The Theil Index in Sequences of Nested and Hierarchic Grouping Structures: Implications for the Measurement of Inequality through Time with Data Aggregated at Different Levels of Industrial Classification

by Pedro Conceição, James K. Galbraith and Peter Bradford

Pedroc@uts.cc.utexas.edu Galbraith@mail.utexas.edu bradford@mail.utexas.edu

University of Texas Inequality Project, LBJ School of Public Affairs

The University of Texas at Austin

Austin TX 78713

UTIP Working Paper Number 15

May 9, 2000

ABSTRACT

This paper discusses the implications of the decomposition property of the Theil index in sequences of nested and hierarchic grouping structures, formalizing general results applicable to a generic sequence of grouping structures. A specific application to data on wages and employment by industrial classification to measure the evolution of wage inequality through time will be explored, analyzing the links between Theil indexes computed at different levels of n-digit SIC codes. A dynamic analysis shows the extent to which a between group Theil statistic tracks the evolution of inequality within industries, and estimations are provided as to the amount of information gained by using ever more disaggregated grouping structures to assess the dynamics of overall inequality. The empirical illustration provides a monthly time-series for industrial earnings inequality in the US is computed at 2, 3 and 4-digit SIC codes from January of 1947 to March of 1999.

1- INTRODUCTION

Conceição and Galbraith (1998) argue that the Theil index, when applied to measure the between-industry dispersion of wages, can be used to construct long and dense time-series of inequality. The usage of the Theil index in such a way measures inequality between sectors, but fails to capture the level of inequality within each of the sectors. However, these authors provide formal criteria under which the between sector Theil index tracks the overall movement of inequality, concluding that, under some very general conditions, the dynamics of overall inequality can be captured by using only the between sector component of the Theil index.

This paper deepens and extends that argument, formally exploring the fractal properties of the Theil index and presenting a more compelling empirical illustration. The fractal property of the Theil index results directly from its unique characteristic of perfect decomposability, which allows for the separation of inequality into a between and a within groups components, provided that the groups are mutually exclusive and completely exhaustive (MECE). The fractal property will be explored in section 2, in the context of nested and hierarchic groups – of relevance when dealing with groups of firms aggregated at different levels of SIC codes – and also of non-hierarchic grouping structures. Section 3 provides a dynamic analysis, looking at the way in which the components of the Theil index co-change over time. In particular, this section analyzes the relationship between the dynamics of the Theil index computed at different levels of aggregation. Finally, section 4 provides an empirical illustration, based on a monthly time-series of inequality for the US, from January 1947 to April 1999, measured by industrial earnings Theil indexes computing the between-industries inequality at 2, 3 and 4 digit SIC codes.

2- STATICS: THE FRACTAL NATURE OF THE THEIL INDEX AND INEQUALITY DECOMPOSITION ACROSS SEQUENCES OF GROUPING STRUCTURES

The Theil index (Theil, 1967) is normally written as:

[1] $$T = \frac{1}{n}\sum_{p=1}^{n}\frac{y_p}{m_Y}.\log\left(\frac{y_p}{m_Y}\right)$$

where n is the number of individuals in the population, y_p is the income of the person indexed by p, and m_Y is the population's average income. Theil's T can also be expressed as:

[2] $$T = \sum_{p=1}^{n}\frac{y_p}{Y}\log\left[\left(\frac{y_p}{Y}\right)\Big/\left(\frac{1}{n}\right)\right]$$

with Y representing the population's total income, $Y = \sum_{p=1}^{n} y_p$.

Expressing the Theil in the less familiar form [2] highlights a possible intuitive interpretation of the Theil index as a direct measure of the discrepancy between the distribution of income and the distribution of individuals between mutually exclusive and completely exhaustive (MECE) groups, as suggested in Conceição and Ferreira (2000). In other words, a group that has the same share of income as the group's share of individuals does not contribute to inequality. In the specific case where individuals are not grouped, each individual's "population share" is merely $1/n$. This "population share" is compared with the individual's income share y_p/Y. If the population and income shares are equal, then there is no contribution by person p to the Theil index. If the shares are different from each other, then person p either has more or less income than the $1/n$ "fair share" (which

implies that at least one other, but quite possibly more than one, individual also has less or more than the "fair share" of income) and Theil's *T* increases from zero.

Expressing the Theil index with [2] also highlights the Theil's self-similar nature for any grouping structure chosen to aggregate individuals. After grouping all the individuals into *m* generic MECE groups, overall inequality can be completely and perfectly decomposed into a between-group component (T_g, where *g* is a "tag" identifying a specific grouping structure) and a within-group component (T^W_g). Thus:

[3] $$T = T_g + T^W_g$$

The self-similar nature of the Theil index becomes evident when one notes that:

[4] $$T'_g = \sum_{i=1}^{m} \frac{Y_i}{Y} \log\left[\left(\frac{Y_i}{Y}\right) \Big/ \left(\frac{n_i}{n}\right)\right]$$

Now *i* indexes not an individual but a group, with n_i representing the number of individuals in group *i*, and Y_i the total income in group *i*. Note that the structure of [4] (that gives the inequality between the groups defined by the grouping structure *g*) is exactly the same as the structure of [2], which defines inequality among individuals. Thus, the structure of the Theil index at measuring inequality between individuals is similar to the structure of the Theil index measuring inequality between groups.

There is yet another level to explore: the within group component of overall inequality, T^W_g, which is given by a weighted average of the Theil indexes for each group, the weights being each group's income shares:

[5] $$T^W_g = \sum_{i=1}^{m} \frac{Y_i}{Y}{}^{i} T$$

The Theil index for each group, ${}^{i}T$, corresponds to the inequality only between those individuals that are members of group *i* and is given by:

[6] $$ {}^{i}T = \sum_{p=1}^{n_i} r_{ip} $$

with:

[7] $$ r_{ip} = \frac{y_{ip}}{Y_i} \log\left[\left(\frac{y_{ip}}{Y_i} \right) \Big/ \left(\frac{1}{n_i} \right) \right] $$

In [6] and [7] each individual is indexed by two subscripts: *i* for the unique group to which the individual belongs, and *p*, where, in each group, *p* goes from 1 to n_i. Since ${}^{i}T$ only measures inequality between the individuals of group *i*, the relevant shares to be compared are y_{ip}/Y_i and $1/n_i$. But the *structure* of the inequality measure remains the same as the structure of the Theil index that accounts for the inequality between all the individuals in the population [2] and the inequality between groups [4]. The difference is that in [6] the context in which inequality is measured is limited to group *i*.

So far, we considered only one grouping structure: we partitioned the population into *m* groups, aggregating individuals in categories that are MECE. But these *m* groups can also be aggregated-up with a new grouping structure into a number of higher order groups. And these last groups may also be aggregated with yet another grouping structure into even higher order groups, and so forth. At each level of aggregation, the Theil index can be used not only to compute inequality between groups, but also to link the inequality measured at one level with that at any other level. We will show that the structure of the Theil index maintains, at every level of aggregation, its self-similar nature. We will derive

some of the implications that result from the self-similar behavior of the Theil that are useful when exploring the extent to which the evolution of a between-group component can approximate the dynamics of overall inequality.

There are several circumstances where the application of a sequence of grouping-structures may be of interest. For example, since the population of the United States can be aggregated by county, using the Theil index permits the decomposition of overall inequality in the US into a between-county component and a within county component. But counties can be aggregated into states, and states can be aggregated into regions; thus, the Theil index allows the within/between decomposition of inequality to be performed at ever-higher levels of aggregation. This exercise allows, for example, the determination of the contribution to overall inequality that can be attributed to asymmetries in the level of average income across regions. While the results presented in this paper will be general – applicable to any sequence of nested and hierarchic grouping structures – our main interest is in the sequence of grouping structures that is embodied in an industrial classification system. The reason is that, as Conceição and Galbraith (1998) argue, the between-industry component of the Theil index can provide, under certain general conditions, a good approximation of the dynamics of overall inequality. In this paper we will look at inequality computed at different levels of SIC codes, and we try to determine the "information gain" when we move towards a higher number of SIC digits. This section and the next will be focused on getting analytical results. In section 4 we provide an empirical application, using long and dense time-series of inequality (monthly, from 1947 to 1999) of industrial earnings computed at two, three and four digits levels of aggregation.

We will first address the specification of a sequence of nested and hierarchic grouping structures. Starting with the grouping of individuals, every grouping structure at this level is just a partition of the population into *m* groups, where we also need to specify the number of individuals in each group, n_j, where *j* indexes the group, and goes from 1 to *m*. Thus, to specify a grouping structure at the individual level we need to give only the structure of the partition of the population: $\{m; n_1; n_2; n_3; \ldots, n_m\}$. At a second level of aggregation, the second grouping structure needs to specify a similar partition of the *m*

groups into a number lower than *m* of groups, and also the second order group to which each of the n_j belongs. If we think about how to index individuals, with two grouping structures we need to attach to each individual two indexes, to uniquely designate to which group each individual belongs at each level of aggregation.

Figure 1 illustrates the specification of a sequence of grouping structures with three levels of aggregation. It is more convenient to make the specification going from the highest to the lowest level of aggregation. The highest-level grouping structure has *m* groups, where *k* – that indexes the groups at this level – goes from 1 to *m*. In the immediately lower-level grouping structure, for each group *k*, *j* (which indexes the groups within *k*) goes from 1 to m_k. Continuing to an even lower level of aggregation, for each group *j* that is part of *k*, *i* goes from 1 to m_{kj}. And finally, at the individual level, each person in *i* (where *i* is in *j* and *j* is in *k*) is index by *p*, where *p* goes from 1 to m_{kji}.

