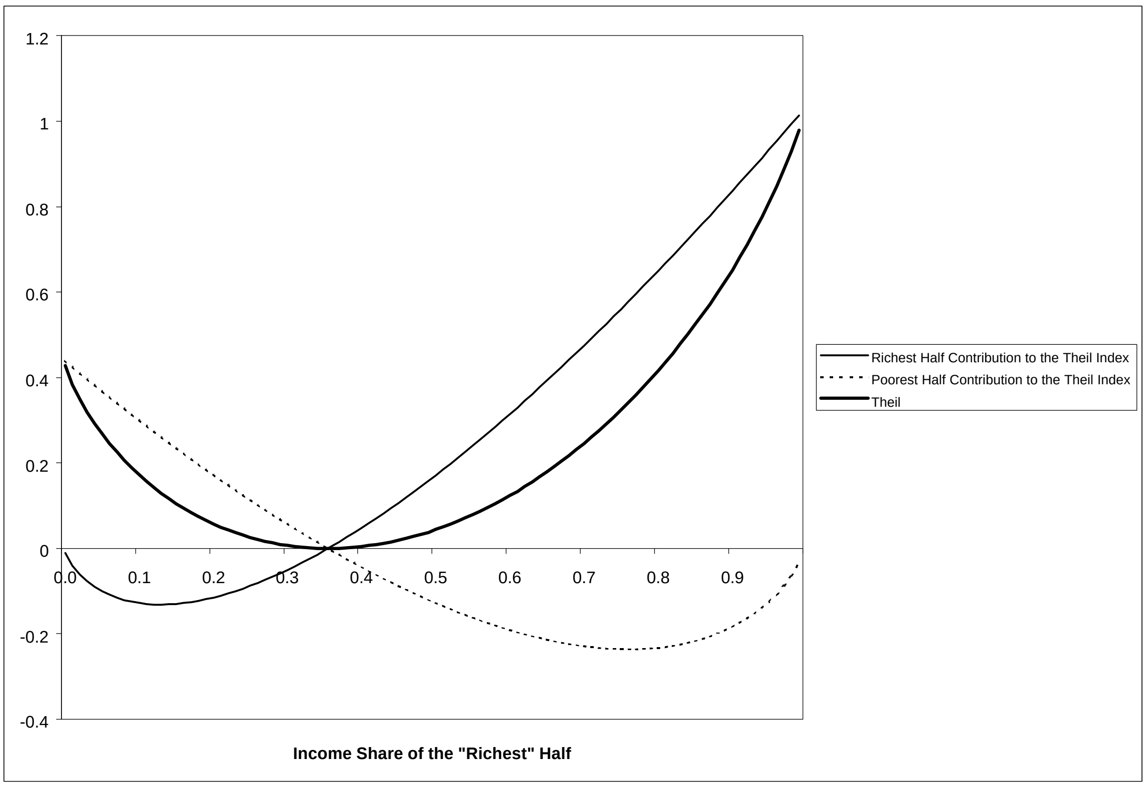

Figure 3- Deconstruction of the Theil Index

The thick line represents the Theil index, already shown in Figure 2. The thinner lines represent each group's contribution to the Theil index. The thin solid line shows the contribution of the "richest" group, and the thin dashed line of the "poorest" group's contribution. If we look to the right of the hypothetical point of perfect equality (the point at where all lines cross) the richest group's contribution is always positive, and the poorest group's contribution is always negative. This results, naturally, from the fact that to the right of the point of perfect equality, the rich group's income share is always higher that its population share, and vice-versa for the poorest group. To the left of the point of perfect equality, the situation is reversed.

Figure 3 illustrates several interesting points. For example, the positive contribution is always higher than the negative contribution, which makes the summation of the two contributions always positive[7]. Additionally, the positive contribution is almost linear. In fact, if we were to ignore the negative contribution to the Theil index, the measure of inequality that would result would not be much different from the other proposed measures of inequality suggested above, which behave linearly. The negative contribution has two important effects. First, it means that the Theil index approaches zero as the distribution becomes more equal faster, first, and then slower, than in a linear way. Second, the deviance from linear behavior is due to the shape of the negative contribution, which gives the Theil curve its concavity. The negative contribution is always due to the group that has less than its fair share of income. Thus, if we start from the right of the figure (where the income share of the "richest" half is one) the negative contribution is due to the poorest half's contribution. As we move to the right there is a sharp reduction of the Theil index as income is transferred from the rich to the poor. This reduction accelerates until the negative contribution reaches its minimum[8]. From then on, although the existence of a negative contribution still makes the pace at which the Theil index goes

[7] While this illustration does not prove that this result is general (and it is, indeed, general) it helps to understand how the interplay between the logarithmic transformation plus the income share weights work in conjunction to produce a measure of inequality that is always positive.

[8] It can be shown that the minimum of the poorest half's contribution is attained when the rich group's income share is equal to $1\text{-}n_{poor}/e$, and the minimum for the richest half contribution when the income share is given by n_{rich}/e. Thus, the minimum for the poorest half occurs when the income share of the richest half is .76 and the minimum for the richest half's contribution when the income share of the richest half is .13; the fact that these values do indeed correspond to the minimum of each contribution can be readily checked in the figure.

to zero initially more rapidly than in the linear case, this pace decelerates as compared with the situation before the minimum is reached. The existence of a minimum in the negative contribution, and the change in the sign of the first derivative, is, indeed, required to achieve a convergence to zero as the distribution moves towards equality.

The objective of this section was to provide an intuitive interpretation for the Theil index. We saw that the Theil index can be understood as a summary measure of inequality that gives a number that reflects the extent to which the structure in the distribution of income across groups differs from the distribution of population across those same groups. When the structures are the same (each group has the same share of income as its share of population) the Theil index attains its minimum (zero). If one of the groups has the same share of income and population, this group's contribution to the Theil index is zero. Groups that have higher shares of income than population shares contribute positively to the Theil index; those that have lower shares of income than population contribute negatively. Still, the positive contributions are always higher than the negative contributions, so that the Theil index is always positive overall. The negative contributions provide the non-linearity that make the Theil index sensitive to transfers of income from poor to rich, a sensitivity that increases the large the amount that is transferred and the wider the dispersion between rich and poor.

The *a priori* motivation to introduce the logarithm was somewhat arbitrary, although some reasons why it would be important to perform a logarithmic transformation on the ratio of the shares were given: the logarithm of one is zero, the logarithm is a monotonic transformation that staggers the ratio of the shares, the measure of inequality that results from applying the logarithm is always positive and continuous (with first derivative always defined). Still, other transformations could have been used yielding essential the same results. The value of the logarithm, though, will become apparent *a posteriori*, when we explore in the next section further properties of the Theil index, especially the fact that the Theil index allows for a perfect and complete decomposition of the inequality measure across groups.

2-APPLICATIONS: DECOMPOSING THE WORLD DISTRIBUTION OF INCOME

To further explore the analytical potential of the Theil index, we will continue to use the same data set as in the previous section. However, we will group countries according to the continent to which they belong, providing us with five groups, instead of the two we considered before. Figure 4 plots the population and income shares for each continent in 1970, providing a different perspective of world inequality than the one given by Figure 1.

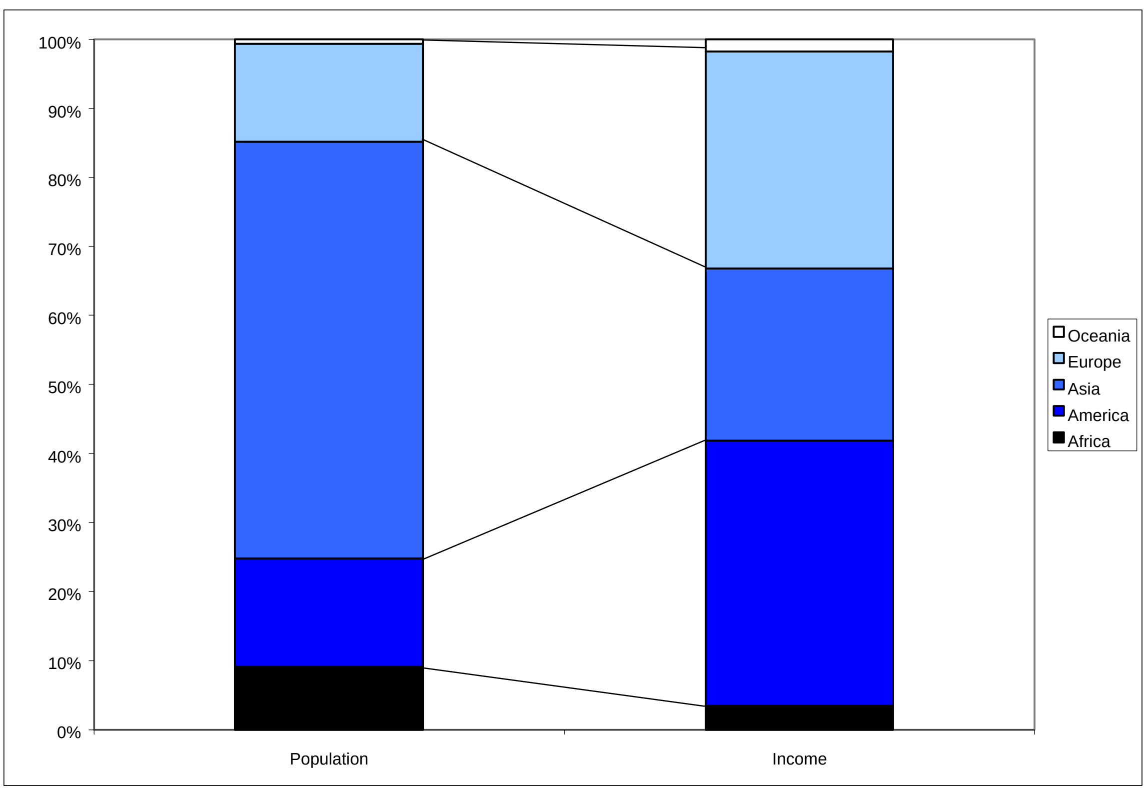

Figure 4- World Inequality: Population and Income Shares of Five Continents in 1970.

Table 2 provides the values needed to compute the Theil index. The first column represents the population share of each continent, and the second column the income shares. Africa, for example, has 9% of the population of the countries we are considering, but only 3% of the income. Europe, with a share of population less than 50% larger than Africa, has income share more than ten times as large as Africa's. In the Americas and in Australia income shares are more than two times as large as the corresponding population shares. Asia, like Africa, has a lower income share than its population share, although the ratio in Asia's case is more favorable than for Africa.

Table 2- Population and Income Shares for 5 Continents in 1970 and Theil Index Between The 5 Continents[9].

	Population Share	Income Share	Log of the Ratio of Shares	Contribution to the Theil Index
Africa	0.09	0.03	-0.98	-0.03
America	0.16	0.38	0.90	0.34
Asia	0.60	0.25	-0.88	-0.22
Europe	0.14	0.31	0.80	0.25
Oceania	0.01	0.02	1.09	0.02
Theil Index				**0.36**

From Figure 4 and the textual description based on the population and income shares we can characterize the inequality in the distribution of income across continents, but still fail to produce an inequality measure. To give us a measure of inequality, we can, again, compute the Theil index. We apply the formula defined in equation [2]; now we have five terms (instead of two), one for each continent. The third column in Table 2 provides the logarithm of the ratio of the shares. The continents for which income shares are lower than

[9] The values in the table have been rounded to ease the presentation; direct computations of the difference of the logs of shares with basis on the first and second column will lead to different results than those presented in the third and fourth column.

population shares yield a negative number, for the reasons explored in the previous section. The final column provides the ratios already weighed by the income shares, representing each continent's contribution to the Theil index. After applying the weights, the (positive) contributions of the groups with larger income shares overwhelm the negative contributions. While in the third column the numbers are (in absolute value) very close to each other, the weighing by the income shares works to produces a positive Theil index, which results from the summation of the five terms in the fourth column.

