REGIONAL INCOME INEQUALITY

IN THE UNITED STATES,

1913-2003

by

Estelle Sommeiller

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics

Fall 2006

REGIONAL INCOME INEQUALITY

IN THE UNITED STATES,

1913-2003

by

Estelle Sommeiller

Approved: ______________________________
Saul D. Hoffman, Ph.D.
Chair of the Department of Economics

Approved: ______________________________
Conrado M. Gempesaw II, Ph.D.
Dean of the College of Business and Economics

Approved: ______________________________
Conrado M. Gempesaw II, Ph.D.
Vice Provost for Academic and International Programs

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

James L. Butkiewicz, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Yves Crozet, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

William R. Latham III, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

René Sandretto, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Thomas Piketty, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ____

Emmanuel Saez, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

ACKNOWLEDGEMENTS

My gratitude and acknowledgments are dedicated to Professor Thomas Piketty's exceptional talent for writing passionating economic books. I would have never undertaken the task of writing about income inequalities without the clarity of his comments and the detail of his appendices in his impressive publication entitled Les hauts revenus en France au XXème siècle. Inégalités et redistributions 1901-1998.[1] Professor Emmanuel Saez also deserves the acknowledgment of my appreciation for lending me support whenever I asked for it.

Warm and grateful thoughts to Professor William Latham III, whose genuine interest for my work rekindled my enthusiasm on a daily basis. The accuracy of his comments, the timeliness of his answers, and his attachment to the values of professionalism made my dissertation work enjoyable and certainly more productive.

Professor James Butkiewicz, Professor Yves Crozet, and Professor René Sandretto also provided me with very valuable comments that greatly affected the organizational structure of this thesis.

---

[1] Suggested translation: France's Top Incomes in the 20th Century. Inequality and Redistributive Issues, 1901-1998.

The financial support from the LET[2], the prestigious research center directed by Yves Crozet, and from the Department of Economics at the University of Delaware allowed me to purchase data and get started with my doctoral dissertation.

From a personal perspective, I would like to thank my surrounding friends who shared with me the ‘concentrate’ of life that my doctoral paper was, with times of satisfaction and frustration, excitement and exhaustion.

Last but not least, I would like to give my warmest regards to my parents, whose humbleness and humility have never lessened the trust and faith they put into my success. I would like to dedicate my work to both of them.

[2] LET stands for ‘Laboratoire d’Economie des Transports’ (Laboratory of Transportation Economics).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The main contribution of this dissertation is the construction of a new homogeneous set of panel data by state cross-sections and annually from 1913 to 2003, using the *Statistics of Income* publications by the U.S. Internal Revenue Service. This database represents the top 10 percent of the income distribution, but data from other sources are needed to account for average income. Meanwhile, the new income database of the top decile offers an alternative estimate of average income figures used by Barro and Sala-i-Martin to study the same topic.

In order to address the issue of income convergence across the United States over the long-run, three types of convergence are distinguished: 1) the β convergence of average income in comparison with the β convergence of the top decile, 2) the σ convergence, and 3) the convergence of top incomes towards the lower decile of the income distribution.

In the case of the β convergence, we found evidence confirming convergence within the top decile, and more mitigated results for convergence among state average incomes. The results showed that income inequality is positively correlated to average income, and negatively correlated to economic growth rates.

In the case of the σ convergence, the trend over time opposes two income groups. On the one hand, average income and income of the top percentile both recorded a decline in dispersion across states (except after the mid 1980s). On the other hand, the incomes of fractiles 90-95, and 95-99 percent featured with a rise in dispersion across states.

Finally, the convergence of the top decile towards the bottom decile is emphasized from 1965 to 2003 (likewise for the top and bottom quartiles). The dispersion indicators of the lower layers of the income distribution were estimated by generalizing the Pareto assumption to the full income distribution. What emerges from these estimates is that the top income shares did not grow faster than the low income shares from 1965 to 1984. This trend was totally reversed and reached a peak in 1988 in all states, then decreased again until 1985, and finally reached a local maximum in 2000.

Chapter 1

INTRODUCTION

Income inequality can be – and is – conceived of in many ways. The topic has ignited passionate debates in the academic literature, crossing the borders of economics, sociology, politics, law, history, and geography, just to name a few. While this study will not attempt to shed light on all disciplinary aspects of the income inequality issue, both geographical and historical facets of income inequality within the United States will lie at the core of the analysis.