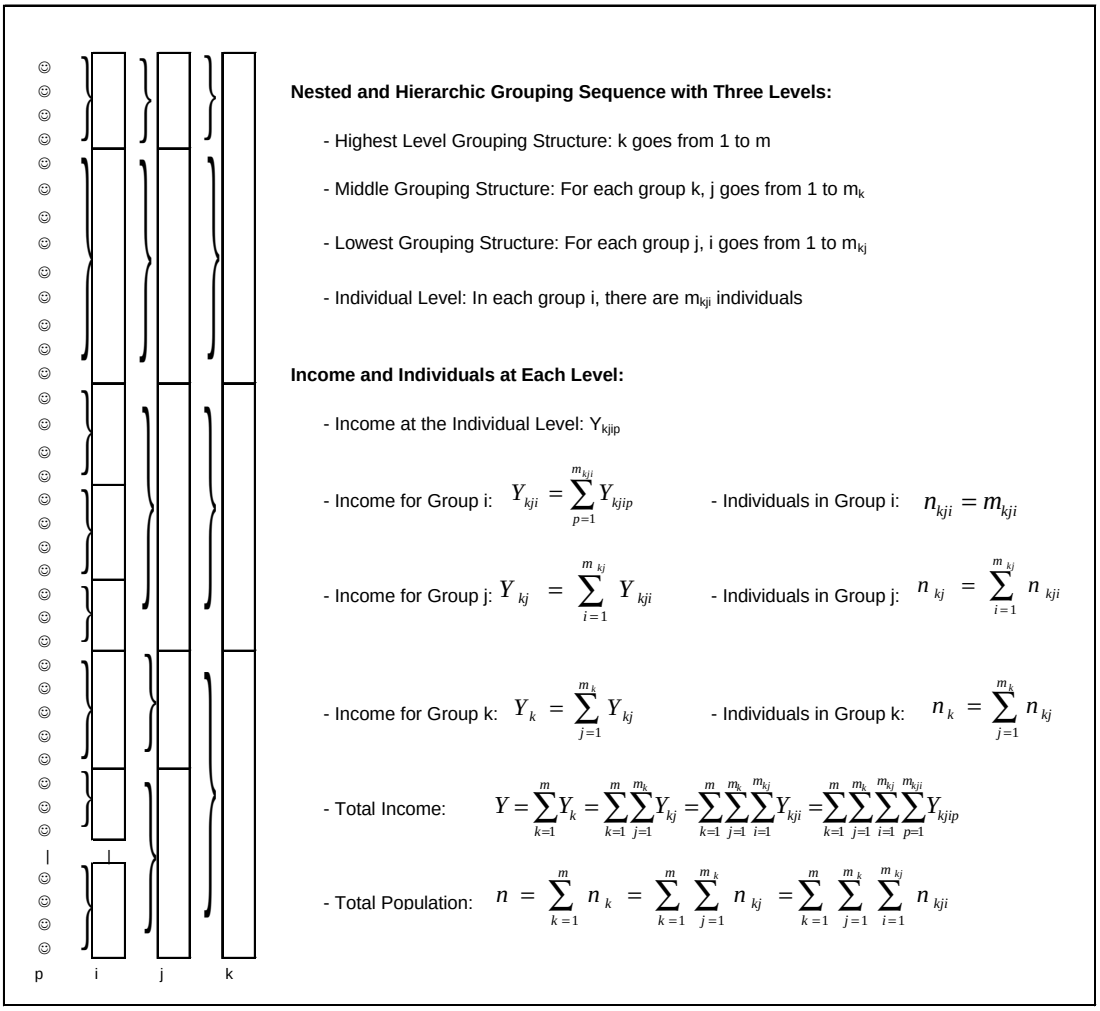

Figure 1- Specifying a Sequence of Nested and Hierarchic Grouping Structures with Three Levels of Aggregation

Consequently, each individual's income is indexed by four subscripts: Y_{kjip}. The total income for group *i*, Y_{kji}, is given by the summation of Y_{kjip} when *p* goes from 1 to m_{kji}. Figure 1 provides details on how to compute income and population at different levels of aggregation.

The Theil index is given by:

[8] $$T = \sum_{k=1}^{m} \sum_{j=1}^{m_k} \sum_{i=1}^{m_{kj}} \sum_{p=1}^{m_{kji}} \frac{Y_{pijk}}{Y} \log\left[\left(\frac{Y_{pijk}}{Y} \right) \Big/ \left(\frac{1}{n} \right) \right]$$

The nested structure entails that we need to use a series of summations in sequence, and the hierarchic nature of the sequence of grouping structures means that the upper limit in each summation is dependent on the index of the previous summations.

More generally, we can have more than the four levels (the individual and three sequential grouping structures) considered above. Consider that we have $l>1$ levels, from the individual to the highest level of aggregation. We can index each grouping structure in the sequence G by g, where $g=1,2,3,\ldots,l$. At each level of aggregation g the indexation is provided by the subscript i_g, with the highest level of aggregation being indexed by $g=1$ and the individual level by $g=l$. Therefore, at the highest level of aggregation, each group is indexed by i_1, with i_1 going from 1 to m. At the immediately lower level of aggregation, $g=2$, i_2 goes from 1 to m_{i_1}, for each i_1 in the grouping structure $g=1$. For the generic grouping structure g, i_g goes from 1 to $m_{i_1 i_2 i_3 \ldots i_{g-1}}$ in each group i_{g-1} in g-1, which is part of group i_{g-2} in g-2, and so forth all the way up to the highest level of aggregation where $g=1$. At the individual level, each individual's income is indexed by a sequence of subscripts, each subscript corresponding to one of the grouping structures in the sequence: $Y_{i_1 i_2 i_3 \ldots i_g \ldots i_l}$. The Theil index for the generic sequence of l hierarchic and nested grouping structures is given by:

[9] $$T = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \sum_{i_3=1}^{m_{i_1 i_2}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \cdots \sum_{i_l=1}^{m_{i_1 \cdots i_{l-1}}} \frac{Y_{i_1 i_2 \ldots i_g \ldots i_l}}{Y} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_g \ldots i_l}}{Y} \right) \Big/ \left(\frac{1}{n} \right) \right]$$

The Theil index given in [9] can be decomposed, at any level of aggregation g, into a between group and within group component, as we saw in [3] above:

[10] $$T = T_{g}^{B} + T_{g}^{W}$$

The tag g is a number between l and 1 identifying which of the grouping structures in the sequence G was chosen for a breakdown of the Theil index. In other words, g provides the level of aggregation at which the between group/within group decomposition of the Theil index was performed.

The between group component T_{g}^{B} is given by:

[11] $$T'_{g} = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \sum_{i_3=1}^{m_{i_1 i_2}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \log \left[\left(\frac{Y_{i_1 i_2 \ldots i_g}}{Y} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g}}{n} \right) \right]$$

Where the total income for each of the groups at level g is given by:

[12] $$Y_{i_1 i_2 \ldots i_g} = \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \sum_{i_{g+2}=1}^{m_{i_1 \cdots i_{g+1}}} \sum_{i_{g+3}=1}^{m_{i_1 \cdots i_{g+2}}} \cdots \sum_{i_l=1}^{m_{i_1 \cdots i_{l-1}}} Y_{i_1 i_2 \ldots i_g i_{g+1} i_{g+2} i_{g+3} \ldots i_l}$$

and the population of each of the groups at level g is given by[1]:

[13] $$n_{i_1 i_2 \ldots i_g} = \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \sum_{i_{g+2}=1}^{m_{i_1 \cdots i_{g+1}}} \sum_{i_{g+3}=1}^{m_{i_1 \cdots i_{g+2}}} \cdots \sum_{i_{l-1}=1}^{m_{i_1 \cdots i_{l-1}}} m_{i_1 i_2 \ldots i_g i_{g+1} i_{g+2} i_{g+3} \ldots i_{l-1}}$$

[1] Note that each individual counts as 1, so that $m_{i_1 i_2 i_3 \ldots i_{l-1}}$ gives the number of individuals in each of the groups of the most disaggregated grouping structure. While every individual counts as 1, each person's income is different, and equal to $Y_{i_1 i_2 i_3 \ldots i_g \ldots i_l}$, which leads to the fact that, in terms of the population aggregations, the summations goes only to the level l-1, while for income they go all the way to level l.

The within group component, T^W_g is, as we saw above, merely a weighted sum of the Theil index between individuals in each group, where the weights are the income shares of the groups:

$$[14] \qquad T_g^W = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \sum_{i_3=1}^{m_{i_1 i_2}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \left({}^{i_1 i_2 \ldots i_g} T \right)$$

The Theil index within each of the groups at level g is given by:

$$[15] \qquad {}^{i_1 i_2 \ldots i_g} T = \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \sum_{i_{g+2}=1}^{m_{i_1 \cdots i_{g+1}}} \sum_{i_{g+3}=1}^{m_{i_1 \cdots i_{g+2}}} \cdots \sum_{i_l=1}^{m_{i_1 \cdots i_l}} \frac{Y_{i_1 i_2 \ldots i_g \ldots i_l}}{Y_{i_1 i_2 \ldots i_g}} \log \left[\left(\frac{Y_{i_1 i_2 \ldots i_g \ldots i_l}}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{1}{n_{i_1 i_2 \ldots i_g}} \right) \right]$$

One other measure of inequality of interest is, for each level g, the Theil index within each group of g that accounts for the inequality between the groups at the immediately lower level of aggregation, T_g. In other words, T_g represents the within-group inequality at level g that considers only the sub-groups at the immediately lower level of aggregation g+1 as "individuals". For the grouping structure g, T_g is given by:

$$[16] \qquad T_g = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \sum_{i_3=1}^{m_{i_1 i_2}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \left({}^{i_1 i_2 \ldots i_g} T^w \right)$$

where

[17] $$ {}^{i_1 i_2 \ldots i_g}T^{w} = \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g i_{g+1}}}{n_{i_1 i_2 \ldots i_g}} \right) \right] $$

Note that ${}^{i_1 i_2 \cdots 1 g}T^{w}$ is different from ${}^{i_1 i_2 \cdots 1 g}T$ in that ${}^{i_1 i_2 \cdots 1 g}T^{w}$ accounts only for the inequality between groups at level g+1, for each group in g. The interest of T_g lies in the fact that it accounts for the inequality at level g as if the grouping structure g was the only one being considered, with the "individuals" grouped with this grouping structure being the groups at level g+1. For example, if we were aggregating individuals across US regions, and considering g the aggregation at the state level and g+1 the aggregation at the county level, then T_g would give us the within-state component of inequality associated with the dispersion across counties in each state.

One important result – which will be useful to determine the information gain or loss when we measure inequality with ever more disaggregated groups – is that, at any level g>1, the following expression is valid:

[18] $$ T_g = T_{g-1} + T_{g-1} $$

The formal proof of expression [18] is left for Appendix 1, but the intuition behind this decompositions of the between group component can be understood with the help of Figure 2. In words, the between group Theil at level g is the summation of the between group Theil at the immediately higher level of aggregation (g-1) plus the within group inequality at this higher level that measures the dispersion across the groups at level g. Again using the example mentioned above, where individuals are aggregated across US regions, expression [18] tells us that between-county inequality is equal to the between-state inequality plus the across-county inequality within states.