World inequality across continents is not as large as inequality between the two groups of rich and poor countries depicted in Figure 1. The Theil index measuring inequality across continents in 1970 is .36, while for inequality across the two groups in section 1 is .46. The reason is that in every continent there are rich and poor countries. Even though the distribution of income across continents is far from equal, considering a geographic, rather than average GDP, criterion yields a lesser level of inequality across groups.

Therefore, the number of groups we consider produces different measures of inequality. This inequality across groups is a component of the more fundamental level of overall inequality across countries (which are, in our data, the fundamental unit of analysis). Our objective now is to produce that comprehensive measure of inequality across countries, and see how that is associated with inequality between groups of countries.

Expressing equation [1] with the Theil index, we get:

$$T_{\text{World}} = T'_{\text{World}} + T_{\text{remaining}} \tag{3}$$

So far we were able to compute T'_{World} for two grouping structures, but $T_{\text{remaining}}$ remains to be addressed. One natural way to move forward is to apply exactly the same procedure at a lower level of aggregation than that of a continent. In other words, we can compute the Theil index that measures inequality between countries within each continent.

To illustrate the procedure, we will start with Oceania, for which we have data for four countries. First, Figure 5 shows the graph with the population and income shares for the four countries in Oceania in 1970.

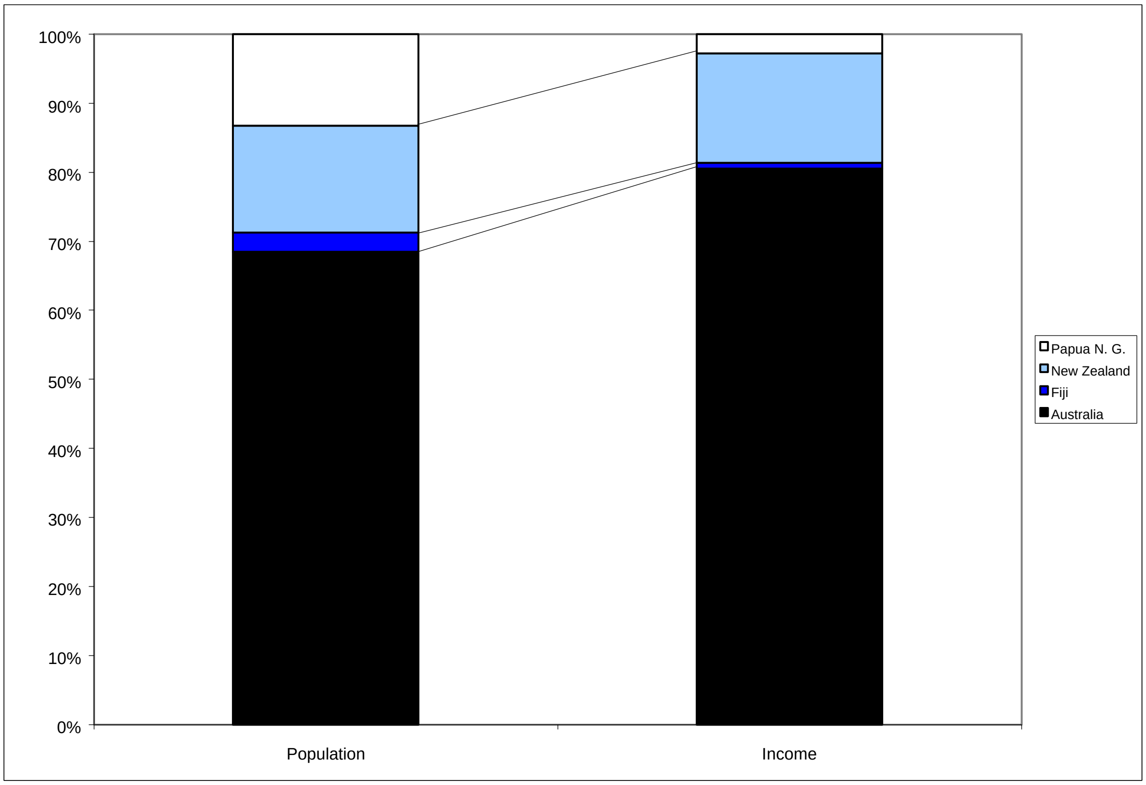

Figure 5- Inequality in Oceania in 1970: Population and Income Shares for Four Countries.

Similarly to what we have done across continents, Table 3 shows the values of the population and income shares for each country in Oceania. Note that our context now is exclusively that of one continent, so when we speak of shares, we are now referring to the population and income proportion that each country has of the total population and income *within* Oceania. To define clearly the context under which the shares are calculated is crucial to understand, and to compute correctly, the Theil index.

Table 3- Population and Income Shares for Four Countries in Oceania in 1970 and Theil Index Between these Countries

	Population Share	Income Share	Log of the Ratio of Shares	Contribution to the Theil Index
Australia	0.68	0.81	0.16	0.13
Fiji	0.03	0.01	-1.26	-0.01
New Zealand	0.15	0.16	0.03	0.00
Papua N. G.	0.13	0.03	-1.57	-0.04
Theil Index				**0.08**

An interesting feature of the Australian continent is that New Zealand has a fair share of income: note how the population and income shares are the same, around 15%[10]. Therefore, New Zealand's contribution to the Theil index between countries must be negligible. In contrast with New Zealand, which has a "fair share" of income, Table 3 shows that Papua N. Guinea has an extremely low share of Oceania's income. While this country has 13% of the continent's population – close to that of New Zealand's share – its income share is only 3% of the continent's income. Again, the Theil index gives a measure of the dispersion of income between the countries we considered in Oceania, summarizing the discrepancies in the shares of population and income for each country. The Theil index is obtained through the summation of the values in the fourth column of Table 3.

The procedure described in detail for Oceania can be replicated for the four remaining continents, so that we can obtain a measure of inequality between countries within each continent. Additionally, this procedure can be extended to other years. Table 4 shows the inequality across countries within each continent for 1970, 1980 and 1990. Continents ranked from the most unequal to the most equal in 1990. Asia comes across as the most unequal continent consistently since 1970. This is largely driven by the impact of Japan (with, in 1990, 4% of the continent's population, but 28% of income), and of China and India (with a combined population of 70%, but only 41% of the continent's income). We will explore below in more detail the dynamics of inequality in Asia. Next come the

[10] Of course this does not mean that the distribution of income *within* New Zealand is equal.

Americas, where the US, with 36% of the population and 68% of the income in 1990, drives that continent's Theil index. Africa, while poor, has a more equal distribution of income among its countries than do the richer Asia and the Americas. Europe's position as the most equal continent is not surprising, given the relative homogeneity of the European countries income distribution, reflected in similar shares of that continent's income and population shares for each country.

Table 4- Income Inequality Across Countries in Five Continents as Measured by the Between-country Theil Index

	1970	1980	1990
Asia	0.42	0.43	0.43
America	0.22	0.18	0.26
Africa	0.16	0.14	0.17
Oceania	0.08	0.10	0.12
Europe	0.09	0.08	0.10

We are not interested so much in the substantive conclusions one could draw from this type of analysis, but more on the exploration of the analytical potential of the Theil index. For a study that looks into the global distribution of income following a similar approach, and provides a more thorough substantive analysis, see Theil (1996).

Our next question is how to join the information we have in order to move towards a more comprehensive measure of world inequality. To recapitulate, we now know how much is the inequality between continents for 1970, as presented in Table 2. That procedure can be replicated for other years, yielding the following Theil measures for between continent inequality for 1970, 1980 and 1990: .36, .35 and .29. We also know what is the inequality between countries within each continent, as shown in Table 4. How can we combine the information we now have to produce a worldwide measure of between country inequality?

Let us start by writing once again the T' formula as it was applied to compute the between continent inequality (we include only the terms for Africa and Australia to simplify the presentation):

$$T' = w_{Africa}.[\log (w_{Africa} / n_{Africa})] + w_{Australia}.[\log (w_{Australia} / n_{Australia})] + \ldots$$

Looking at the Theil index formula above, it is clear that the structure of the formula is that of a weighted summation of direct inequality measures. The weights are the income shares for each continent (the proportion of world income that each continent has), and the direct measure of inequality is the logarithm of the ratio between the income and population shares. We also know that T' only gives us the inequality between continents. Following the same structure let us consider the following formula:

$$T'' = w_{Africa} \mathrm{x} T_{Africa} + w_{Americas} \mathrm{x} T_{Americas} + w_{Asia} \mathrm{x} T_{Asia} + w_{Australia} \mathrm{x} T_{Australia} + w_{Europe} \mathrm{x} T_{Europe}$$

This formula for T'' has the same structure as the expression for T': a weighted summation of a series of inequality measures, where for T'' the inequality measure is the Theil index for each continent that measures the inequality between countries *within* that continent. Thus, while T' provides a measure of inequality *between* continents, T'' gives a combined measure of inequality *within* all the continents. Computation is trivial from the formula above. The information in Table 2 (the second column gives each continent's income shares) and in Table 4 (the first column gives the within continent inequality) provides all the required data for 1970; Table 4 also provides the within continent inequality for other years, which after being multiplied by the income shares of the continents for that year provide the within continent contributions to world inequality. The results are presented in Table 5. Asia still is the major contributor to within continent world inequality, followed by America. However, given the large share of world income in

Europe, this continent's contribution is larger than that of Africa and Oceania, even though inequality within Europe is lower than in these other two continents (see Table 4).

Table 5- Within Continent Contributions to World Inequality[11]

	1970	1980	1990
Asia	0.11	0.12	0.14
America	0.09	0.07	0.09
Europe	0.03	0.02	0.03
Africa	0.01	0.01	0.01
Oceania	0.001	0.002	0.002
T´´	**0.23**	**0.22**	**0.27**

If we are interested in the world inequality between countries, then, for every year, $T'_{World}=T' + T''$ provides us with a measure of the dispersion in the distribution of income among all the countries in the world. The results are presented in Figure 6. The between continent contribution is represented by the area in each column shaded in black. Figure 6 indicates that world inequality (measured only as the between country inequality in the distribution of income) has been decreasing since 1970. We can also see that the decrease in world inequality has been driven by a decrease in the between continent component of world inequality. In 1970 the between continent contribution to world inequality was .36 (as we saw in Table 2) but by 1990 it was only .29 (the computations are not presented here, but the process is the same for the other years as presented in Table 2). In fact from 1970 to 1990, as Table 5, shows the within continent contribution increased. The graphic representation of Figure 6 helps to assess the relative contributions of each continent's within component contribution to world inequality. Clearly, Oceania and Africa do not make much difference; the African contribution has remained constant, and although the

[11] The values presented in this table were rounded; looking to rounded values can lead to misleading conclusions. For example, from 1970 to 1980 the contribution of Oceania increased only 10%, from about .0014 to 0.0015, but rounding makes it look like it doubled from .001 to .002. Below we will give more precise determination of the dynamics of inequality.

within country contribution to world inequality of Oceania increased from 1970 to 1990, the scale of this continent's contribution is so small that it did not have impact in the world distribution of income across countries.