Before proceeding further, the conceptual framework needs to be set properly, and to this we turn first.

1.1 Income Inequality: Meaning and Measurement

1.1.1 How is Income Defined?

First, income has to be comparable from one perspective or another: over time, across geographic space, or among households, just to name a few. In other terms, income needs to be expressed in one single unit of measurement, whether it refers to the choice of a particular currency for international comparisons, the base year of a price index for temporal comparisons, or the percentage of a total if one considers income shares. Furthermore, the unit of account may be selected carefully. Champernowne and Cowell (1998, p. 68) illustrate the point with the following example.

> We cannot assume that an income of x_1 going to 'person' 1 always counts as less than income of x_2 going to 'person' 2 just because x_2 is a larger number than x_1, since 'person' 1 might be a single individual and 'person' 2 a sprawling household of three adults and ten children, but incomes must be defined in such a way that we can always decide for any pair which is the larger or that they are equally large, even when the two 'persons' receiving them are demographically dissimilar: that is what is meant by the income being comparable.

Second, one has to be clear about the income sources counted in the definition of income, as it varies from one statistical agency to the other, or from one year to another within the same agency. Income usually includes wages, business and property incomes, dividends, interest, and social benefits. But official income may exclude amounts that contribute to the individual's overall living standard, such as transfer payments, non-cash income (the use of company car, subsidized meal vouchers, food stamps), and so on. In other words, the income definition needs to be clearly stated. The income definition used here is discussed in section 3.3.2 of this thesis.

1.1.2 Income Recipients: Who is to Count?

Individuals, households, families, tax units… The choice of the basic unit for the income recipients may not be as straightforward as it appears. Based on the same income tax returns for their inquiry on inequality in the United States, Kuznets (1953), and Piketty and Saez (2004) did not choose the same unit. Kuznets reduced tax returns to a per capita basis, while Piketty and Saez debated Kuznets' choice and created instead a tax units series for homogeneity purposes. What are the arguments from both sides? On the one hand, Kuznets (1953, p. xxxiii) argued that a tax return "does not represent the number of income recipients, since there may be more than one recipient per return (and the number cannot be ascertained from the available data). (…) Nor does the income tax return measure, in and of itself, a family or spending unit, however defined, since a family may file more than one return". On the other hand, Piketty and Saez (2003, p. 4) pointed out a downward bias that the per capita basis introduces in the income series, and prefer to use tax units.[1] "Because our data are based on tax returns, they do not provide information on the distribution of individual incomes within a tax unit. As a result, all our series are for tax units and not individuals. A tax unit is defined as a married couple living together (with dependents) or a single adult (with dependents), as in the current tax law."

1.1.3 Income Inequality Measures

Income inequality can be measured in *absolute* terms. The poverty line, for instance, is defined as the income threshold below which an individual or a family is considered poor. Inequality needs to be measured in *relative* terms whenever two income distributions are to be compared. There are many relative measures of inequality. In the particular case of poverty, a relative measure may be the percentage of the population below the poverty line. There are many other inequality indicators expressed in relative

[1] Section 3.3.3 will discuss the details of this issue.

terms, such as percentile distributions, the Gini coefficient, the Theil, Atkinson, and Gibrat indexes, etc.

However, none of these inequality indicators informs about the causes of income inequality. Multiple factors may generate income inequality: life cycle effects (age), gender or race discrimination, education and training, inherited wealth, economic circumstances, individuals' characteristics (IQ, talent), the leisure-labor choice, and so on. This critical remark does not mean that inequality measures are invalid; rather, it points out the need for clarification of definitions and consistency of comparisons prior to investigating the causal aspects of the subject.

1.2 A Double-Contradiction Rekindles the Debate

The starting point of this study is a contradiction that emerges from two currents of the literature: the income inequality literature, and the regional convergence literature. On income inequality, the natural decrease of inequality predicted by Kuznets (1955) does not receive strong empirical support. On regional convergence, neo-classical economists support convergence whereas geographers marshal evidence for divergence, at least in the case of the United States at the sub-national level.

This thesis sheds light on the subject through the construction and use of a new panel data set of 4,641 income inequality observations (the 50 U.S. states plus the District of Columbia over 91 years from 1913 to 2003). Various inequality indicators are derived from this database. At the national level, the time-series data here are very similar to those of Piketty and Saez (2004). At the state level, there is no possible comparison with Piketty and Saez's series as their data set does not allow for sub-national differentiation. Many empirical studies used panel data on income inequality within the United States, but none, to my knowledge, covers such a comprehensive range of years for each state.