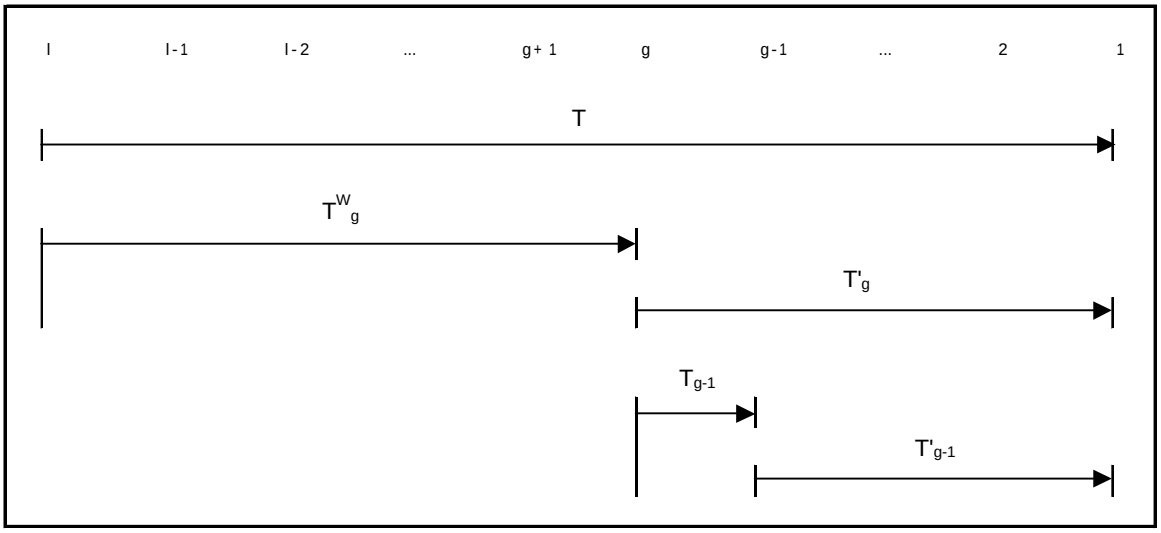

Figure 2- Decomposition of the Theil Index for a Generic Grouping Structure *g*, which is Member of a Sequence of Nested and Hierarchic Grouping Structures G

Given [18], overall inequality, *T*, can be perfectly and completely partitioned into the between group component at level *g* plus the summation of all components T_s for which $1<g<=s<l$[2]:

[19] $$T = T'_g + T_g + T_{g+1} + T_{g+2} + T_{g+3} + \ldots + T_{l-2} + T_{l-1} = T'_g + \sum_{s=1}^{l-g} T_{l-s}$$

The within group component T^W_g can thus be decomposed into a summation of T_s components, with $1<g<=s<l$. Each T_s gives the contribution to overall inequality, *T*, of the within-group Theil at level *s* across the groups at level *s*-1, except for T_{l-1} which gives, for

[2] We are considering only contiguous grouping structures, where *g* changes in increments of one, but the results can be generalized to increments larger than one.

each group at the lowest level of aggregation l-1, the between individual inequality[3]. In the special case where g=2, we obtain, since T_{2} is equal to T_1:

[20] $$T = T'_2 + T_2 + T_3 + \ldots + T_{l-2} + T_{l-1} = \sum_{s=1}^{l-1} T_s$$

Expression [20] provides yet another way to decompose total inequality as a summation of measures of inequality associated uniquely with each of the grouping structures in the nested and hierarchic sequence.

The between-individual inequality is not known in many empirical applications[4]. Therefore, it is convenient to isolate the between-individual component from overall inequality, which can be performed using the "usual" Theil index property allowing for a separation into a between group and within group components: $T = T_{l-1} + T^{W}_{l-1}$ which is equal to $T = T_{l-1} + T_{l-1}$. In this decomposition the between group component does not provide any information on the levels of inequality at each of the increasingly higher levels of aggregation. Expression [20] provides a decomposition of the between-group inequality at the lowest level of aggregation into a summation of the ever-increasing levels of aggregation. For example, continuing with the example we have been considering, suppose that at the lowest level of aggregation we have information on wages by plant. Then we can decompose the between-plant wage inequality in the US into a summation of the following: between-plant/within-county; between-county/within state; between-state/within the US.

To conclude this section, we consider next a special case of the sequence of grouping structures we have been considering. Thus far, the hierarchy of the sequence of

---

[3] Obviously T^{w}_{l} is zero, since there are no groups at the "within-individual" level.

[4] Even household surveys, for example, fail to capture inequality between individuals within the household.

nested grouping structures has been paramount. Counties are within states, and it would not be possible to ignore this hierarchy. However, in several cases the hierarchy is not relevant, as when we are interested in grouping individuals according to, say, demographic dimensions. To give a specific example, consider the case where we group individuals according to three variables: sex, age and educational level. The order of the groups is irrelevant. Relaxing the hierarchic condition corresponds to a less general situation than the one we have been considering. Figure 3 helps to understand how to specify the grouping structure; we only need to determine to which of the cubic cells each individual belongs, since the order of the groups is irrelevant.

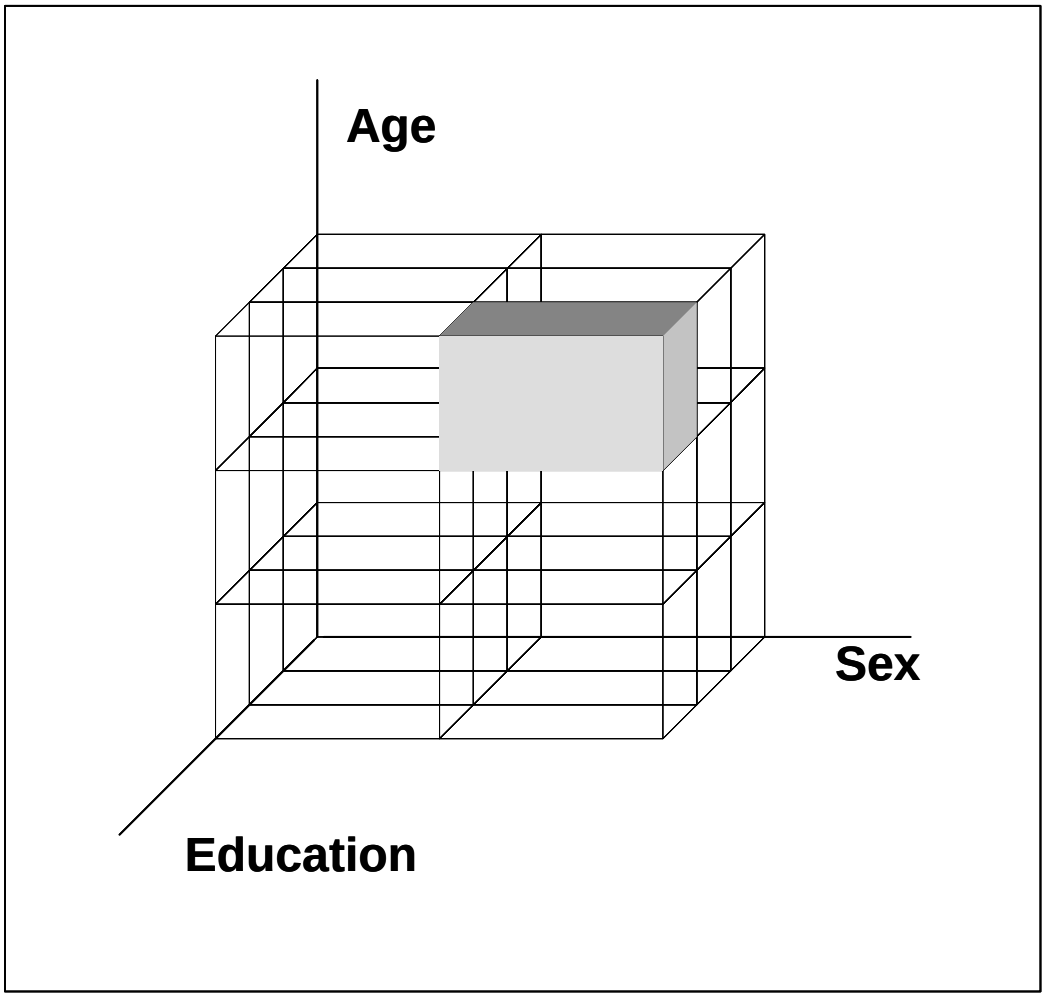

Figure 3- Specifying a non-hierarchic Sequence of Grouping Structures.

As we mentioned above, between-individual inequality is rarely know so our main interest will lie in discussing the between-cell inequality[5], which is given by:

$$[21] \qquad T' = \sum_{s=1}^{S}\sum_{e=1}^{E}\sum_{a=1}^{A} \frac{Y_{sea}}{Y} \log\left[\left(\frac{Y_{sea}}{Y}\right)\Big/\left(\frac{n_{sea}}{n}\right)\right]$$

where *s* indexes sex, *e* education levels and *a* age groups. Note that the fact that the grouping structures are not hierarchic means that the number of groups in each grouping structure (*S* for sex, *E* for education, and *A* for age) is independent from the other grouping structures.

The contribution of each of the between-group components to the Theil index follows from the usual Theil index decomposition:

$$[22] \qquad T\text{¢} = T\text{¢}_s + T\text{¢}^W{}_s = T\text{¢}_e + T\text{¢}^W{}_e = T\text{¢}_a + T\text{¢}^W{}_a$$

Expression [22] provides the between-group contribution of each of the grouping structures one at the time. For example, $T\text{¢}_s$ is the inequality across sexes and $T\text{¢}^W{}_s$ the inequality across age and education group for each sex class. Expression [23] provides a way decompose $T\text{¢}$ as a summation of *all* the between group components, plus an interaction term:

[5] Even if between-individual inequality is known, often the research question attempts to determine the extent to which overall inequality can be accounted for dispersion across groups.

[23] $$T' = \frac{1}{3}\left(T'_s + T'_e + T'_a\right) + \sum_{s=1}^{S}\sum_{e=1}^{E}\sum_{a=1}^{A}\frac{Y_{sea}}{Y}\log\left[\left(\frac{Y_{sea}}{\sqrt[3]{Y_s Y_e Y_a}}\right)\Big/\left(\frac{n_{sea}}{\sqrt[3]{n_s n_e n_a}}\right)\right]$$

Expression [23] may be of interest when one is interested in determining each grouping structure's contribution to inequality. Or, in other words, when one wants to establish how much of overall inequality can be attributed to differences across sexes, educational levels or age groups. This objective is usually achieved in the literature by running a regression analysis where individual income[6] is the independent variable, and where sex, education and age are used as regressors, as in, for example, Katz and Murphy (1992). The residual of the regression captures the within group inequality, and the remaining of the variation (captured in the coefficient) is associated with between-group dispersion. Expression [23] provides an alternative (or complementary) way to look at the contribution of inequality across different groups. A possible advantage of expression [23] is that it is a purely accounting formula, partitioning inequality across groups, while the usage of a regression analysis involves statistical estimation for an end that is exclusively of accounting.