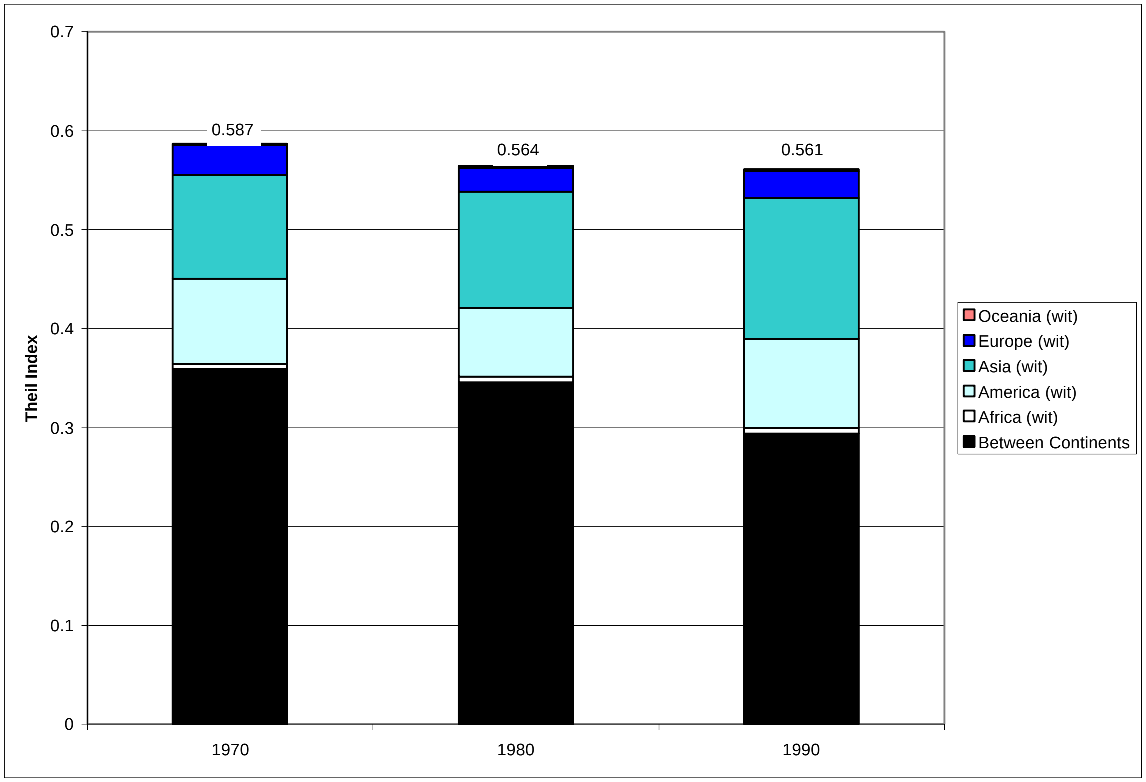

Figure 6- Decomposition of the Between-Country World Income Inequality

So is T'' all we need to add to T' to have the comprehensive measure of inequality, T_{World}, defined in equation [3]? No, because so far we have only measured inequality between countries, and not individuals. Extending the structure of the Theil index formula to T_{World} suggests the following expression as a comprehensive measure of world inequality:

[4] $$T_{\text{World}} = T' + T'' + \sum_{all\ countries} w_{country} \times T_{country}$$

The third term to the right of the equality sign [4] corresponds to the inequality between individuals. The formula $\sum_{all\ countries} w_{country} \times T_{country}$ extends our argument to the individual level: inequality between individuals is measured as a weighted average of the Theil index for each country. $T_{country}$ represents the Theil index among each country's individuals, and the weights are each country's income share of the world's total.

Since we have defined the Theil index only in terms of groups, we need to look into how the formula needs to change when it is applied to individuals. To do so, let us start by defining the following terms:

- y_i: income for individual i;
- $n_{country}$: country's population (number of individuals in the country);
- $y_{country}$: country's total income ($y_{country} = \sum_{i=1}^{n_{country}} y_i$)

With these definitions, the formula for a country's Theil index is:

$$[5] \qquad T_{country} = \sum_{i=1}^{n_{country}} \left[\frac{y_i}{y_{country}} \log \frac{\left(\frac{y_i}{y_{country}} \right)}{\left(\frac{1}{n_{country}} \right)} \right]$$

The income shares of each individual are just that individual's income divided by the country's total income. The "population share" is now just one (a single individual) divided by the country's population. It does not really make sense to speak about population shares when we are considering individuals as the unit of analysis. Still, formally the idea is the same as when we consider grouped individuals. We want to measure the extent to which the income distribution differs from the population distribution; when we consider individuals, the population distribution is simple: each person counts as one. Therefore, we have equality when the distribution of income is such that each person has the same amount, an amount that has to be equal to the country's income divided by the country's population. This is the only condition under which the Theil index for a country is zero.

Some manipulations of equation [5] will show the country's Theil index in a more familiar form. If the average country income is given by:

$$m_{country} = \frac{y_{country}}{n_{country}}$$

Then we can transform equation [5] into the more familiar expression for the Theil index:

$$[6] \qquad T_{country} = \frac{1}{n_{country}} \sum_{i=1}^{n_{country}} \left[\frac{y_i}{m_{country}} \left(\log \frac{y_i}{m_{country}} \right) \right]$$

To compute a comprehensive measure of the world inequality, we could apply the Theil index at the individual level to all the people in the world. The formula, following equation [5], is:

$$T_{world} = \sum_{i=1}^{n_{world}} \left[\frac{y_i}{y_{world}} \log \left(\frac{y_i}{y_{world}} \Big/ \frac{1}{n_{world}} \right) \right]$$

where n_{world} is the total world's population and y_{world} the overall worldwide income. This expression can be decomposed into three parts. First, T' measures inequality between continents:

$$[7] \qquad T' = \sum_{c=1}^{m_{continents}} \left[\frac{y_c^{continents}}{y_{world}} \log\left(\frac{y_c^{continents}}{y_{world}} \Big/ \frac{n_c^{continents}}{n_{world}} \right) \right]$$

where $m_{continents}$ is the number of continents, $y_c^{continents}$ is the total income in continent c, and $n_c^{continents}$ total population also in continent c. Second, T´´ measures the inequality within continents:

$$[8] \qquad T'' = \sum_{c=1}^{m_{continents}} \left[\frac{y_c^{continents}}{y_{world}} T_c^{continents} \right]$$

where $T_c^{continents}$ is the inequality between countries for continent c, which is given by:

$$T_c^{continents} = \sum_{p=1}^{m_c} \left[\frac{y_c^p}{y_c^{continents}} \log\left(\frac{y_c^p}{y_c^{continents}} \Big/ \frac{n_c^p}{n_c^{continents}} \right) \right]$$

where y_c^p and n_c^p are the income and population of each country p belonging to continent c; continent c includes m_c countries, and the continent's aggregated income and population are represented as in the previous expressions. Finally, the remaining component of inequality corresponds to the distribution of income across individuals within each country. This remaining component would be obtained by computing expression [6] for all countries (within country inequality) and summing all these within country Theils weighed by each countries income share. This clearly requires knowledge of the income distribution across individuals for each country, which we do not have in the current data set and, in general, is difficult (if not impossible) to gather for all the countries in the world.

Still, if we are interested in considering only the level of inequality in the distribution of income across countries, the decomposition properties of the Theil index can be used to explore analytical characteristics in the dynamics of income distribution across countries and continents.

We already saw how the decomposition of the Theil index into between continents and within continents component's helped us to understand that there is a convergence in the levels of income across continents, but a divergence in the inequality within continents. The divergence within continents is driven by the contribution of Asia to world inequality, which increased from .11 in 1970 to .14 in 1990. However, Table 4 showed that inequality within Asia had increased only from .42 in 1970 to .43 in 1990. Clearly, the contribution to world inequality of each continent is a function of two factors: a pure distribution effect (the level of inequality within countries in that continent, *DT*) and of a "continent's share of world income" effect (the way in which the weight of each continent's inequality enters into the world Theil index, *Dw*). Equation [8] shows that each continent's total contribution to changes in inequality is given by *DT* x *Dw.*

Table 6 shows the relative contributions of pure inequality changes and changes in each continent's shares of world income to world inequality. From 1970 to 1980 inequality decreased within every continent, except in Asia and Oceania, but the increase of inequality within Asia was of only 3%. However, given that, from 1970 to 1980, the Asian share of world income increased 1.09 times, the overall impact of the within Asian inequality in world inequality augmented 12%. The increase of the Asian contribution from 1980 to 1990 was equally driven by the augmentation of the share of world income in this continent. Note that inequality within the Asian continent did hardly change from 1980 to 1990, but that this continent's share of world income increased more than 20%. In fact, all other continent's shares of income decreased from 1980 to 1990. The within continent increase from 1980 to 1990 was driven, for all continents except Asia, by increases in the pure inequality effect. In other words, from 1980 to 1990 all continents with the exception of Asia became more unequal; Asia's inequality remained stable, but a large increase in the continent's share of world's income gave more prominence to Asia's high level of inequality.

Table 6- Decomposition of the Contributions of the Within Component Changes to World Inequality

	1970-1980	1980-1990
	Pure Inequality Change (ΔT)	
Africa	0.88	1.23
America	0.82	1.40
Asia	1.03	1.00
Europe	0.86	1.24
Oceania	1.18	1.26
	Continent's Share of World Income Change (Δw)	
Africa	1.12	0.91
America	0.99	0.92
Asia	1.09	1.21
Europe	0.94	0.92
Oceania	0.92	0.97
	Total Change (ΔTxΔw)	
Africa	0.99	1.12
America	0.81	1.29
Asia	1.12	1.21
Europe	0.80	1.14
Oceania	1.09	1.22

We will conclude this exploration of analytical applications of the decomposition properties of the Theil index with an analysis that combines groups with individual countries. The application will be to Asia. We have noted before that inequality within Asia is: 1) the highest; 2) relatively stable; 3) Asia's share of world income increased almost 32% from 1970 to 1990. Three countries dominate Asia. China and India have almost two thirds of the continent's population, and Japan has almost one third of the continent's income. However, in terms of dynamics, the period from 1970 to 1990 was characterized equally by the emergence of the Asian tigers, which we will define here as those regions that in 1990 achieved levels of income per capita higher than 5 000 USD (Hong Kong, Korea, Malaysia, Singapore and Taiwan). Therefore, it would be interesting to decompose the Asian inequality measure (which has remarkably stable between 1970 and 1990) to see what the effect of the dominant countries and of the Asian tigers was along these two decades.

Table 7 shows the rich dynamics that hide behind a relatively stable inequality measure. Asian inequality is decomposed for each year between the contributions of three countries considered separately (China, India and Japan) and the contributions of two groups (the Asian tigers and the remaining countries). For each group there is the between component and the within group component. The income and population shares of each country and group are also represented. The first interesting fact is that the joint contribution to Asian inequality of China and India has remained stable, at -.22 in 1970 and -.23 in 1980 and 1990. China lost population share throughout; it gained income share in 1980, but lost again 1% of income share in 1990. India gained population share from 1980 to 1990, and also 1% in income share; however, India's share of income was at the peak in 1970 (18%), with a lower share of population than in 1990.