1.3 The Analyses of Convergence

The panel data on top incomes is used here to address the question of income convergence in the United States over the past century. Did the richest states experience relatively slower rates of growth while the poorest states recorded faster rates of growth, with both ends eventually converging to a common average? Does the same conclusion hold when considering top incomes instead of average incomes? If true, how does it impact inequality across states? Furthermore, convergence can be approached in terms of income dispersion. What conclusions can be drawn from the comparison between the dispersion of average income and the dispersion of top income? Finally, another aspect of convergence raises the issue of the estimation of the income distribution itself, as it has been only partially known so far. Suppose that we can derive, one way or another, all deciles of the spectrum. In that case, we would be able to address the following question. To which extent does top decile converge towards or diverge away from the lower deciles of the income distribution? How did one evolve with respect to the other over time?

1.4 Organizational Structure

The remainder of this study is organized as follows. Chapter 2 reviews the existing literature and explains how it relates to the subject of this thesis. Chapter 3 presents the raw data as they appear in the *Statistics of Income* tables, and describes the methodology used to derive the fractile tables and all other inequality measures. Cross-sectional and trend-over-time analyses are presented in Chapter 4. Chapter 5 starts with an overview of β convergence, continues with the estimation of growth and inequality equations, and ends with an analysis on income dispersion. Finally, Chapter 6 recognizes the need for studying more than the top 10 percent of the income distribution, and estimates the lower fractiles (down to the lower decile) based on one assumption about the functional form of the Lorenz curve. The Gini coefficients, and two other inequality measures ('top-to-bottom' ratio and an intermediary indicator) are also estimated and briefly discussed. Chapter 7 concludes the study with a summary and suggestions for future research.

Chapter 2

LITERATURE SURVEY:

THE NEED FOR DISPERSION INDICATORS

Theoretically, the Kuznets' curve and the neo-classical model of growth are perfectly consistent. Empirically, the ambiguity of conclusions dominates the debate, and the sections below briefly illustrate where we stand.

2.1 On Income Inequality: Inverted-U or Sinusoidal Curve?

2.1.1 Kuznets and the Natural Decrease of Income Inequality

Kuznets (1953) pioneered the development of empirical studies based on income tax data sorted by income ranges, which is a common way to highlight income differences. Kuznets used individual income tax returns to derive top income shares series (top 1%, top 3%, top 5% and their intermediate percentage bands) from 1913 to 1948, but the series is fairly complete only for 1919-1946. In percent, the shares of upper income groups are defined as the ratio of income amounts reported by each group over total income. Individual tax returns, despite the term 'individual', do not necessarily reflect the income of individuals, rather the income of households filing a tax return. For consistency purposes, Kuznets decided to estimate series based on individuals, not tax units. To do so, Kuznets divided the total income reported in each income bracket by the total number of individuals represented by all tax returns in that bracket. For instance, a tax return of a widow with no dependents reporting $10,000 would be replaced by an individual with $10,000 of income while a family of four with $10,000 of income would be replaced by four identical individuals with $2,500 of income each.[1]

Kuznets (1953) was restricted to the top tail of the distribution because taxes were paid only by the upper income classes of the population. Another data limitation is that the size of income shares depends upon the unit used in the distribution (individual, family, consuming unit, etc.) Under these circumstances, Kuznets (1953, p. xxxv) made three major findings: 1) The income shares of upper income groups drastically declined between 1939 and 1950; 2) this decline was highly correlated with the drop in the shares of upper income groups in total savings of individuals; and 3) short term changes in these shares were associated with business cycles. Note that the declining trend of upper shares in the United States at that time serves as a prelude to the famous inverted-U curve.

What follows briefly explains what the Kuznets' curve is. Using the long-term perspective of development stages, Kuznets (1955) uses the partitioning of employment (among the usual three sectors of agriculture, manufacturing, and services activities) in order to explain how the development process first widens and then narrows income inequalities over time. Theoretically, as the industrial revolution triggers labor to

---

[1] For more details about this example, see Piketty and Saez (2003, p. 37).

migrate from rural areas to more urbanized and industrialized areas, the first stages of economic development result in a widening of the income inequality gap, as described by the upward-sloping portion of the Kuznets' curve. At the later stages of development, the industrial gains concentrated in urban areas are eventually shared among more people (decrease in inequality) as the flow of workers moves from rural areas of low-productivity levels into the industrialized metropolis, hence the downward sloping curve.