The fact that the interaction term remains reflects the idea that it is not possible to express T' exclusively with between-group components, given that there is always a within-group contribution to inequality. However, [23] provides a way to detach this within-group component from the combined contribution of all the between-group contributions.

The structure of the interaction term is similar to the Theil index structure, but each cell's income and population (Y_{sea} and n_{sea}) are now divided by the geometric mean of the income and population across groups, while in the usual Theil structure they are divided by totals of income and population. An interpretation of this interaction term follows from a graphical description of the geometric mean, and is given in Appendix 2.

[6] Or income at the lowest available level of aggregassion.

In general, with l non-hierarchic grouping structures, the expression for inequality across groups at the lowest level of aggregation is[7]:

$$[24] \qquad T' = \frac{1}{l-1}\sum_{s=1}^{l-1} T'_s + \sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\cdots\sum_{i_{l-1}=1}^{m_{l-1}} \frac{Y_{i_1 i_2 \ldots i_{l-1}}}{Y} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_{l-1}}}{\sqrt[l-1]{\prod_{s=1}^{l-1} Y_s}}\right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_{l-1}}}{\sqrt[l-1]{\prod_{s=1}^{l-1} n_s}}\right)\right]$$

which can equally be written as:

$$[25] \qquad T' = \frac{1}{l-1}\sum_{s=1}^{l-1} T'_s + \sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\cdots\sum_{i_{l-1}=1}^{m_{l-1}} \frac{Y_{i_1 i_2 \ldots i_{l-1}}}{Y} \log\left(\frac{m_{i_1 i_2 \ldots i_{l-1}}}{\sqrt[l-1]{\prod_{s=1}^{l-1} m_s}}\right)$$

where $m_{i_1 i_2 \ldots i_{l-1}} = Y_{i_1 i_2 \ldots i_{l-1}} / n_{i_1 i_2 \ldots i_{l-1}}$ and $m_s = Y_s / n_s$. Therefore, group's $(i_1,i_2,\ldots,i_{l-1})$ contribution to the interaction term is zero whenever that group's average income is equal to the geometric mean of the average income of $i_1,i_2,\ldots,i_{l-1}$.

This section was devoted to an exploration of the implications of the decomposition property of the Theil index for sequences of grouping structures. The analysis was static, and aimed at, essentially, finding relationships between the Theil indexes computed at different levels of aggregation. Particularly important was to note the relationship between the Theil indexes computed at consecutive levels of aggregation. In the next section, these results will be explored in a dynamic analysis, where we look at the way in which the Theil index changes over time as income and population evolve.

---

[7] Note that the number of groups for the grouping structure g is now merely m_g, g=1,…,l-1.

3- DYNAMICS

In the dynamic analysis of this section, we will be looking only at inequality across groups, that is, we will not analyze dynamics of inequality across individuals. As in the discussion of non-hierarchic grouping structures in section 2, we will consider between-individual inequality as a residual that is either not possible to measure or of little substantive interest. We will also consider only fixed grouping structures, so that the number and relationship of the different levels of aggregation do not change over time.

In this context, the sources of between-group inequality variation over time are associated with income and population effects. Each effect is reflected in the way the Theil index responds to changes in the income and population shares. To analyze the Theil index response to income and population changes over time with generality we take the time derivative of the following expression:

[26] $$T' = \sum_{j=1}^{m} w_j \log\left(\frac{w_j}{e_j}\right)$$

where the group's income and population shares are[8]:

[8] Note that while *y* and *n* have no subscripts, the results we will obtain are valid even if *y* and *n* represent only a group within a higher level grouping structure. To keep the notation as simple as possible we will omit the inclusion of explicit subscripts at higher levels of aggregation.

[27]
$$\begin{cases} w_j = \dfrac{y_j}{y}, y = \sum_{j=1}^{m} y_j \\ e_j = \dfrac{n_j}{n}, n = \sum_{j=1}^{m} n_j \end{cases}$$

Since the grouping structure is constant, the Theil change over time is given by:

[28]
$$\dot{T}' = \sum_{j=1}^{m} \left(\frac{\partial T'}{\partial w_j} \dot{w}_j + \frac{\partial T'}{\partial e_j} \dot{e}_j \right)$$

The first term in the summation corresponds to the income effect, while the second to the population effect. The rates of change of the shares are given by:

[29]
$$\begin{cases} \dot{w}_j = w_j \left(g_j - g \right) \\ \dot{e}_j = e_j \left(p_j - p \right) \end{cases}$$

where g_j is the rate of change of income for group *j* and *g* is the overall rate of change of income, p_j is group's *j* rate of population change and *p* is the overall rate of population change:

[30]
$$\begin{cases} g_j = \dfrac{\dot{y}_j}{y_j}, g = \dfrac{\dot{y}}{y} \\ p_j = \dfrac{\dot{n}_j}{n_j}, p = \dfrac{\dot{n}}{n} \end{cases}$$

The intuition behind [29] is now immediate: the change in the shares of group *j* is proportional to the difference between the rate of change of income (population) in group *j* and the overall rate of income (population) change. If group's *j* income changes at the same rate as the change in overall income, then the share remains the same. If the growth rates differ, the change depends also on the level of the shares.

The dependency of the Theil on income and population changes is given by:

[31]
$$\begin{cases} \dfrac{\partial T'}{\partial w_j} = \log\left(\dfrac{w_j}{e_j}\right) + 1 \\ \dfrac{\partial T'}{\partial e_j} = -\dfrac{w_j}{e_j} \end{cases}$$

which in conjunction with [29] and [30] gives:

[32]
$$\dot{T}' = \sum_{j=1}^{m} w_j \left\{ \left(g_j - g\right)\left[\log\left(\frac{w_j}{e_j}\right) + 1\right] - \left(p_j - p\right) \right\}$$

Expression [32] shows how the Theil index reacts to changes in income and population. Essentially, the change in the Theil depends on the difference between the rates of change of the shares of income and population of each group with the overall rates of change of income and population. If a group's share of income and population change at the same rate as the change in overall income and population, then this group does not contribute to changes in the Theil index.

If g_j is different from *g* or p_j is different from *p*, then group *j* contributes to changes in the Theil. The way in which changes in income and population in group *j* affect the

Theil index depends on whether group *j* is a "poor" or a "rich" group (poor in the sense that the share of income is substantial lower than the share of population). The multiplicative factor $[\log(w_j/e_j)+1]$ determines whether (g_j-g) contributes positively or negatively to the change in Theil. If $[\log(w_j/e_j)+1]>0$ (meaning that we are dealing with a "rich" group) then if $g_j>g$ the income effect increases the Theil index, because a "rich" group gains income at a rate higher that the overall population, increasing inequality. Since (p_j-p) is preceded by a minus sign, the population effect works in a symmetric way (if a "rich" group gains population at a rate higher than the overall population growth, then inequality decreases). When $[\log(w_j/e_j)+1]<0$ we are dealing with "poor" groups, and the effect of income and population changes is opposite of the one described above.

The effect of $[\log(w_j/e_j)+1]$ goes beyond determining the sign that precedes (g_j-g), since it also weighs more heavily changes at the lowest shares of income. Figure 4 shows the dependency of $[\log(w_j/e_j)+1]$ on w_j, assuming that $e_j=0.5$.

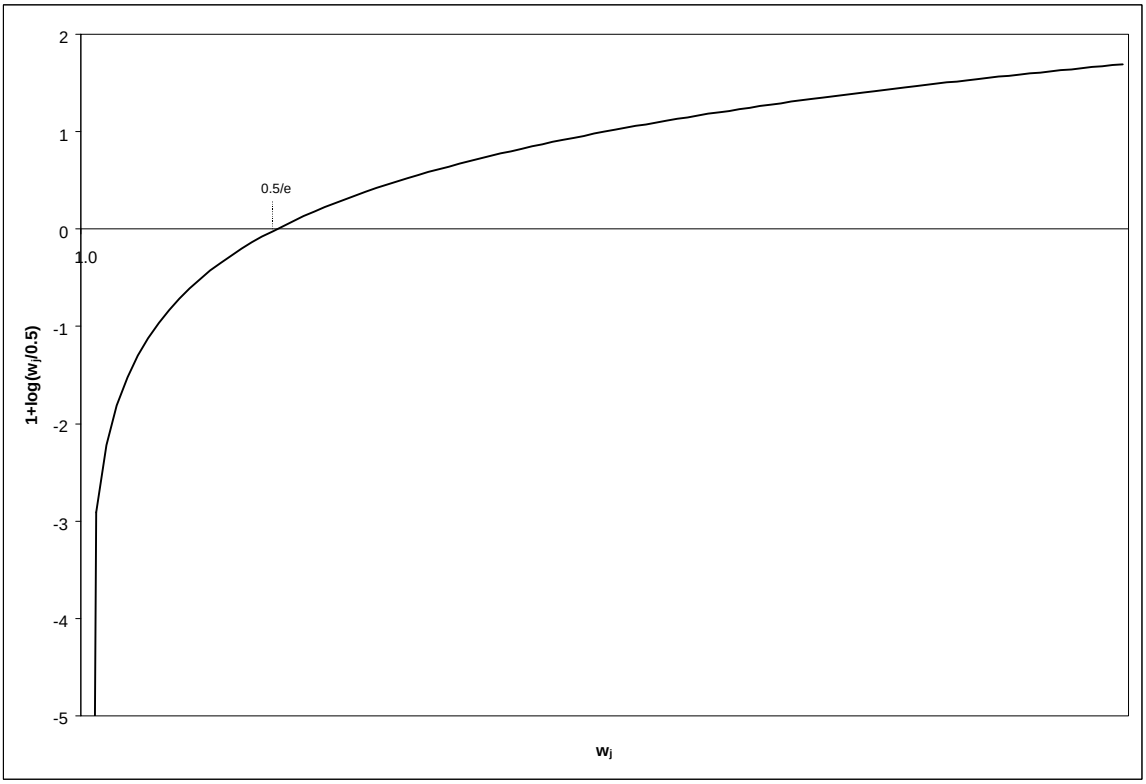

Figure 4- Multiplicative Factor of the Income Effect: Dependency on the Income Share

When the $w_j=e_j/e$ then $[\log(w_j/e_j)+1]=0$; wage shares lower than e_j/e lead to a negative $[\log(w_j/e_j)+1]$ and the reverse happens when the w_j is higher than e_j/e[9]. Additionally, the shape of the curve is such that when $[\log(w_j/e_j)+1]$ is negative (that is, when we are dealing with "poor" groups) the expression $[\log(w_j/e_j)+1]$ weighs more heavily on (g_j-g) than when $[\log(w_j/e_j)+1]$ is positive ("rich" groups).