Table 7- Decomposing the Dynamics of Inequality in Asia

	1970			1980			1990		
	Population Share	Income Share	Contribution to Theil	Population Share	Income Share	Contribution to Theil	Population Share	Income Share	Contribution to Theil
China	0.43	0.23	-0.14	0.42	0.25	-0.13	0.40	0.24	-0.13
India	0.29	0.18	-0.08	0.29	0.16	-0.10	0.30	0.17	-0.10
Japan	0.05	0.31	0.55	0.05	0.31	0.55	0.04	0.28	0.52
Tigers (btw)	*0.03*	*0.05*	*0.03*	*0.03*	*0.08*	*0.07*	*0.03*	*0.10*	*0.12*
Tigers (wit)			0.002			0.005			0.005
Other (btw)	0.19	0.22	0.02	0.21	0.21	0.003	0.22	0.21	-0.01
Other (wit)			0.05			0.03			0.02
Asian Theil			**0.42**			**0.43**			**0.43**

Japan's situation was the same in 1970 and in 1980. This country contributed each of these years with .55 to Asian inequality. However, in 1990 its contribution dropped to .52, as the country's income share fell from 31% to 28%. For the other Asian countries, the between group contribution to Asian inequality is rather small, since the population and income shares of this group are very close. In 1970 this group's income share was slightly higher than the groups population share, which meant that the between group contribution to inequality was of .02. In 1990 the income share is 1% below the population share, and so the between group's contribution is, again a rather small, -.01. Within this group of other countries inequality has been dropping, so the within group contribution to inequality decreased from .05 in 1970 to .02 in 1990.

Perhaps more interesting is the effect of the Asian Tigers. The population share of this group remained around 3%, but the income share increased from 5% in 1970 to 10% in 1990. Consequently, the between group contribution of the Tigers to Asian inequality rose from .03 in 1970 (not very different from the .02 of the other Asian countries) to .12 in 1990. The Tigers are relatively equal among each other, with the within Tigers inequality being around .005 in the more recent years.

The combination of groups with individual countries makes the analysis somewhat more confusing and less intuitive, largely because of the negative contributions to the Theil index. In fact, if we consider only groups, then the decomposition of the Theil index into an overall between groups component and the several within groups components produces only positive values, which add to the overall Theil index, as we saw in Figure 6. Still, if we have in mind the interpretation of the Theil index explored in section 1, a similar chart to that of Figure 6 could be produced using the contributions of countries and groups (for these, the within and the between components, where the between components can also be negative) presented in Table 7. We attempted to do precisely that in Figure 7. Again, the interpretation must be cautious, because the overall inequality level results from the summation of all the components, with the negative components (which appear below the horizontal axis) to be subtracted to the positive components. While this type of chart may not be adequate to provide an analysis of the evolution of the Asian level of inequality, it certainly calls attention to the dynamics of the evolution of Asian inequality described above with the help of Table 7. The interplay of the several countries and groups chosen can be easily discerned. It is possible to see, for example, that the negative contributions of China and India are almost constant. Also, these are the only negative contribution up to 1980; in 1990 the between component of the other countries comes also as a negative contribution. More salient is the overwhelming weight of Japan in driving inequality in Asia, with the gain in weight of the between component of the Asian Tigers also clearly visible. Finally, the reduction in the contribution associated with the within component of the “other countries” group is also clearly exposed.

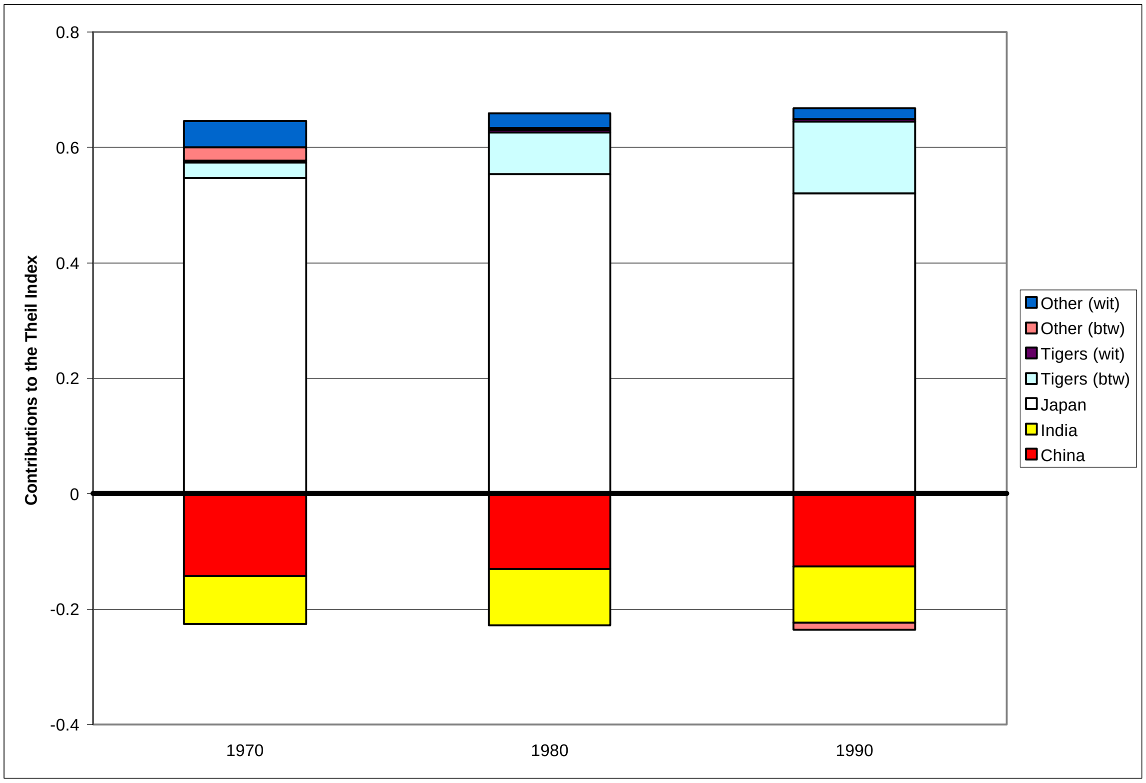

Figure 7- Decomposing the Asian Theil Index: Contributions of China, India, Japan, the Asian Tigers and of the Other Asian Countries

So far we explored the potential for analytical work taking advantage of the perfect decomposition of inequality between and within groups of the Theil index. The decomposition properties of the Theil index derive from the characteristics of the logarithm, which has an ability to transform multiplications into summations. Can this decomposition of inequality be achieved with other measures of inequality? And what is the relationship between the Theil index and other measures of inequality, particularly the Gini coefficient?

The answer to the first question is a qualified no. The qualification derives from the fact that the Theil index is one the measures of a family of "entropy based measures of inequality". Only inequality measures that are members of the "entropy based" family allow for a perfect decomposition of inequality into a between group and a within group component (Shorrocks, 1980). What are the other elements (measures of inequality) of the "entropy based" family? We saw that the Theil index measuring inequality between *m* groups, where group's *i* income share is w_i and population share is n_i, can be written as:

$$T' = \sum_{i=1}^{m} w_i \log \frac{w_i}{n_i}$$

A similar measure would be one in which the role of the income and population shares are switched:

$$L' = \sum_{i=1}^{m} n_i \log \frac{n_i}{w_i}$$

This measure of inequality is called Theil's second measure, being an example of another member of the "entropy based" family. In general, the between group component of entropy based measures of inequality takes the form:

$$E_a' = \frac{1}{a(1-a)} \sum_{i=1}^{m} \left[1 - \left(\frac{w_i}{n_i} \right)^a \right]$$

an expression which is not defined for α=1 and α=0, but that can be shown to be transformed into T' , when α approaches 1, and L', when α approaches 0. Even though it may not be obvious the way in which E_a' turns into T' and into L', it is clear that the intuition behind E_a' is the same as that of the measures we have been considering: once again, we are trying to measure the discrepancy between the shares of income and the population shares across groups.

In fact, any (objective) measure of inequality needs to convey this discrepancy between the income and population shares across groups, which means that, formally at least, there is a large degree of similarity between inequality measures. The formal similarity between the Theil index (or, more generally, the entropy based measures of inequality) and the Gini coefficient can be understood if we write the Gini – as suggested by Theil (1967) – in the following form:

$$G' = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} n_i n_j \left| \frac{w_i}{n_i} - \frac{w_j}{n_j} \right|$$

This expression shows that the Gini across groups is the result of a comparison, across all groups, of the ratio between income and population shares[12].

The shortcoming of the Gini, and the unique advantage of the entropy based measures of inequality, is that the *within* group component cannot be neatly added to the between group component. The entropy based measures of inequality are the only for which total inequality can be expressed as:

$$E_a = E_a' + \sum_{i=1}^{m} n_i \left(\frac{w_i}{n_i} \right)^a E_a^i$$

[12] We need the ½before the summation because otherwise we would be double counting. Theil (1967) provides an easy way to get a better feeling of how the Gini in this form measures inequality. Assume that all income is in the first group, so that w_1=1 and w_i=0 for i=2,…,m. Then G´=1-n_1, which approaches 1 (Gini's maximum value) as the population share of the group that has all the income decreases. The Gini is 1 when all the income is with a single individual.

The summation represents the contribution to total inequality of the all the within group's inequality. It is easy to see that when α=1 (when we have the Theil index) the weight of each group's within inequality contribution is that group's income share, and when α=0 the weights are the population shares.

The next section further extends the illustration of the analytical potential of the Theil index, describing a software application that can ease the computation of the Theil index at different levels of aggregation.

3- EXTENSIONS: EXPLORING THE REGIONAL PATTERNS OF INEQUALITY IN THE US

The Theil index, as we saw, allows for a perfect and complete decomposition of the total level of inequality into the inequality within the sub-groups of the population, the within-group contributions, and the between-groups contribution.

Figure 8 shows a partition of the individuals of a population, $Ind_1,\ldots Ind_n$, into groups, $Group_1,\ldots, Group_n$, which are in turn aggregated into broader groups, $Group_a$ through $Group_z$. This population is, therefore, divided into two levels, which is enough for the purpose of showing the fractal behavior of the Theil Index although the formulation we derive may be applied to any number of levels.

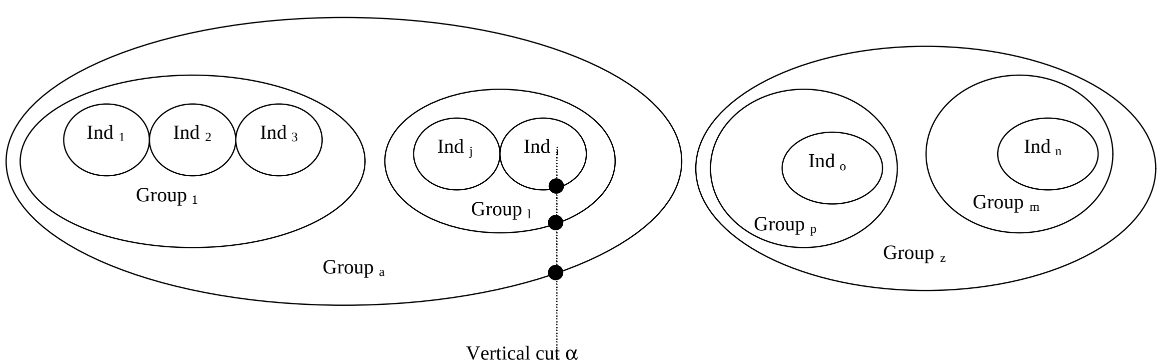

Figure 8 – A population of n individuals is hierarchically divided into groups.

The vertical cut α, with origin in individual i, intercepts the boundary of individual i (its source point), but also the boundaries of group l and group *a* (recall that *a* is at higher level than l and contains it). We now proceed to represent the same information but using a tree, as shown in Figure 9.