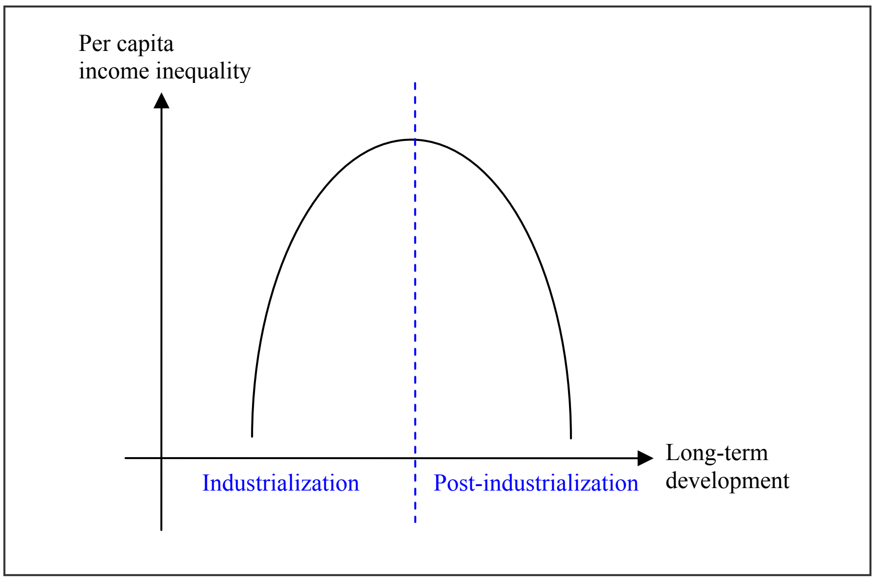


Figure 2.1 Kuznets' Curve

Kuznets (1958) investigated the parallels between the distribution of income within the United States (for 1919-21, 1929, 1940, 1950 and 1955) and the distribution of income across countries. For each year, the 48 states are divided in 6 groups of 8, sorted by the descending order of per capita personal income as defined by the Department of

Commerce[2]. Kuznets also calculated an annual inequality indicator that divides the state-group averages over the national average. In general, Kuznets' conclusions for U.S. states are very similar to those drawn for international comparisons, although there are two key differences. The first shows that the level of per capita income inequality was narrower in the United States than that among other nations. The second finds larger amplitude in the sectoral composition in the interstate comparisons than in the international comparisons.

When sub-national regions of the U.S. economy are considered, one should expect them to lie on the decreasing portion of the Kuznets curve, so that the low-income states are associated with greater income disparities than the more developed states. Perloff, Dunn, Lampard and Muth (1960, p. 522) illustrate this point for 1955: "the sixteen states with the lowest per capita income derived some 5.9% of their total manufacturing wages and salaries from 'machinery (including electrical)'; the sixteen states with middle level income, 16.3%; the sixteen states with the highest per capita incomes, 20.1%."

Similarly, Williamson (1965) collected U.S. county-level data for 1950 and 1960 (p. 19 and 20), and found that:

> On the average, the eight lowest income states have a coefficient of inter-county inequality approximately two and one-half times that of the richest seven. The same pattern holds true for the 1960 data, where again severe interregional differentials are associated with relatively low levels of development.

---

[2] In Kuznets' time, the Department of Commerce defined personal income as the sum of wages, salaries, and other labor income; entrepreneurial income; dividends, interest, rent and royalties; and transfer income; less personal contributions for social insurance – or an approximation to it in the case of the Leven estimates for the earlier years.

Overall, the inverse-U curve finds some empirical support more at the sub-national than at the international level[3]. But the sub-national studies usually suffer from a lack of data over the long-run, thereby weakening the conclusions for convergence.

Kuznets' expectations of natural regional convergence are consistent with the neoclassical steady state of conditional convergence[4] across regions, assuming the absence of factors hindering the flows of capital and labor from one place to the other.

2.1.2 Piketty: the Accidental Pattern of Income Inequality

Piketty (2001)[5] uses individual income tax returns (1915–98), wage tax returns (1919–98) and inheritance tax returns (1902–94) to create a homogeneous data set of income inequality, wage inequality and wealth inequality. His findings question the arguments supporting the belief of a natural fading away of income inequality over time. "Mostly accidental" is how Piketty qualifies the decline in income inequality that took place during the first half of the twentieth century in France. The overall decline is related to the historical hazards of capital income and progressive taxation on very large fortunes.