What is the meaning of the "cut-off" point $w_j=e_j/e$ for which $[\log(w_j/e_j)+1]=0$? This is the point at which group *j*'s contribution to the Theil index attains its minimum, as illustrated in Figure 5. Conceição and Ferreira (2000) provide an intuitive interpretation of this "cut-off" point.

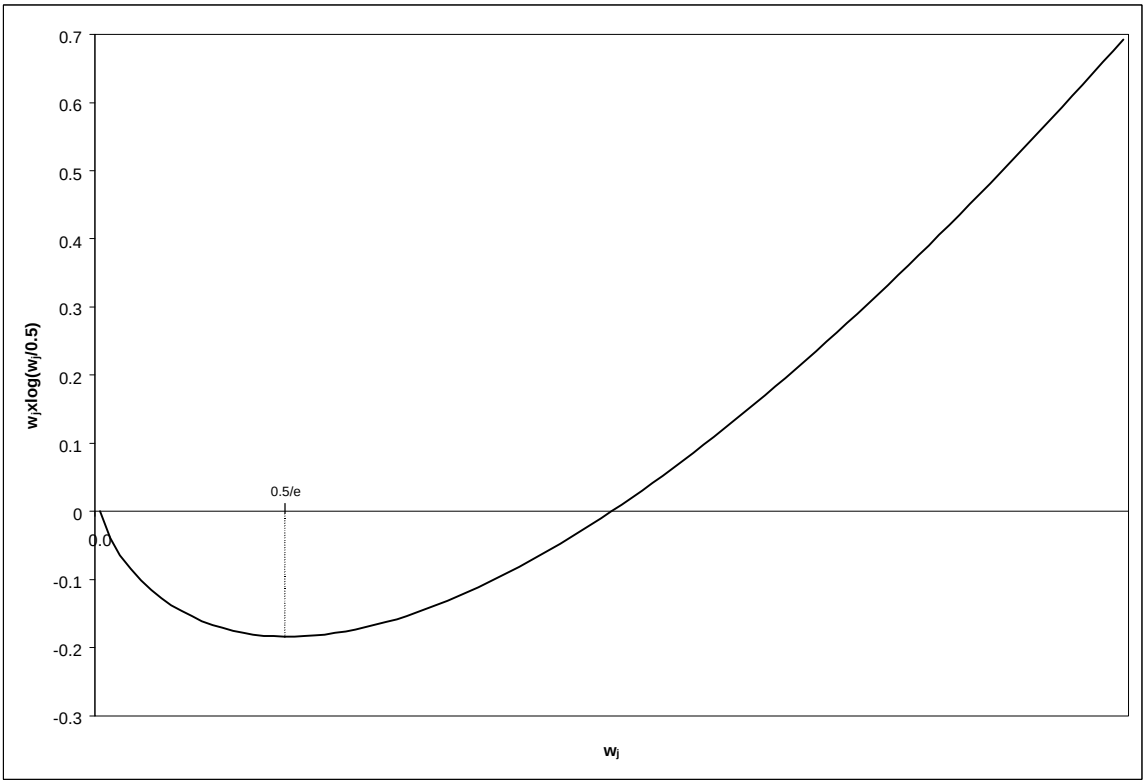

Figure 5- Group *j*'s Contribution to the Theil Index: Dependency on the Income Share

[9] Here "e" is Neper's number.

Yet another way to express the change over time of the Theil index is given by:

[33] $$\dot{T}' = p - gT' - \sum_{j=1}^{m} w_j \left[p_j - g_j \log\left(\frac{w_j}{e_j} \right) \right]$$

which shows that changes in the Theil can be separated into a "macro" component (*p-gT'*) and a "micro" component, which depends on the dynamics of the distribution of income and population across groups. This decomposition is explored in Conceição and Galbraith (forthcoming). The "micro" component is composed of two summations: $\sum_{j=1}^{m} w_j p_j - \sum_{j=1}^{m} w_j g_j \log\left(w_j / e_j\right)$. The structure of the micro component mirrors the structure of the macro component. The first summation adds the group's weighted rates of population change, where the weights are the income shares. The second summation is a modified Theil index, where the weights of $\log(w_j/e_j)$ are, instead of w_j, the changes in the wage shares (note that $w_j(t+1)=w_j(t)+g_j w_j(t)$; the weights are, then: $w_j(t+1)-w_j(t)$). So, in a way, the micro component is, as the macro component, a difference between the rates of population change and the rate of income change combined with the Theil index.

The general results on the between-group dynamics of the Theil index are important to formalize the relationship between the dynamics of the Theil index at different levels of aggregation. From [18], we know that the relationship between the levels of the between-group Theil index at consecutive levels of aggregation can be expressed as:

[18] $$T_{s+1} = T_s + T_s$$

this expression is important because it shows that the information gain associated with calculating the Theil index at a lower level of aggregation is a single additive factor. Since the Theil index is always positive or zero, we can also see that at a lower level of aggregation the Theil index will always be equal or higher than that computed at a higher level of aggregation. Additionally, expression [18] shows that, unless the distribution of income across the groups at level *s* within each *s*+1 group is homogeneous, the level of the Theil index at the two levels can be quite different.

However, our interest is in the dynamics of inequality. From [18], it is obvious that the rate of change at the lower level of aggregation is also the simple summation of the change at a higher level plus an additional term. In other words, the informational gain on the dynamics of the Theil index associated with a less aggregated grouping structure is given by:

[34] $$\dot{T}'_{s+1} - \dot{T}'_{s} = \dot{T}_{s}$$

The sign and scale of the term $\dot{T}_s$ is now ambiguous, because the information gain associated with considering a lower level of aggregation can either increase or decrease the rate of change at the higher level. The explicit expression for $\dot{T}_s$ is given by:

[35] $$\dot{T}_s = \sum_{i_1=1}^{m} \cdots \sum_{i_s=1}^{m_{i_1 \cdots i_{s-1}}} \frac{Y_{i_1 \cdots i_s}}{Y} p_{i_1 \cdots i_s} - gT_s + \sum_{i_1=1}^{m} \cdots \sum_{i_{s+1}=1}^{m_{i_1 \cdots i_s}} \frac{Y_{i_1 \cdots i_{s+1}}}{Y} \left[p_{i_1 \cdots i_{s+1}} - g_{i_1 \cdots i_{s+1}} \log\left(\frac{w_{i_1 \cdots i_{s+1}}}{e_{i_1 \cdots i_{s+1}}} \right) \right]$$

where

[36] $$\begin{cases} w_{i_1 \ldots i_{s+1}} = Y_{i_1 \ldots i_{s+1}} \big/ Y_{i_1 \ldots i_s} \\ e_{i_1 \ldots i_{s+1}} = n_{i_1 \ldots i_{s+1}} \big/ n_{i_1 \ldots i_s} \\ g_{i_1 \ldots i_s} = \dot{Y}_{i_1 \ldots i_s} \big/ Y_{i_1 \ldots i_s} \\ p_{i_1 \ldots i_s} = \dot{n}_{i_1 \ldots i_s} \big/ n_{i_1 \ldots i_s} \end{cases}$$

Expression [35] shares the general structure of the rate of change of the between-group Theil shown in [33], but is slightly more complex, given that T_s is not a "pure" between-group Theil. In the context of [35], micro now means at level s+1. Therefore, the first two terms to the right of the equality sign in [35] still reflect macro behavior. The first term is a weighted summation of the population shares at level s, with the weights being the income shares of all the groups in s. The second term is the overall income growth rate "corrected" by T_s, following the structure of the macro term in [33]. The micro term is the weighted summation of the differences between the growth in population shares and corrected wage shares at level s+1. The weights are the income shares of the groups at level s. The correction in the change rate of the wage shares is given by logarithm of the ratio between the shares of each group s+1 in each s and the corresponding population shares.

The information gain on the dynamics of the Theil index when one moves to a less aggregated grouping structure is zero whenever $\dot{T}_s$ is zero. From [35] it can be shown that if the following expression occurs for every group in s then $\dot{T}_s = 0$:

[37] $$g^{i_1 \ldots i_s} T^w - p_{i_1 \ldots i_s} = \sum_{i_{s+1}=1}^{m_{i_1 \ldots i_s}} w_{i_1 \ldots i_{s+1}} \left[g_{i_1 \ldots i_{s+1}} \log\left(\frac{w_{i_1 \ldots i_{s+1}}}{e_{i_1 \ldots i_{s+1}}} \right) - p_{i_1 \ldots i_{s+1}} \right]$$

Expression [37], once again, can be interpreted as equality between macro behavior (in this case, at level s) and micro dynamics (at level s+1). Formally, [37] is

structurally similar to [33] (if we make $\dot{T}' = 0$ in [33]), and this interpretation follows from the discussion after [33].

While [37] is a sufficient condition for $\dot{T}_s = 0$, it is not a necessary condition. In fact, even if [37] is not verified for every group in *s*, $\dot{T}_s$ can be zero due to the interaction of negative and positive rates of change in [35]. Therefore, the ability to move further with general analytic results is limited by the possibility of complex interactions across the rates of growth of the shares of population and income of different groups at different levels of aggregation. Still, we were able to establish and understand the general structure of these interactions. The next step, developed in the following section, is to look at empirical results.

4- EMPIRICAL APPLICATION: MONTHLY US INTER-INDUSTRY WAGE INEQUALITY, 1947-1999

Informed by the knowledge of the static and dynamic behavior of the Theil index in sequences of nested grouping structures developed in sections 2 and 3, we devote this concluding section to an empirical application. The specific sequence of nested and hierarchic grouping structures that we consider is the industrial classification system, where firms are aggregated into industrial sectors. Industrial sectors can be defined at different levels of aggregation, the number of sectors at each level being indexed by one-digit SIC codes (the highest level of aggregation), two-digit codes, and so on. We will consider the Theil index at two, three and four digit SIC codes.

The continuous and smoothed monthly time-series for wage inequality across manufacturing sectors are presented in Figure 6. The tags of the between-industry Theil index in Figure 6 correspond to the number of SIC-codes digits considered. Smoothing was performed using a centered 12-month moving average. The original time-series are not continuous because the grouping structures (SIC classifications) change over time. Therefore, the construction of the continuous series presented in Figure 6 was obtained by joining the end points of the consecutive time-segments with consistent classification structures. The differences between the continuous and the non-continuous series are small, and especially the dynamics are not affected by considering one or the other (see Figure 7, which shows the discontinuous series).