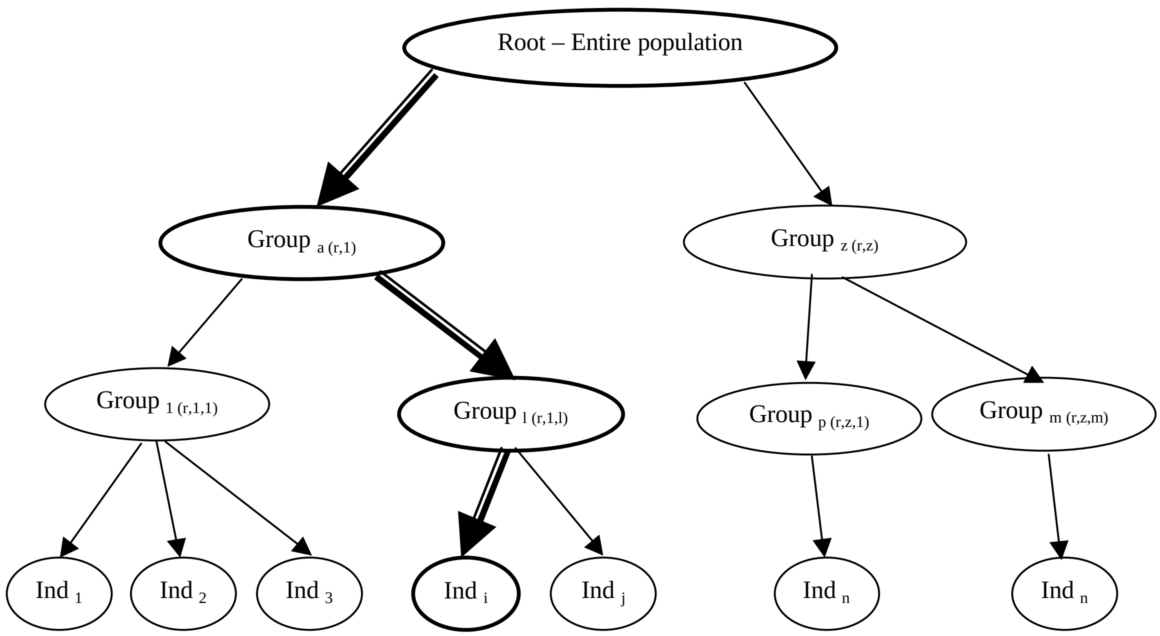

Figure 9– The same population hierarchically organized represented by a tree.

Let the root of the tree represent the entire population and the leaves the individuals. To be coherent with the previous representation individual *i* should belong to a branch with nodes corresponding to groups l and *a*, the groups to which it is linked given the actual decomposition of the population, in this order from the root to the leaf. Applying the same reasoning to the entire population we achieve the tree depicted in Figure 9. Note that individual *i* indeed belongs to a branch originated in the entire population and going through groups *a* and l before getting to the level of the individual.

The formulation of the Theil Index to a population using the conceptual framework of a tree allows for a renaming of the groups according to their position in the hierarchy. For instance, group *a* may now be called group *r*,1 since it is the first sub-group of the root node. Group *p* may be called group (*r*,*z*,1) since it is the first subgroup of group *z*, which in turn is originated in node *r*.

Therefore, the Theil Index applied to the entire population can be written as:

$$T_r = \sum_{i=1}^{nodes\_root} \frac{w_i}{Y}.\log\left(\frac{w_i}{Y} \bigg/ \frac{n_i}{n}\right) + \sum_{i=1}^{nodes\_root} \frac{w_i}{Y}.T_{(r,i)}$$

where nodes_root represents the number of children nodes of the root node and i^{th} group in the population has n_i people and an aggregated income of w_i. This equation means that the Theil Index applied to the root node, T_r, is evaluated by summing two components. First, we consider the inequality between the children nodes of the root node. We will call this component the between-groups component. Second, we add the inequality within each of these nodes, which is obtained through a recursive evaluation of the Theil Index at lower levels of aggregation. We shall call this the inequality within-groups. This formulation must be applied to evaluate the Theil Index at lower levels of the tree until a leaf is reached, in which case the within-groups inequality is zero. Using a more functional notation:

```
TheilIndex (Tree)
    IF Tree is just a leaf THEN TheilIndex = 0
    ELSE
    FOR EACH Child of the Root Node
        TheilIndex = Sum [ wi/W.log(wi/W.N/ni) + wi/W.TheilIndex(Tree from child i) ]
END
```

An application to evaluate the Theil Index has been devised, in the context of the University of Texas Inequality Project and will be made available at http://utip.gov.utexas.edu. This application runs as an EXCEL macro and computes the Theil Index over a population of individuals hierarchically organized using a tree representation. Given the structure of the population and information on the income levels, the application evaluates the overall inequality and computes the between and within contributions of each group of the population.

The remainder of this section is devoted to a presentation of results obtained using this software application. The illustration will be performed with data from the Bureau of Labor Statistics (collected by the Census), and will allow a study of the evolution of inequality in the US from 1969 to 1996. We have used population and household income levels to compute income inequality in the US territory for the period considered, with data at the county level.

To explore the potential of the Theil Index we have structured the population into three hierarchical levels. First, we have considered nine large regions typically used by the Census surveys, which are shown in Figure 10.

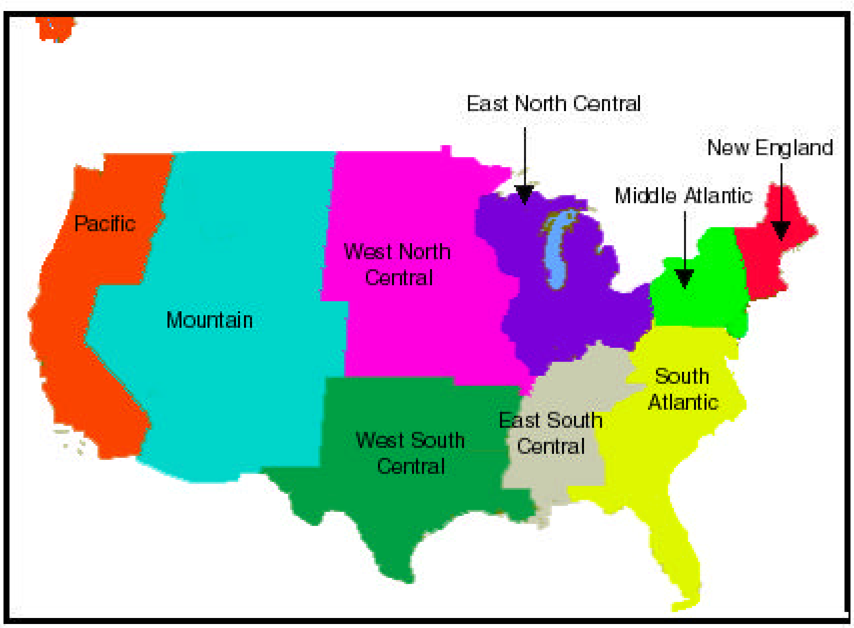

Figure 10- Map of the nine Census regions in the US.

To explore the analytical potential of the Theil Index we have further divided these regions into states, as represented in Figure 11. We have considered 50 states in the US, thus including the Alaska and the Hawaii. Beyond this partition of the US population we have also considered 3084 counties. Therefore, our unit of analysis is the county, for which we have data for both the population and the household income since 1969 up to 1996.

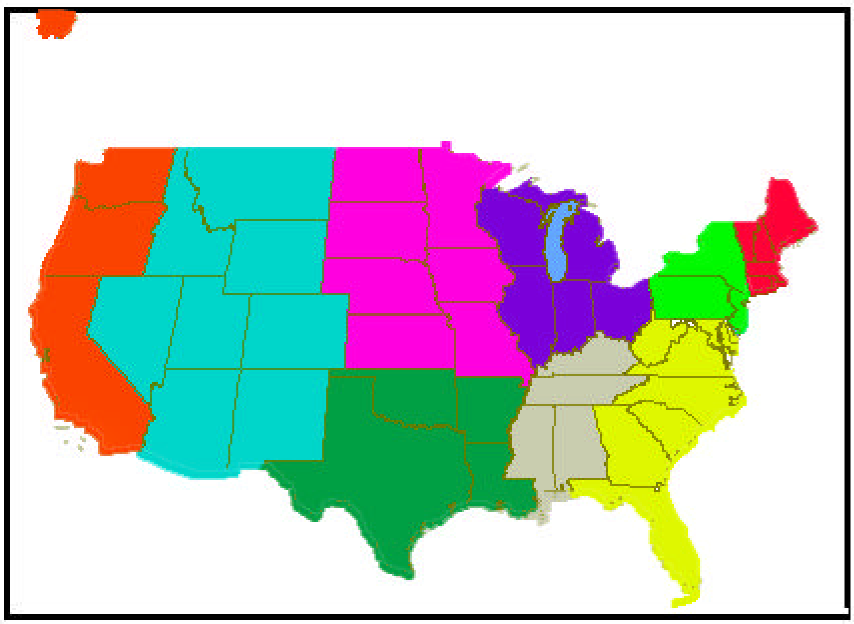

Figure 11- Map of the US showing the nine Census regions and the 50 states considered in the analysis.

This hierarchy of the US population is depicted in Figure 12. We consider four levels of nodes. There is the root node, which contains only the node representing the US from which we will extract the overall level of inequality in the US. There is a second layer of nodes that comprises the children nodes of the US node, which represent the Census regions. These have 50 children on aggregate, which represent the 50 states considered in our analysis. Finally we have the leaf nodes that represent the counties in the US and are children of the state nodes.

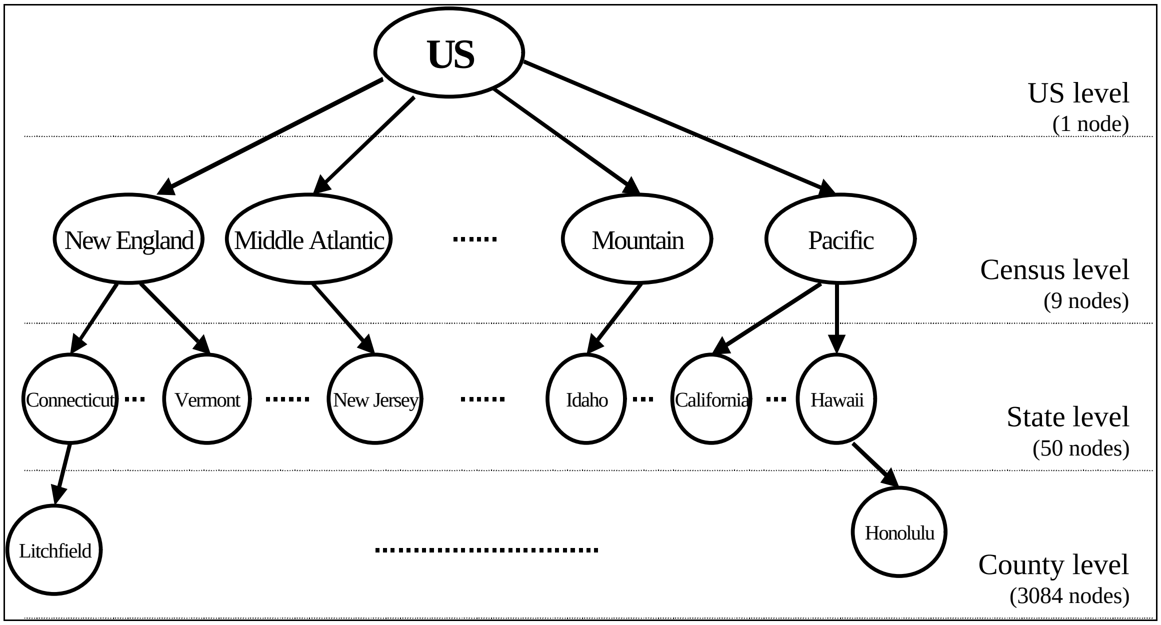

Figure 12- The US population hierarchically represented by a tree.