---

[3] See Barro (2000, p. 9).

[4] Absolute convergence refers to the tendency for the living standards of *different* countries to become more equal over time, independently of the particular characteristics of individual countries. Conditional convergence features the tendency of living standards within groups of countries with *similar* characteristics to become more equal over time.

[5] The original version is the following: Piketty Thomas. Les hauts revenus en France au XXème siècle. Inégalités et redistributions 1901-1998. Grasset. 2001.

Piketty's work (2001) triggered a series of other empirical analyses applying the same methodology to different countries, as summarized by Atkinson (August 2003):

> There has recently been a revival of interest among economists in the distribution of top incomes. The pioneering study by Piketty (2001) produced estimates of the long-run distribution of top incomes for France. Following his lead, Atkinson (2002) made estimates for the United Kingdom, and Piketty and Saez (2003) made estimates for the United States. Estimates are now also available for Germany (Dell, 2002), Canada (Saez and Veall, 2003), Netherlands (Atkinson and Salverda, 2003), India (Banerjee and Piketty, 2001), and Australia and New Zealand (Atkinson and Leigh, 2003). In using data from the income tax records, these studies use similar sources to the earlier work of Bowley (1914) and Stamp (1916 and 1936) in the UK, and Kuznets (1953) in the US. (See also the survey by Kraus, 1981.) The findings of the recent papers are however of added interest, since the data provide estimates covering nearly all of the twentieth century – a length of time series unusual in economics. Moreover, the techniques are considerably more developed.

Thomas Piketty and Emmanuel Saez (2004) shed light on the empirical aspects of income inequality in the United States, using the Internal Revenue Service's (IRS) annual publications on *Statistics of Income* to create a homogeneous time-series data set on the upper shares of income and wages from 1913 to 1998. The methodology used by Piketty and Saez follows very closely that used by Kuznets (1953). There are several differences, though. The first difference lies on the definition of the upper income shares, based on individuals for Kuznets, and tax units for Piketty and Saez. In their paper (long version, 2004, p. 4), a "tax unit is defined as a married couple living together (with dependents) or a single adult (with dependents), as in the current tax law." Because "Kuznets did not correct for the re-ranking," he "misclassified in the top shares large families with high total income but moderate income per capita."[6] Relative to Saez and Piketty's, Kuznets' shares are therefore underestimated. This discrepancy does not impact the pattern of upper income evolution over years though, because the decrease in the number of individuals per tax unit occurred across all income groups over the century.

The second difference deals with the treatment of capital gains and numerous other data adjustments. They received much less attention in Kuznets' work (1953) than in Piketty and Saez (2004). One should be aware, however, that the IRS micro-files released from 1960 to 1995, and improving the statistics of the initial IRS tables, significantly helped Saez and Piketty in making the corrections to their estimates, were not at the disposal of Kuznets in his time.

The evidence for rising inequalities in the U.S. since the 1970s leads the authors to re-interpret the Kuznets curve as a sinusoidal graph. "A new industrial revolution has taken place, thereby leading to increasing inequality, and inequality will decline again at some point, as more and more workers benefit from the new innovations." This argument matters even more once applied to the state-level analysis as the technological revolution occurred first on both the east and the west coasts. The lack of empirical evidence supporting the theoretical reduction in inequalities over time also holds in the convergence literature.

---

[6] Piketty and Saez (2003, p. 37)

2.2 On Regional Convergence: Two Disciplines Compared

Two academic disciplines, (neoclassical) economics and (human) geography, predict two opposite outcomes: regional convergence for the former school and regional divergence for the latter. On the one hand, the Solow model suggests that two economies with similar technology, savings rates, and population growth rates should converge to the same capital-labor ratio and level of per-capita income in the long run. This model finds empirical support in numerous studies such as Barro and Sala-i-Martin's (1995). Barro and Sala-i-Martin applied the neoclassical model to the U.S. states. "The overall evidence weighs heavily in favor of convergence: poor states tend to grow faster in terms of per capita income and product, and within sectors as well as for state aggregates." In that sense, the growth literature validates the income inequality debate begun by Kuznets (1955).