Figure 6- Between-Industry Theil Index at Three Levels of Aggregation in Manufacturing

Figure 6 shows that the differences in the levels of the Theil index can, indeed, be substantial, when one compares the level of T'_2 with that of T'_3. But the figure also suggests that the dynamics are quite similar. Additionally, Figure 7 indicates that the difference across the Theil indexes at different levels of aggregation also follows a similar dynamics.

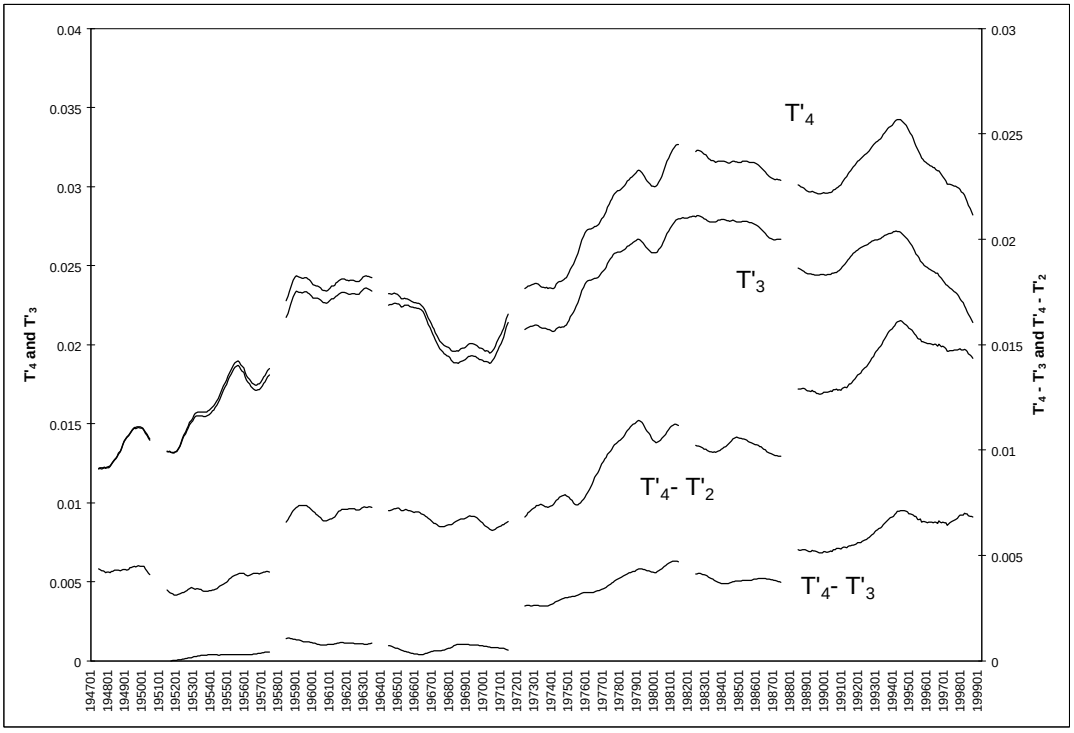

Figure 7- Levels and Differences Across Between-Industry Theil Index for Manufacturing at Different Levels of Aggregation

The differences across levels of the Theil represent the "information gain" associated with computing inequality with groups at lower levels of aggregation. The interpretation of this difference was explored in section 2 above. Since we are interested in comparing the dynamics, rather than the levels, we need to compare the month-to-month growth rates of the several time-series. Figure 8 shows the similarity between the monthly rates of change of the Theil indexes computed at the three levels of aggregation. We also show a comparison between the rates of change of T_3 and of $T_3 - T_2$.

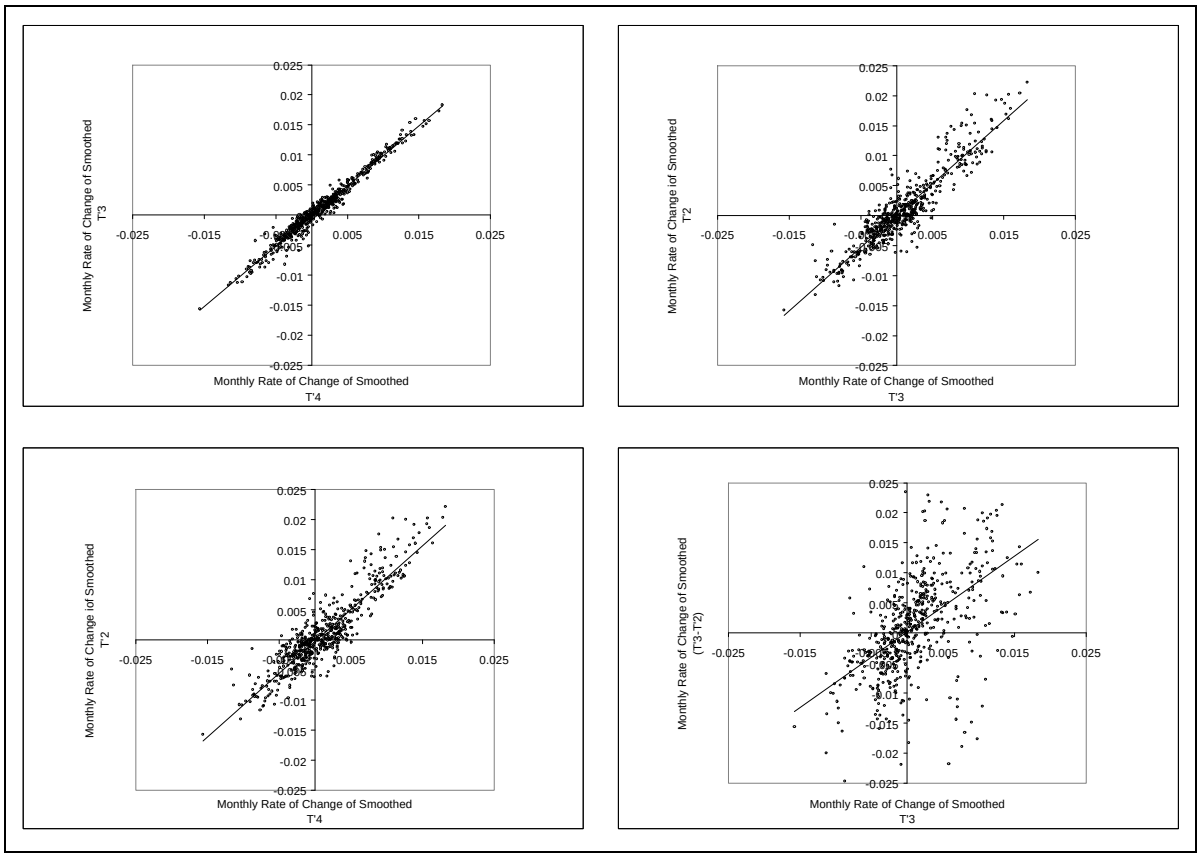

Figure 8- Comparing the Monthly Rates of Change at Different Levels of Aggregation for Manufacturing

The visual depiction of Figure 8 suggests that the monthly rates of change of the Theil indexes computed at the three levels of aggregation are, indeed, very similar. The similarity is much lower in the comparison between the rates of change of $T\text{¢}_3$ and of $T\text{¢}_3$ - $T\text{¢}_2$. The results of the univariate regressions suggested by each of the charts in Figure 8 help to determine more precisely the information gain on the dynamics of inequality associated with considering less aggregated groups. Table 1 shows that the intercept is, in all cases, zero, and the slope is close to one, for values of the adjusted R-squared higher than .8 in all cases.

Table 1- Results of the OLS Univariate Linear Regressions for Manufacturing

	Intercept		Slope		Adj.
	coeff.	t-stat	coeff.	t-stat	R^2
$\Delta T'3=f(\Delta T'4)$	**0.00**	-5.28	**1.00**	128.50	**0.97**
$\Delta T'2=f(\Delta T'3)$	**0.00**	-0.31	**1.06**	55.98	**0.85**
$\Delta T'2=f(\Delta T'4)$	**0.00**	0.00	**1.06**	50.07	**0.82**

N=543

The results presented thus far consider the entire time period under consideration as a whole. However, as we noted above, the grouping structure changed. Table 2 shows the changes in grouping structure over time. It also shows the differences in the number of groups considered at different levels of aggregation.

Table 2- Evolution of the Manufacturing Grouping Structures over Time

		47-50	51-57	58-63	64-71	72-81	82-87	88-99
SIC-4	total	38	44	111	115	203	206	226
	unique	5	5	62	66	151	155	176
SIC-3	total	35	41	**79**	**80**	**117**	**117**	**125**
	unique	33	39	**79**	**80**	**117**	**117**	**125**
SIC-2	total	**18**	**18**	**18**	**18**	**18**	**18**	**20**
	unique	**18**	**18**	**18**	**18**	**18**	**18**	**20**

Table 2 differentiates, for each level of aggregation, the total number of groups from those that are unique for that grouping structure. The difference between the total number and the number of unique groups results from the fact that not all the groups at a higher level of aggregation are split into smaller bins as we move towards a lower level of aggregation. The pairings (grouping structure, time frame) for which the total number of groups is equal to the number of unique groups are indicated in bold – naturally, at the highest level of aggregation all the values are in bold.

Table 2 shows there is a large difference in the number of groups when we move from an aggregation of industries at the 2-digit SIC level to the 3-digit level. Especially in later periods, the difference in the number of unique groups is very large. Still, the results of Table 1 suggest that the dynamics of inequality can be captured quite well at the higher level of aggregation when we take the entire time period under analysis as a whole. Do the results change when we break down the time period under analysis into the time frames for which the grouping structures are consistent? Table 3 shows the results of univariate regressions circumscribed to the grouping-structure consistent time frames[10].

Table 3- Comparing the Dynamics During Consistent Grouping-Structures Time Frames in Manufacturing

	$\Delta T'3=f(\Delta T'4)$			$\Delta T'2=f(\Delta T'3)$			$\Delta T'2=f(\Delta T'4)$		
	Intercept	Slope	R^2	Intercept	Slope	R^2	Intercept	Slope	R^2
47-50	0.00	1.00	1.00	0.00	1.03	0.85	0.00	1.03	0.85
51-57	0.00	1.00	1.00	0.00	1.03	0.85	0.00	1.07	0.92
58-63	0.00	1.02	0.98	0.00	1.04	0.86	0.00	0.75	0.80
64-71	0.00	1.08	0.98	0.00	1.04	0.86	0.00	1.11	0.90
72-81	0.00	0.99	0.97	0.00	1.04	0.86	0.00	1.05	0.58
82-87	0.00	0.94	0.82	0.00	1.05	0.86	0.00	0.70	0.36
88-99	0.00	0.92	0.83	0.00	1.05	0.87	0.00	0.96	0.75

The results of Table 3 confirm that the information gain associated with considering lower levels of aggregation is small. The intercept is zero in all cases, and the slope is close to one, within a less than .05 range, in most cases. The R^2 values are equally high. Note that for the relationship between the changes in the Theil computed with 3 and 4 digits, in 1964-1971, when there are 66 unique groups at the 4-digit level (out of a total of 115) the R^2 is .98 (at the 3-digit level there are only 80 groups). Even in the next period, 1972-1981, when the number of unique groups at the 4-digit level increases to 151 (out of a total of 203, with 117 groups at the 3-digit level) the R^2 decreases only slightly

[10] Statistical significance results are not as good as those reported when the time periods are combined, due, in large part, to the lower number of data points in each interval. R-squared values are not adjusted,

to .97. Therefore, there is virtually no information gained on the dynamics of cross-industry industrial earnings inequality when we move from 3 to 4 digits.