The first step in order to use the automated software application for the evaluation of the Theil Index is to express this tree structure in a spreadsheet format. The form shown in Figure 13 shows the required formatting. We call this the input form since it provides the format to supply the application with the population structure and the income levels. The four large boxes on the top are four buttons that we will explain later. Below there is a table that comprises four main areas. First, in column 2, we list the names of the nodes in the tree. So, we start by introducing the US, then the East North Central region, the East South Central region and the remaining 7 Census regions. After these, we move to the next level in the tree and we start listing the states from the Alabama to the Wyoming. Similarly, after the states we move to the leaf level of the tree and we start listing the counties from Autauga (the first county in Alabama) to Weston (the last county in Wyoming).

Next, in columns 3 to 7 we reserve space for the accumulated population for each node of the tree for each year. For simplicity, we have just shown the first and the last two years of our analysis, 1969, 1970, 1995 and 1996. The next set of columns, from column 8 to 12, is the space for the accumulated income, as before for the population. As expected, columns 3 to 12 we have just filled in the row corresponding to leaf nodes, since we only know the population and income levels for the simpler units of analysis, the counties in our case. Finally, column 14 codifies the structure of the tree. For each node we indicate the row of its parent, keeping in mind that parents must always appear in the list of nodes before their suns (mathematically, the number in (row *i*, column 14) cannot be larger the *i*).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2		Unfold		Eval Pop&Inc				Eval Pop&Inc %				Eval Theil Index		
3														
4														
5			Population					Income						
6		Node	1969	1970		1995	1996	1969	1970		1995	1996		Parent
7														
8		US												
9		East North Central												8
10		East South Central												8
11		Middle Atalntic												8
12		Mountain												8
13		New England												8
14		Pacific												8
15		South Atlantic												8
16		West North Central												8
17		West South Central												8
18		Alabama												10
19		Arizona												12
20		Arkansas												17
21														14
22		Wyoming												12
23		Autauga	25166	24606		39154	40277	66834	73240		697164	748246		18
24		Baldwin	56951	59474		119966	124049	156304	171023		2437328	2641005		18
25		Barbour	23818	22655		26626	26839	50192	55999		434397	452678		18
26														
27		Washakie	7486	7557		8581	8548	24413	26120		159592	166320		67
28		Weston	6235	6266		6570	6583	22942	23195		135099	136190		67
29														

Figure 13- Input form for data on population and income for inequality evaluation.

In order to run the Theil Index application we must start by pressing the button “Unfold”, which automatically creates the output book for the various results we will compute. Figure 14 shows the first sheet of this book, which is the result of clicking the button “Eval Pop&Inc” in the previous form. It shows a table with the same structure as the input table but it is now fully filled in. Indeed, this option simply unfolds the calculations for the population and income levels for every node in the tree according to the structure defined in the 14th column.

Unfold

Eval Pop&Inc

Eval Pop&Inc %

Eval Theil Index

	Population				Income				
Node	1969	1970	1995	1996	1969	1970	1995	1996	**Parent**
US	201802000	203494394	262158993	264574445	771577699	829305152	6123309518	6465150819	
East North Central	32836000	33153862	35589394	35747042	132559002	139189180	836721442	876389801	8
East South Central	15175000	15382176	18678358	18940289	46872382	50851112	383278325	402856875	8
Middle Atlantic	13476000	13576659	14979299	14911553	50499253	54186149	348428153	362466965	8
Mountain	16670000	16888362	20581702	21189296	63324776	69428308	509649374	535070793	8
New England	16662000	16837726	19788724	19894400	67077806	71502099	522473202	545728204	8
Pacific	25859000	26383851	41318101	41798432	113779040	122098797	993488946	1051638572	8
South Atlantic	24227000	24657944	38577823	39126298	87830052	96335432	891694286	945635406	8
West North Central	37161000	37525296	41722909	41947427	147064947	157270981	1009941397	1066422734	8
West South Central	18936000	19248378	30845727	31161680	61391111	66043524	631738385	670647370	8
Alabama	3440000	3448846	4251963	4287178	9431163	10225435	82057865	85881582	10
Arizona	1812000	1838877	2480919	2506293	5020649	5494172	49011111	47527037	12
Arkansas	1737000	1794312	4308188	4434340	6024161	6847940	87527224	94607350	17
Wyoming	329000	333785	478532	480011	1170525	1284818	10028823	10362181	12
Autauga	25166	24686	39154	40277	66334	73240	697164	743246	18
Baldwin	58951	59474	119966	124049	158304	171023	2437328	2641005	18
Barbour	23818	22655	26626	26839	50192	55309	434387	452679	18
Bibb	14994	13788	17962	18239	29273	31929	282888	298032	18
Washakie	7466	7557	8581	8949	24413	26120	153532	166320	67
Weston	6235	6268	6570	6583	22942	23196	135899	136190	67

Figure 14- "Data" sheet from the output book of the Theil Index application.

The same information will also be useful in the form of percentages. For this reason, the next outputs, achieved by pressing the button "Eval Pop&Inc %" are the shares of population and income for every node in the tree, as shown in Figure 15. The population share of a particular node is the ratio between the population it has and the aggregated population of its parent. The same applies to the income levels. From Figure 15 we may see that the East North Central region had 16.3% of the US population is 1969 and 17.2% of the income for the same year. By definition, the share of population and income for the entire US is 1, as shown in row 8 of this table for every year considered.

Unfold | Eval Pop&Inc | Eval Pop&Inc % | Eval Theil Index

	Population Shares				Income Shares				
Node	1969	1970	1995	1996	1969	1970	1995	1996	Parent
US	1	1	1	1	1	1	1	1	
East North Central	0.163362	0.162923	0.135679	0.135111	0.171803	0.167838	0.136645	0.135555	8
East South Central	0.075497	0.075197	0.071248	0.07121	0.060749	0.061318	0.062592	0.062328	8
Middle Atalntic	0.067044	0.066718	0.056754	0.056361	0.065436	0.065339	0.056411	0.056065	8
Mountain	0.082934	0.082992	0.080034	0.080088	0.082849	0.083719	0.083068	0.083381	8
New England	0.082895	0.082743	0.075484	0.075194	0.086936	0.086702	0.085325	0.085029	8
Pacific	0.12865	0.129261	0.157599	0.157825	0.147463	0.14723	0.162234	0.162663	8
South Atlantic	0.120531	0.121173	0.147154	0.147884	0.113832	0.116164	0.145623	0.146267	8
West North Central	0.184879	0.184405	0.15915	0.158547	0.19059	0.189641	0.164934	0.164949	8
West South Central	0.094208	0.094589	0.116897	0.11778	0.080343	0.082049	0.103168	0.103764	8
Alabama	0.226689	0.225448	0.228177	0.227554	0.201209	0.201086	0.214099	0.213378	10
Arizona	0.114757	0.114284	0.118237	0.118281	0.07854	0.079134	0.088491	0.088183	12
Arkansas	0.09173	0.09325	0.14058	0.142301	0.097178	0.100635	0.138552	0.141027	17
Wyoming	0.019736	0.019765	0.022807	0.022654	0.018311	0.018503	0.019713	0.019222	12
Autauga	0.007316	0.007132	0.009187	0.009395	0.007087	0.007163	0.008496	0.008702	18
Baldwin	0.016556	0.01724	0.028148	0.028935	0.016573	0.016725	0.029703	0.030716	18
Barbour	0.006924	0.006567	0.006247	0.00626	0.005322	0.005476	0.005294	0.005265	18
Washakie	0.022754	0.02264	0.017932	0.017808	0.020856	0.020333	0.015916	0.016051	67
Weston	0.018951	0.018772	0.013729	0.013714	0.0196	0.018056	0.013474	0.013143	67

Figure 15- “Shares” sheet from the output book of the Theil Index application.

Finally, we can produce all the information needed to quickly compute the contribution of each node for the overall inequality, which can be done by pressing the button “Eval Theil Index”. As seen before, each node has two separate contributions. On the one hand, each node contributes to overall inequality because it differs from the other groups and, therefore, it contributes to the inequality among its sibling nodes. On the other hand, each node encompasses within itself an amount of inequality that comes from the inequality among its children nodes. These are shown in Figure 16, with the former represented in columns 3 through 7 and the latter in columns 8 to 12. The table shown in this figure has exactly the same structure as the ones shown before, but it provides the between-groups and within-groups contributions for each node in the tree.

As expected, the between group’s contributions for the US node are zero, because there nothing outside the US in our example and therefore the US does not differ from anything else. Conversely, the within group’s contributions for the counties are all zero as well. We should expect this because the county is our unit of analysis. It is indivisible. Hence, there is no inequality within it. If there was, it would have to be the inequality between its children nodes and the county would not be our unit of analysis any longer since it would have be a group of tinier subgroups.

Unfold | Eval Pop&Inc | Eval Pop&Inc % | Eval Theil Index

	Between Contributions					Vithin Countributions						
Node	1969	1970		1995	1996	1969	1970		1995	1996		Parent
US	0	0		0	0	0.028741	0.026945		0.034563	0.035347		
East North Central	0.008655	0.004988		0.00097	0.000444	0.015394	0.014016		0.023904	0.024431		8
East South Central	-0.013203	-0.012511		-0.008108	-0.008304	0.031228	0.026917		0.027162	0.027197		8
Middle Atalntic	-0.001589	-0.001364		-0.000343	-0.000295	0.015382	0.013381		0.025963	0.02684		8
Mountain	-8.5E-05	0.00073		0.00309	0.00336	0.025212	0.024416		0.03855	0.037646		8
New England	0.004138	0.004052		0.010457	0.010452	0.021834	0.021002		0.034357	0.03504		8
Pacific	0.020125	0.019164		0.004702	0.004911	0.009307	0.009233		0.021188	0.022864		8
South Atlantic	-0.00651	-0.004904		-0.001523	-0.001609	0.02408	0.024367		0.027035	0.027485		8
West North Central	0.005798	0.00531		0.005888	0.00653	0.04544	0.04445		0.058166	0.059052		8
West South Central	-0.01279	-0.01167		-0.01289	-0.013148	0.023553	0.021785		0.027007	0.02789		8
Alabama	-0.02399	-0.022996		-0.013634	-0.013725	0.01819	0.013896		0.012929	0.013337		10
Arizona	-0.029783	-0.029085		-0.025644	-0.025895	0.016029	0.014593		0.015023	0.014143		12
Arkansas	0.005606	0.00767		-0.002014	-0.001269	0.009085	0.009216		0.013766	0.014783		17
Wyoming	-0.001372	-0.001221		-0.002874	-0.003157	0.011552	0.010801		0.017232	0.016682		12
Autauga	-0.000226	3.01E-05		-0.000664	-0.000666							18
Baldwin	1.76E-05	-0.000507		0.001597	0.001835							18
Barbour	-0.0014	-0.000994		-0.000877	-0.000912							18
Bibb	-0.001054	-0.000772		-0.0007	-0.00071							18
Washakie	-0.001816	-0.002185		-0.001898	-0.001668							67
Weston	0.000659	-0.000702		-0.000253	-0.000559							67

Figure 16- "Theil" sheet from the output book of the Theil Index application.