On the other hand, the new economic geography, similarly concerned with the understanding of the origins and effects of regional growth differences, levels criticism at the way neo-classical theory handles the analysis of income inequality at the sub national scale: "neoclassical theory is based on diminishing returns in which any spatial structure (such as the creation of rich and poor regions) is self-canceling (through convergence)". The economists of the new economic geography attempted to bridge the gap at the theoretical level, substituting the assumption of diminishing returns to that of increasing returns to scale (Krugman 1995, 1996, 1998). With those recent improvements, the opposition against regional convergence persists, as the researchers from both disciplines marshal empirical evidence supporting the increase in regional inequality in the United States from the late 1970s on, and conclude in favor of economic polarization and regional divergence (Danziger and Gottschalk, 1993; Amos, 1988; Braun 1991, Coughlin and Mandelbaum, 1988; Rowley, Redman and Angle, 1991, Fan and Casetti 1994).

The next section looks more closely at the arguments debated in both disciplines.

2.2.1 Barro and Sala-i-Martin Advocate β Convergence[7]

Acknowledged as a key reference in empirical literature, the contribution of Barro and Sala-i-Martin (1995) strongly supports the convergence prediction of the Solow model with exogenous technological progress in a closed economy. Barro and Sala-i-Martin used both gross state product and per capita personal income[8], exclusive of

---

[7] Barro and Sala-i-Martin (1995, p. 383) distinguish two definitions of convergence: β-convergence and σ-convergence. "In one view (…), convergence applies if a poor economy tends to grow faster than a rich one, so that the poor country tends to catch up with the rich one in terms of the level of per capita income or product. This property corresponds to our concept of *β convergence*. (This phenomenon is sometimes described as "regression toward the mean".) The second concept (…) concerns cross-sectional dispersion. In this context, convergence occurs if the dispersion _measured, for example, by the standard deviation of the logarithm of per capita income or product across a group of countries or regions_ declines over time. We call this process *σ convergence*."

[8] The BEA's definition is: "Personal income is the income that is received by persons from all sources. It is calculated as the sum of wage and salary disbursements, supplements to wages and salaries, proprietors' income with inventory valuation and capital consumption adjustments, rental income of persons with capital consumption adjustment, personal dividend income, personal interest income, and personal current transfer receipts, less contributions for government social insurance. This measure of income is calculated as the personal income of the residents of a given area divided by the resident population of the area. In computing per capita personal income, BEA uses the Census Bureau's annual midyear population estimates."

all transfers, for 47 U.S. states or territories from 1880 to 1988[9], and found a β convergence rate of about 2 percent a year. According to the authors, factors of production indeed flowed from low- to high-income states, leading the economies further below the steady-state position to grow faster. The evidence of convergence within the United States supports more the neoclassical model of growth than the evidence of convergence over large samples of many countries. The main reason is that one country's features usually contrast more with another country's than regions within the same country.[10]

Note that the β convergence, here tested within the United States, does not imply convergence across states, as Olivier Blanchard[11] pointed out. Indeed, cross-sectional means that fluctuate within an interval shrinking over time can still display increasing standard-deviations. Blanchard highlighted this nuance to better argue against it. He showed that the temporal dimension of Barro and Sala-i-Martin's data can be used to test the hypothesis of convergence in broader terms. Considering 1) dispersion over time, and 2) convergence coefficient over time, "both [ways] strongly suggest convergence [across the U.S. states]". In other words, Blanchard generalized Barro and

---

[9] The authors exclude Oklahoma (no data), Alaska, Hawaii, and the District of Columbia. The BEA Personal State Income data they used are available for 1880, 1900, 1920, and annually from 1929 to 1988.

[10] Barro and Sala-i-Martin (1995) also studied convergence between regions in Europe and in Japan.

[11] See "Comments and Discussion" in Barro, Sala-i-Martin, Blanchard & Hall (1991, p. 159-174).

Sala-i-Martin's conclusions on β convergence to overall convergence across U.S. states. Such results are now taken for granted in the economic literature, and this is exactly what I think can be improved, and this for two reasons.

First, Barro and Sala-i-Martin tested the neo-classical equation using arithmetic means as a proxy for the distribution of income. How to measure income discrepancies is an issue debated carefully in the income inequality literature, and the arithmetic mean hardly belongs to the set of the best statistical tools. A response to this problem is to resort to dispersion indicators, such a quartiles, deciles, or percentiles, revealing, at least partially, the income distribution itself.