More importantly, the comparison with the results at the 2-digit level continues to reveal a high level of similarity. Naturally, the similarity is not as good as the one that exists between the 3 and the 4 digit levels, but note that the number of groups at the 2-digit level being considered is extremely small: 18 from 1947 to 1987 and 20 from 1988 to 1999. At the 4-digit level there are, for later years, almost ten times as many groups, and at the 3-digit level there are between four and five times as many groups as there are at the 2-digit level. Still, the R^2 of the changes in T_3 and in T_2 is high and stable, in the .85 range, even when the number of groups at the 3-digit level increases substantially. The relationship between T_4 and T_2 is more volatile, and during some time-periods there appears to be a considerable information loss associated with considering 2 instead of 4 digits if we are interested in the monthly dynamics, such as in the 1982-1987 period, for which there is the poorest fit.

In general, the results suggest that the information gain associated with computing the between-industry earnings inequality with the Theil index at lower levels of aggregation adds little information on the dynamics of inequality, at least when dealing with monthly data. Results with other classification schemes will vary, of course, but we conclude that standard industrial datasets can often provide a good source of information on inequality dynamics.

Finally, we conclude with a comparison of the Theil index for manufacturing only with the Theil index for the entire economy. Figure 9 shows the evolution of the US Theil for the same time frame considered for manufacturing only. The figure suggests that the dynamics of inequality for the entire US economy was similar to that of manufacturing alone; below we explore in more detail the degree of similarity between the dynamics of inequality in manufacturing and in the US as a whole.

---

and therefore should not be compared directly with those reported in Table 2.

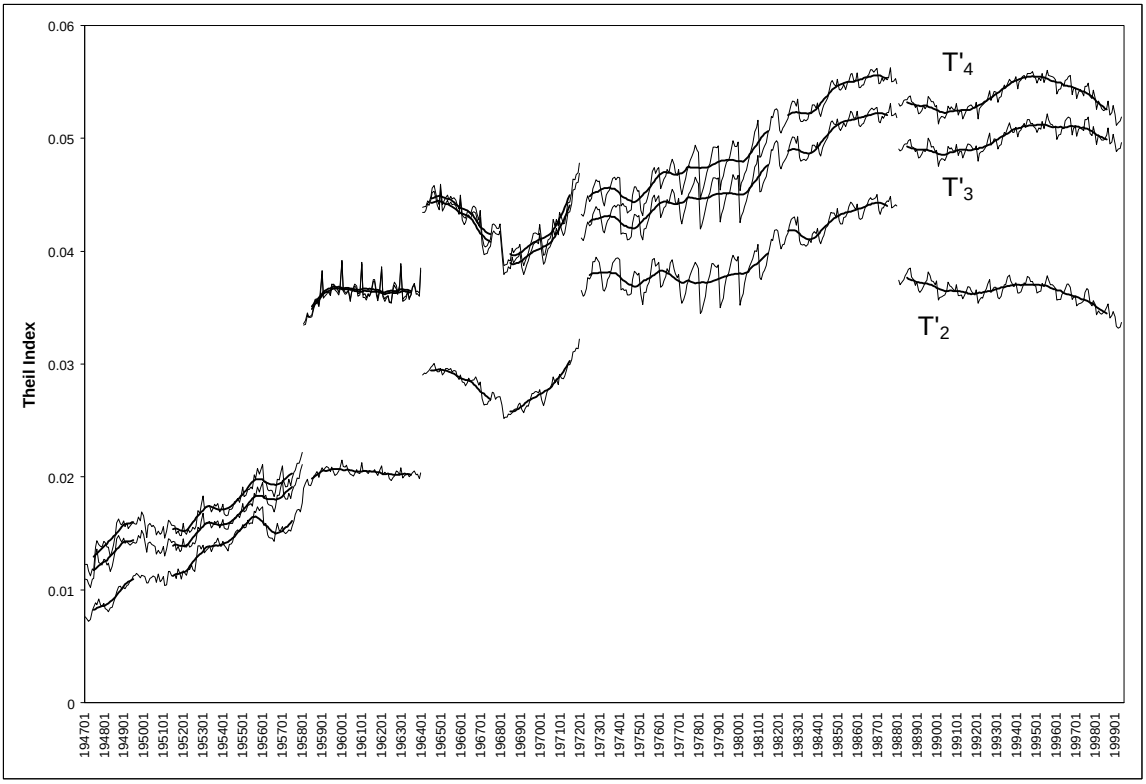

Figure 9- - Between-Industry Theil Index at Three Levels of Aggregation for the US Economy

The discontinuities in each of the series of inequality for the US correspond to transitions between changing industrial classifications. Table 4 shows the evolution of the different grouping structures. It is also important to note from Table 4 the higher number of groups considered to account for the entire US economy, as opposed to manufacturing only. In later years, the number of groups in the US economy is close or above to two times those considered for manufacturing alone.

Table 4- Evolution of the US Grouping Structures over Time

		47-49	50	51-57	58-63	64-67	68-71	72-81	82-87	88-99
SIC-4	total	47	49	55	144	152	153	292	307	369
	unique	6	6	6	64	68	68	153	163	216
SIC-3	total	43	45	51	111	116	117	205	213	247
	unique	39	41	47	108	112	113	201	209	242
SIC-2	total	**21**	**21**	**21**	**28**	**30**	**30**	**45**	**46**	**57**
	unique	**21**	**21**	**21**	**28**	**30**	**30**	**45**	**46**	**57**

Comparing the dynamics of the Theil index for the US economy as whole at different levels of aggregation yields essentially the same results already presented for manufacturing, strengthening our suggestion that little information is gained by considering extremely disaggregated grouping structures to measure the dynamics of inequality month by month. A new issue that we explore now is the extent to which data on manufacturing only provides an indication of the evolution of the US inter-industry inequality as a whole. Using the framework presented before for manufacturing only, we compare below the monthly rates of change of the Theil index in manufacturing with the monthly changes in the entire US economy. Figure 10 provides a graphic illustration.

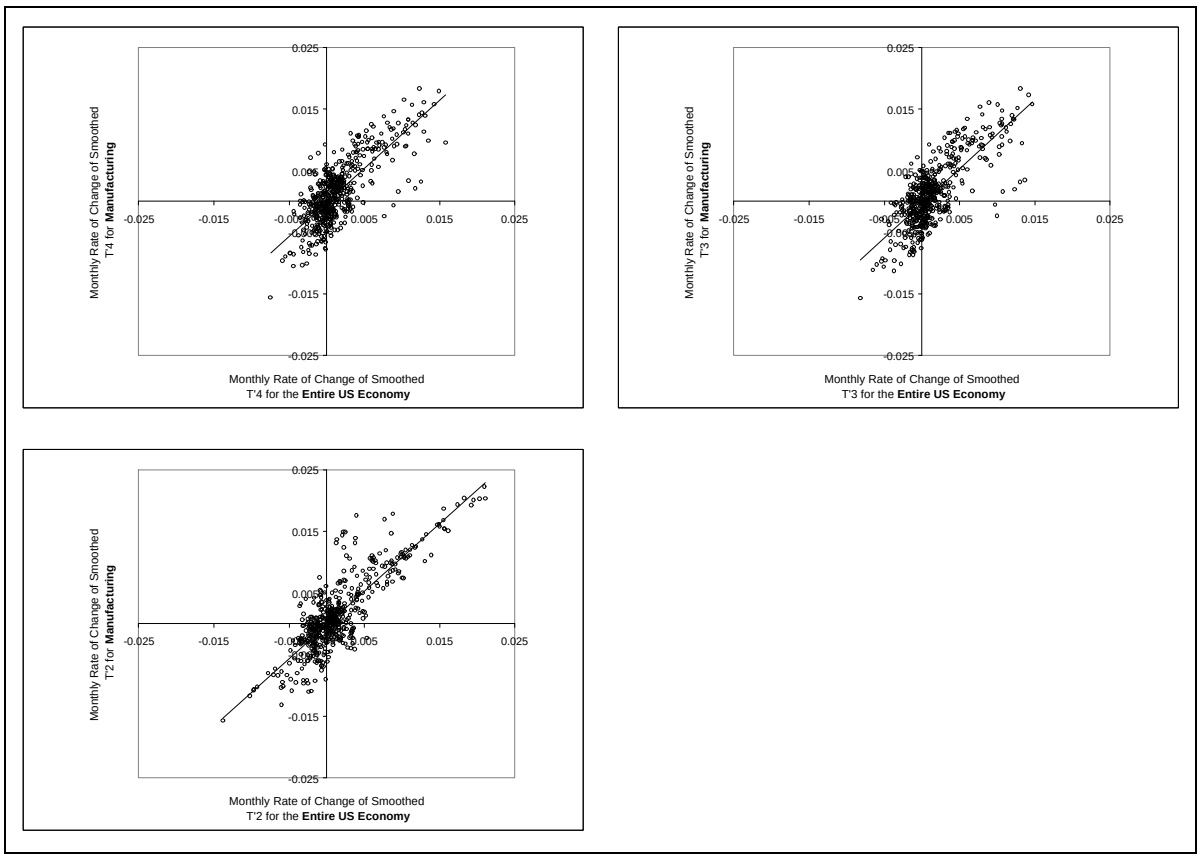

Figure 10- Comparing the Monthly Rates of Change at Different Levels of Aggregation for Manufacturing

In each of the graphs of Figure 10 we are now considering the Theil index at the same level of aggregation, with the monthly changes in the US being represented on the horizontal axis and the monthly changes in manufacturing alone on the vertical axis. The fit to a straight forty-five degree line is not as good as when we consider different levels of aggregation in manufacturing or the US economy as a whole, but the figure still suggests that the dynamics in manufacturing is, on a month by month basis, broadly similar to the dynamics of inequality in the US economy.