After these four steps we have all the information we need to express the Theil Index, or, in other words, the overall inequality, as a function of the within contributions and between contributions at any chosen level in the tree. For sake of simplicity we will focus the rest of our analysis at the Census region level, in which we have nine nodes. Clearly, overall inequality is equal to the summation of the inequality within each region plus the inequality between regions. The inequality within each of these regions is simply taken from the "Theil" sheet of the output book of our application. It is shown on the right hand side of Figure 17 in rows 3 to 11 for each of the years considered in the analysis. The between regions row is computed by summing the between contributions of all the nine regions and appears in the second row of the same table. The total within regions contribution is evaluated adding all the within contributions of the nine regions and is shown in row 13. Finally, the overall inequality is achieved by adding these two terms and is indicated in row 15. The procedure to develop these computations is also shown the same figure.

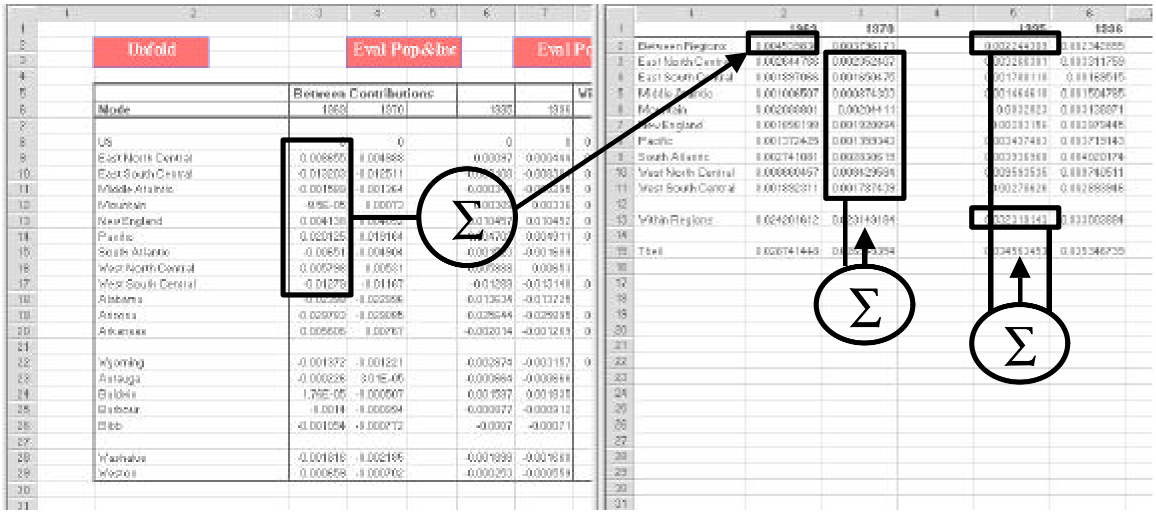

Figure 17- Inequality between the Census regions and within each region evaluated by the Theil Index application.

We now proceed to analyze the results obtained from evaluating the Theil Index over the US population from 1969 to 1996, which gives us a measure of the between county inequality experienced in the country in that period. Figure 18 shows the evolution of the Theil Index through time. We observe that in 1969 the Theil Index was about 0.028 and that during the next 7 years it decreased to its minimum level of 0.021 in 1976. Since then, inequality has always been rising with the exception of the period 1988-1994, for which it has been stable around 0.033, thus above the levels of 1969. Moreover, the inequality in 1986 was again equal to that of 1969. In fact, the period in which the income inequality across counties in the US has its most significant growth is the second half of the 1980’s.

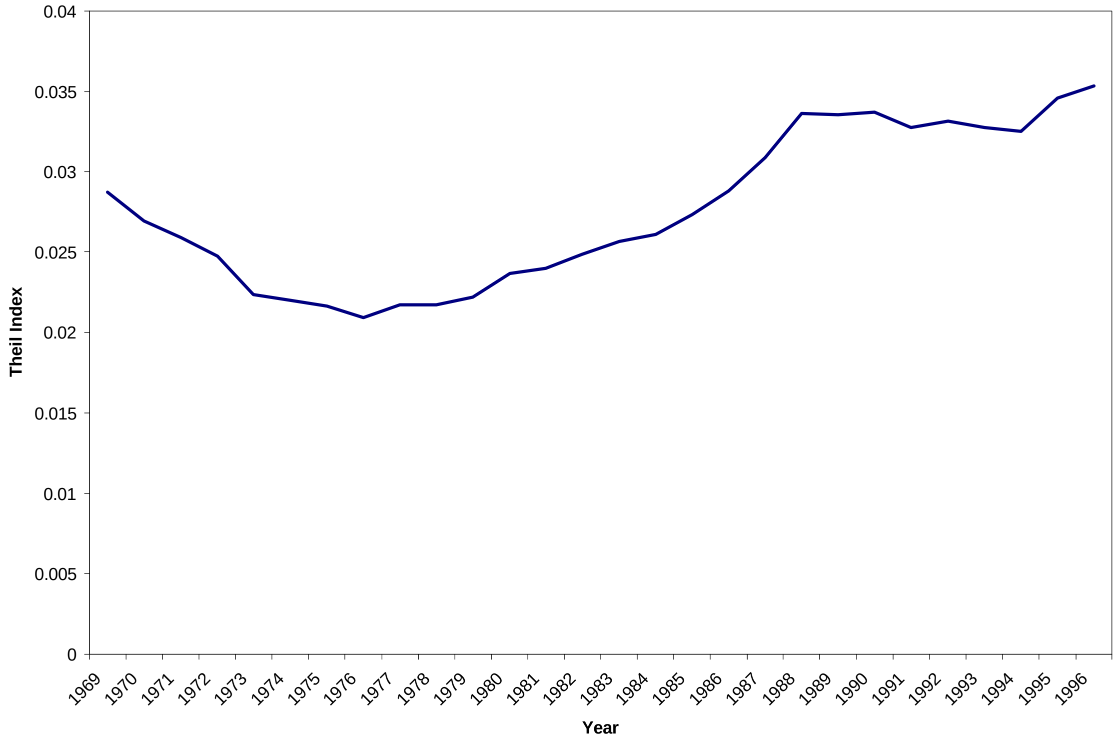

Figure 18- Overall inequality in the US from 1969 to 1996 measured by the Theil Index.

The overall level of inequality can be decomposed into two main components, the inequality between the Census Regions and the inequality within the regions. These may be simply added up to obtain the overall level of inequality given the property of neat decomposition of the Theil Index.

The evolution of these components is shown in Figure 19. We notice that the between regions component has been slowly decreasing for the entire period of analysis. The exception is again the period comprised between 1985 and 1991 during which the increase in the between regions component has contributed for the steepest inequality growth in the US, as described before. However, the dynamics of the overall inequality in the country is largely defined by the inequality within each of the Census regions and not by the differences among them.

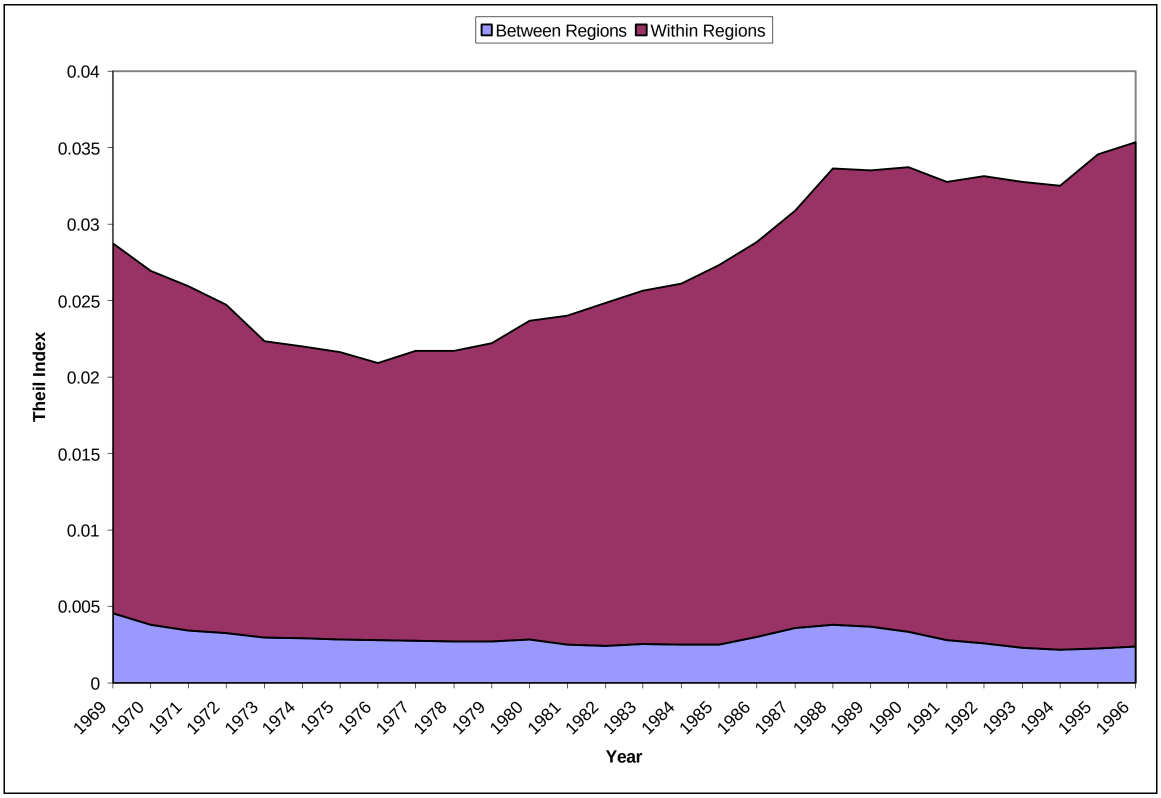

Figure 19- Breakdown of the overall inequality level in the US into the between Census regions and within Census regions components for the period 1969-1996.

These facts are very clear from Figure 20, which represents the evolution of the between regions inequality component and the within regions inequality component separately. The dominance of the within regions component is perceptible from comparing the darker line in this chart with that of Figure 18. The brighter line shows the evolution of the between regions inequality component, which has been decreasing from 1969 to 1996 except for the second half of the 80's when the darker line also exhibits its largest growth.