Table 5 shows the results of univariate regressions using the US Theil as the dependent variable and the manufacturing Theil as the independent variable.

Table 5- Results of the OLS Univariate Linear Regressions Comparing the Theil for the US Economy with the Theil for Manufacturing

	Intercept		Slope		Adj.
	coeff.	t-stat	coeff.	t-stat	R^2
ΔT'4_US=f(ΔT'4_manuf)	**0.00**	-1.04	**1.10**	27.99	**0.60**
ΔT'3_US=f(ΔT'3_manuf)	**0.00**	-2.17	**1.12**	27.30	**0.59**
ΔT'2_US=f(ΔT'2_manuf)	**0.00**	-1.59	**1.09**	35.26	**0.71**

N=517

The adjusted R-squared are lower than when considering different levels of aggregation, but they are still reasonably high (note that we are considering monthly rates of change). The values for the intercept and the slope are equally reasonable, in the sense that they suggest that the dynamics of inequality in manufacturing and the US economy as whole follow similar patterns.

APPENDIX 1: PROOF OF EXPRESSION [18]

We have to show that, for any g, $1 <= g < l$:

$T'_{g+1} = T'_g + T_g$

Writing explicitly the right hand side of the above equation:

$$T'_g + T_g = T'_g + \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \left({}^{i_1 i_2 \ldots i_g}T^w \right) =$$

introducing the explicit expression for ${}^{i_1 i_2 \ldots i_g}T^w$:

$$ {}^{i_1 i_2 \ldots i_g}T^w = \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g i_{g+1}}}{n_{i_1 i_2 \ldots i_g}} \right) \right]$$

we obtain

$$= T'_g + \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g i_{g+1}}}{n_{i_1 i_2 \ldots i_g}} \right) \right] =$$

the income at level g cancels out, because it can be taken out of the inmost summation

$$= T'_g + \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y} \log\left[\left(\frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g i_{g+1}}}{n_{i_1 i_2 \ldots i_g}} \right) \right] =$$

multiplying and dividing the numerator of the argument of the log by Y and multiplying and dividing the denominator by n, we obtain, after expanding the log:

$$= T'_g + \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y} \log \left[\left(\frac{Y}{Y_{i_1 i_2 \ldots i_g}} \right) \Big/ \left(\frac{n}{n_{i_1 i_2 \ldots i_g}} \right) \right] +$$

$$+ \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} \frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y} \log \left[\left(\frac{Y_{i_1 i_2 \ldots i_g i_{g+1}}}{Y} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g i_{g+1}}}{n} \right) \right] =$$

the second summation is just the between group Theil at level g+1, and since the argument of the log in the first summation does not depend on g+1 it can be taken out of the summation sign at this level:

$$= T'_g - \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{1}{Y} \log \left[\left(\frac{Y_{i_1 i_2 \ldots i_g}}{Y} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g}}{n} \right) \right] \sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} Y_{i_1 i_2 \ldots i_g i_{g+1}} + T'_{g+1} =$$

Since $\sum_{i_{g+1}=1}^{m_{i_1 \cdots i_g}} Y_{i_1 i_2 \ldots i_g i_{g+1}} = Y_{i_1 i_2 \ldots i_g}$ we get:

$$= T'_g - \sum_{i_1=1}^{m} \sum_{i_2=1}^{m_{i_1}} \cdots \sum_{i_g=1}^{m_{i_1 \cdots i_{g-1}}} \frac{Y_{i_1 i_2 \ldots i_g}}{Y} \log \left[\left(\frac{Y_{i_1 i_2 \ldots i_g}}{Y} \right) \Big/ \left(\frac{n_{i_1 i_2 \ldots i_g}}{n} \right) \right] + T'_{g+1} =$$

The remaining summation is just $T¢_g$ and consequently:

$$= T'_g - T'_g + T'_{g+1} =$$
$$= T'_{g+1}$$

Which completes the proof.

APPENDIX 2

Consider a representation of the distribution of income across all grouping structures following the framework of Figure 3 but where now each axis contains the distribution of total income across each grouping structure's groups. For example, in the axis that represents the distribution of income across educational groups, e – the index for educational groups – goes from 1 to E, and each group's income, Y_e, is the "length" of the segment that corresponds to group e. Figure 11 illustrates the distribution of income across educational groupings.

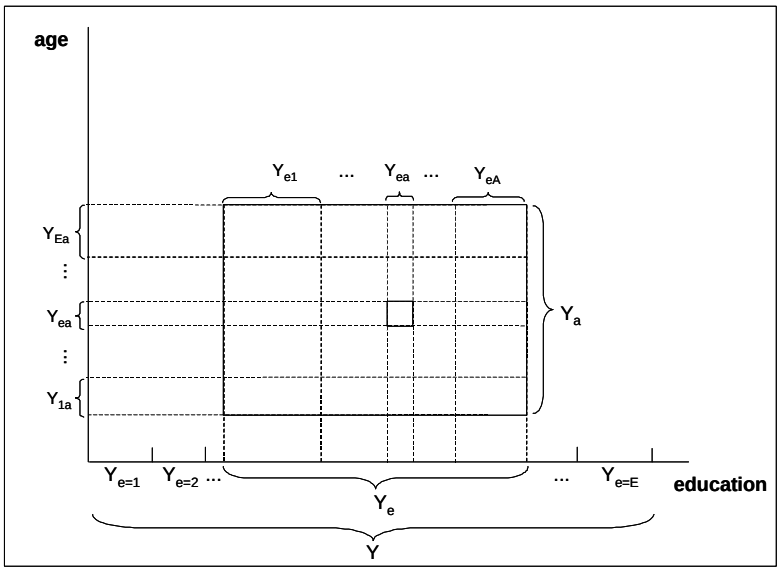

Figure 11- Graphical Representation of the Distribution of Income Across Grouping's Structure

Figure 11 highlights a generic cell (e, a) – sex is omitted to simplify the figure. The length Y_e results from the summation of income across the age groups, where a goes from 1 to A. The length of the segment comprising all the educational groups is Y. Precisely the same occurs with the age grouping structure, so that we obtain a square with area YxY sub-divided in rectangular cells (e, a), each with area Y_exY_a. Each cell (e, a) is also subdivided in smaller rectangles, except for a square that occurs for Y_{ea}xY_{ea}.

In reality, since we have three grouping structures, we are dealing with a cube with volume *Y*x*Y*x*Y*, as represented in Figure 12. This cube results from the juxtaposition of all the cells (*s*, *e*, *a*), each of which has the shape of a parallelepiped (not of a cube necessarily), with volume $Y_s x Y_e x Y_a$. In each cell (*s*, *e*, *a*) Y_{sea} is the only income component that is necessarily the same for all groups, generating a cube (a square in Figure 12).

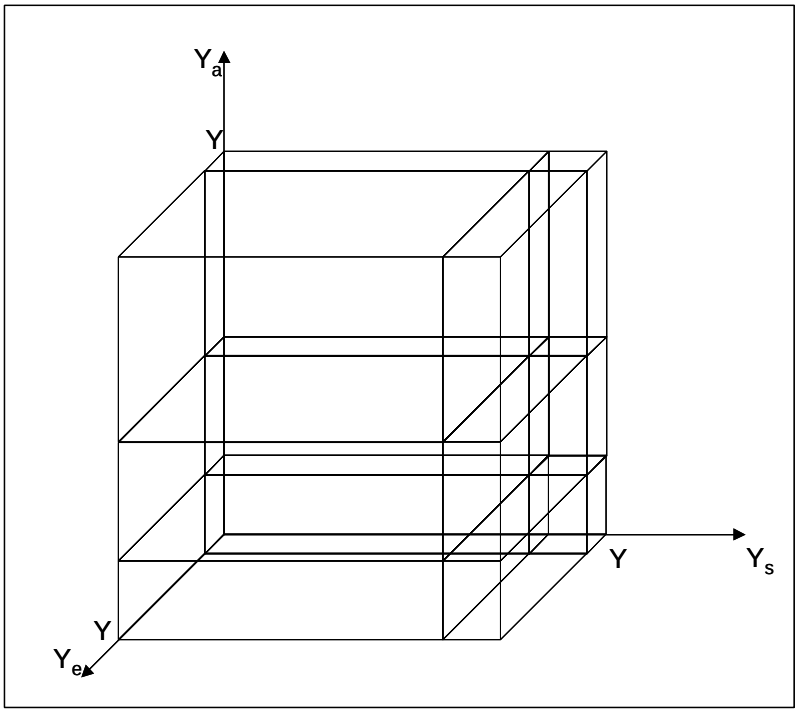

Figure 12- Distribution of Income Across Grouping Structures.

The geometric mean between Y_s, Y_e and Y_a can be interpreted as the length of the sides of a *cube* with the same volume as the parallelepiped $Y_s x Y_e x Y_a$. Therefore, in the expression for the interaction term, Y_{sea} is being compared with the income that would correspond to the hypothetical distribution of income across groups where $Y_s=Y_e=Y_a$. The hypothetical distribution where $Y_s=Y_e=Y_a$ corresponds to the situation where differences across grouping structures are irrelevant, in the sense that cell (*s*, *e*, *a*) is cubic: each group's income is the same regardless of the grouping structure. Precisely the same argument applies to the population ratio between n_{sea} and the geometric mean of n_s, n_e and

n_a. Therefore, the ratio $Y_{sea} \Big/ \sqrt[3]{Y_s Y_e Y_a}$ and the ratio $n_{sea} \Big/ \sqrt[3]{n_s n_e n_a}$ compare the length of the side of the cubes defined by Y_{sea} and by n_{sea} with the side of the hypothetical cubes with areas $Y_s x Y_e x Y_a$ and $n_s x n_e x n_a$.

The cells (*s*, *e*, *a*) for which these two ratios are the same do not contribute to the interaction term[11]. The ratios being the same for a certain cell means that the way in which income and people are distributed across groups in that cell are the same, in the sense that the *income cube* formed by Y_{sea} is in the same proportion to the hypothetical cube with area $Y_s x Y_e x Y_a$ as the population cube formed by n_{sea} is to the hypothetical cube with area $n_s x n_e x n_a$.

[11] Because the argument of the log in the interaction term is one.

REFERENCES

Conceição, P., Galbraith, J. K. (2000), "Constructing Long and Dense Time-Series of Inequality Using the Theil Index", *Eastern Economic Journal*, 26(1): 61-74.

Conceição, P. P. Ferreira (2000), *The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications*, University of Texas Inequality Project Working Paper No. 14; available on the Internet at: http://utip.gov.utexas.edu.

Katz, L., Murphy, K. (1992). "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 106(February): 35-78.

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