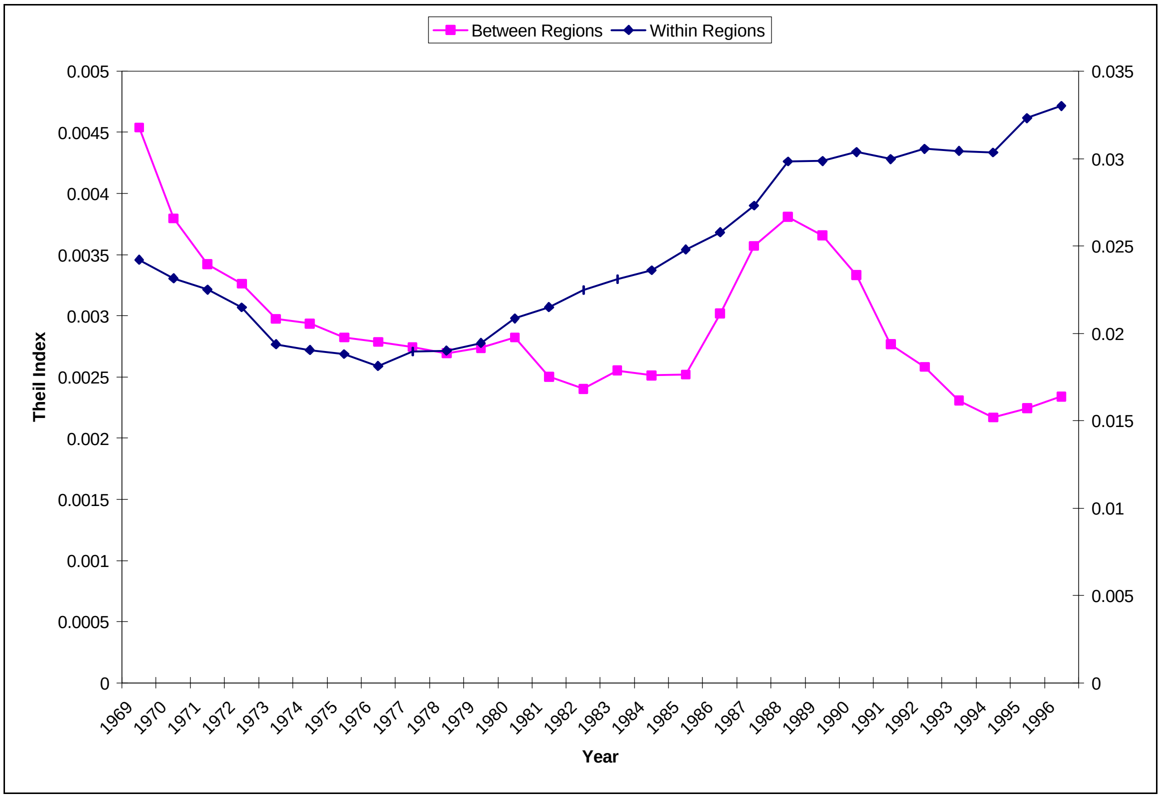

Figure 20- The dynamics of the inequality between the Census regions and of the inequality within the Census regions for the US from 1969 to 1996.

In order to further show the analytical potential of the Theil Index, we have analyzed the evolution of the overall inequality level in the US for the period considered but breaking down that inequality across regions, as shown in Figure 21. Now, we are simply positioning ourselves at the second level of the tree presented in Figure 12 and breaking down the within regions inequality contribution into the inequality within each of the Census regions. Again, and applying the property of neat decomposition of the Theil Index, we may add up these together plus the inequality between the regions to obtain the overall inequality.

Some facts deserve attention. The West North Central region is by far the most unequal region accounting for about 32% of the overall within-regions inequality for every year from 1969 to 1996. The South Atlantic region follows with a constant share of about 12%. Conversely, the Middle Atlantic region is the most equal region over this period.

The other six regions have similar contributions for the within regions inequality component.

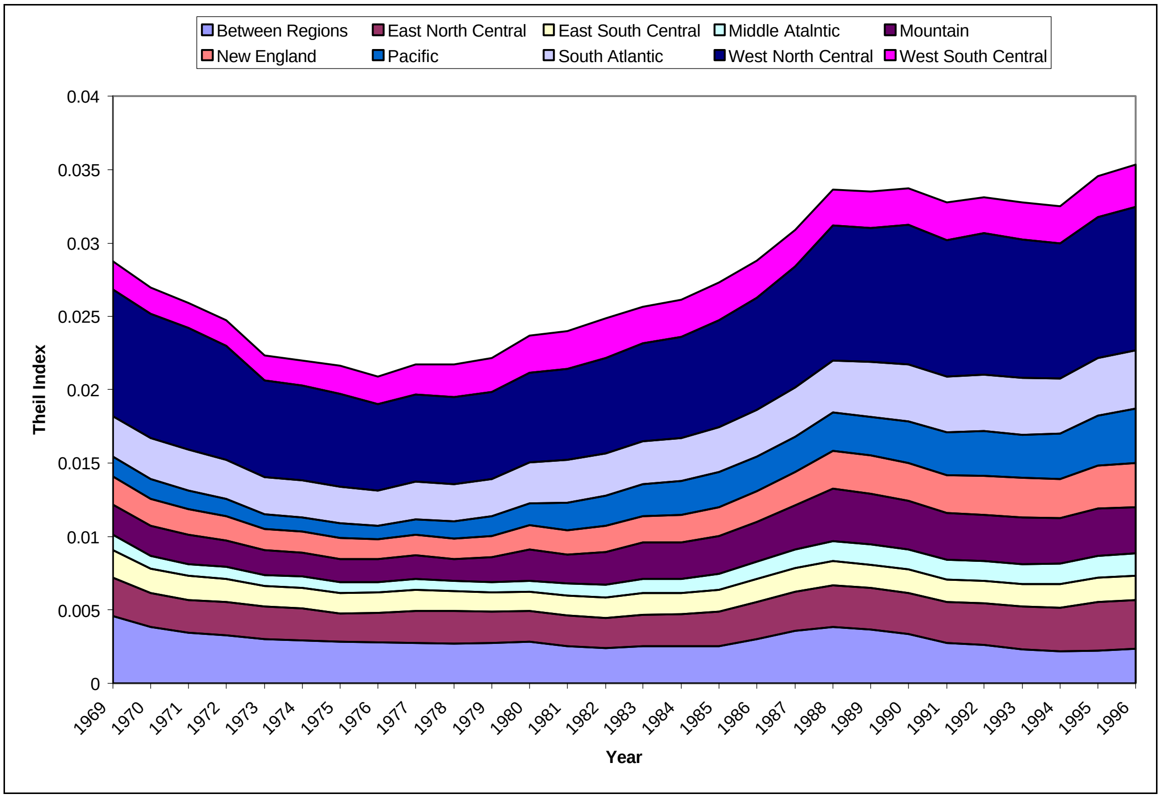

Figure 21- Breakdown of the inequality in the US per Census region from 1969 to 1996.

These same results are presented in Table 8, which shows the average contribution (over time) of each of the components of inequality depicted in Figure 21. The table allows us to see that the between regions component is no more, on average, than 11% of the overall between county inequality in the US from 1969 to 1996.

Table 8- Average over Time and Standard Deviation of the each of the Within Regions Contribution and of the Between Region Contribution

	Average	St Deviation
East North Central	0.101	0.008
East South Central	0.063	0.010
Middle Atalntic	0.041	0.004
Mountain	0.097	0.011
New England	0.081	0.005
Pacific	0.079	0.021
South Atlantic	0.127	0.006
West North Central	0.315	0.024
West South Central	0.095	0.015
Between Regions	**0.109**	**0.025**

The illustrations in this section served not only to show how to use the software application developed to ease the computation of the Theil index at different levels of aggregation, but offered two more features. First, they provided a new interpretation of the Theil index as a recursive measure of inequality that can be understood with the conceptual framework of a tree. Secondly, they showed, in a way similar to the application that was performed in section 2, the analytical potential of the Theil index.

4 - CONCLUSIONS

This paper is an exercise in the exploration of the Theil index. We started by suggesting intuitive interpretations, giving a motivation to construct inequality measures that departs from individuals clustered in groups, rather than from an individual level. The fundamental idea behind the Theil index, thus, is that it provides a way to measure the discrepancy between the structure of the distribution of income (or income) across groups and the structure of the distribution of individuals across those same groups. Groups that have their "fair share" of income contribute nothing to the Theil index. If all groups have their "fair share" of income, the Theil index attains its minimum value: zero. We approached the construction of the Theil index as the result of a quest to construct measures of inequality that could provide a numeric expression to such a discrepancy between the structure of the distribution of income and the structure of the distribution of population.

We then showed how the Theil index can be decomposed into a between group and within groups contributions to an overall inequality measured, and we explored several analytical applications of this property. These explorations extend and complement the work of Conceição and Galbraith (1998), which showed the value of the Theil index as a generator of long and dense measures of inequality. The specific applications included the construction of a measure of the inequality in the distribution of income across countries in the world and across counties in the US.

Finally, we extended the work of the paper into the development of a computer application that takes advantage of the understanding of the Theil index as having a tree structure. The way in which this application works was thoroughly presented in the paper, and the Excel macro will be made available at the UTIP website.

REFERENCES

Alison, P. D. (1978). “Measures of Inequality,” *American Sociological Review*, 43(December): 865-880.

Bourguignon, F (1979). “Decomposable Inequality Measures”, *Economotrica*, vol. 47.

Conceição, P., Galbraith, J. K. (1998), “Constructing Long and Dense Time-Series of Inequality Using the Theil Index”, University of Texas Inequality Project Working Paper No. 1; available at: http://utip.gov.utexas.edu. Also available as The Jerome Levy Economics Institute Working Paper No. 259: http://www.levy.org/publications/pubmainset.html.

Cowell, F. A. (1980). “On the Structure of Additive Inequality Measures”, *Review of Economic Studies*, vol. 47.

Heston, A., Summers, R. (1991). “The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988,” *Quarterly Journal of Economics*, May: 327-368.

Sen, A. (1997). *On Economic Inequality*. Oxford: Clarendon Press.

Shannon, C. E. (1948). “A Mathematical Theory of Communication”, *Bell System Technical Journal*, vol. 27: 379-423.

Shorrocks, A. F. (1980). “The Class of Additively Decomposable Inequality Measures”, *Econometrica*, vol. 48.

Shorrocks, A. F. (1984). “Inequality Decomposition by Population Subgroups”, *Econometrica*, vol. 52.

Theil, H. (1967). *Economics and Information Theory*. Chicago: Rand McNally and Company.

Theil, H. (1996). *Studies in Global Econometrics*. Dordrecht: Kluwer Academic Publishers.

DATA APPENDIX

The table below indicates the countries which were used for the computations of world inequality across nations.

AFRICA	AMERICA	ASIA	EUROPE	OCEANIA
ALGERIA	CANADA	BANGLADESH	AUSTRIA	AUSTRALIA
BENIN	COSTA RICA	CHINA	BELGIUM	FIJI
BURKINA	DOMINICAN	HONG KONG	CYPRUS	NEW ZEALAND
BURUNDI	EL SALVADOR	INDIA	CZECHOSLOVA	PAPUA N. GUI.
CAMEROON	GUATEMALA	INDONESIA	DENMARK	
CAPE VERDE	HONDURAS	IRAN	FINLAND	
CENTRAL AFR	JAMAICA	ISRAEL	FRANCE	
CHAD	MEXICO	JAPAN	GERMANY,	
COMOROS	NICARAGUA	JORDAN	GREECE	
CONGO	PANAMA	KOREA,	HUNGARY	
EGYPT	TRINIDAD&TO	MALAYSIA	ICELAND	
GABON	U.S.A.	PAKISTAN	IRELAND	
GAMBIA	ARGENTINA	PHILIPPINES	ITALY	
GHANA	BOLIVIA	SINGAPORE	LUXEMBOURG	
GUINEA	BRAZIL	SRI LANKA	NETHERLANDS	
GUINEA-BISS	CHILE	SYRIA	NORWAY	
IVORY COAST	COLOMBIA	TAIWAN	POLAND	
KENYA	ECUADOR	THAILAND	PORTUGAL	
LESOTHO	GUYANA		SPAIN	
MADAGASCAR	PARAGUAY		SWEDEN	
MALAWI	PERU		SWITZERLAND	
MALI	URUGUAY		TURKEY	
MAURITANIA	VENEZUELA		U.K.	
MAURITIUS			YUGOSLAVIA	
MOROCCO				
MOZAMBIQUE				
NAMIBIA				
NIGERIA				
RWANDA				
SENEGAL				
SEYCHELLES				
SIERRA LEON				
SOUTH AFRICA				
SUDAN				
TOGO				
TUNISIA				
UGANDA				
ZAMBIA				
ZIMBABWE				