Second, the later trends of increasing income inequalities observed recently in the United States question the neoclassical theory in that respect. The issue is controversial also with regard to the decade when income inequality started to widen: the mid-1970s, versus the mid-1980s. Such an argument carries much weight in the academic circles of human geography.

2.2.2. The Geographers Advocate Regional Polarization

The human geography literature easily gets to the point. Braun (1991) published an article entitled "Income Inequality and Economic Development: Geographic Divergence", Amos (1988) writes about "Unbalanced regional growth and regional income inequality in the latter stages of development", Coughlin and Mandelbaum (1988) "Why have state per capita income diverge recently?", and Fan and Casetti (1994) "The spatial and temporal dynamics of U.S. regional income inequality, 1950-1989."

Using the state-level per capita income data published by the Bureau of the Census, Fan and Casetti (1994) shed light on the short-term dynamics of regional inequalities in the United States between 1950 and 1989. In their 1994 publication, they acknowledge the influence of Kuznets' convergence on the empirical literature. However, they consider the explanation of economic development as "static and deterministic:" The model they suggest lays emphasis on the regional dynamics accounting for the spatial flows of factors of production, that in turn affect the patterns of regional income inequality. The authors stress three successive phases: the "polarization" of factors of production, or regional divergence; the reverse process ("polarization reversal"), or regional convergence; finally, the re-allocation of production factors flowing across the United States and the renewed growth of some traditional core states, such as the recent increase in regional income inequality.

2.3 New State-Level Income Data

The data set constructed here will serve the analyses of convergence in later chapters.

2.3.1 A New Data Set Relates to the Literature

The table below summarizes the inequality and convergence literatures and shows the need for new data at the sub-national level over the long-run.

Table 2.1 A New Data Set to Debate on Inequality and Convergence

Authors studying the U.S.	National Level (data & conclusion)	State Level (data & conclusion)
Kuznets	IRS data & top decile (1953) inverted-U curve (1955)	OBE[12] personal income (1958) convergence (1958)
Piketty and Saez	IRS data & top decile (2004) sinusoidal Kuznets' curve (2004)	---
Barro & Sala-i-Martin	Real GDP (1995) convergence (1995)	BEA personal income (1995) state convergence (1995)
Geographers	Real GDP divergence	BEA or Census income data polarization
Estelle Sommeiller	IRS data & top decile (2006) sinusoidal Kuznets' curve (2006)	IRS and BEA data + estimates of full income distribution convergence and divergence

2.3.2. The New Data Set is Used for Convergence Analyses

The analysis based on the panel data set is twofold. First, the data relative to the upper decile yield measures of inequality between and within states. These inequality indicators and the growth rates of state *i*'s average income are used in regressions to study the β and σ convergence (see Chapter 5). Second, the remaining 90 percent of the

[12] OBE stands for 'Office of Business Economics', and is the former name of the BEA, 'Bureau of Economic Analysis'. Both were/are part of the U.S. Department of Commerce.

distribution is not ignored as lower measures (bottom decile, first 25 percent, median, first 75 percent, and upper decile) are estimated to analyze the convergence (or divergence) of income groups within the income distribution itself.

Chapter 3

DATA AND METHODOLOGY

This chapter describes 1) the set of panel data, 2) the methodology applied, and 3) the inequality indicators that result from the calculations.

3.1 Data: the IRS Publications *Statistics of Income*

The Internal Revenue Service publication, *Statistics of Income*, *Individual Income Tax Returns*, is the primary source of the data used for the 50 states and the District of Columbia, from 1913 to 2003. Two variables are extracted from the IRS tables: the number of individual returns and the total income expressed in current dollars. Both variables are ranked by size of income (i.e. for a certain number of income brackets), and by state. For instance, the first page of the 1917 IRS table appears as:

Table 3.1 An Example of the IRS Tables

TABLE 2a.—PERSONAL RETURNS—DISTRIBUTION OF INCOMES, BY CLASSES; *showing for each class of income the number of returns, net income, contributions, dividends, tax, and relative percentage, by States.*

[Income returned for the calendar year ended Dec. 31, 1917.]

ALABAMA.

Income class.	Number of returns.	Total net income.	Exemptions from normal tax.		Normal tax.	Surtax.	War excess profits tax.		Total tax.	Average rate of tax.
			Contributions.	Dividends.			Personal service (8 per cent).	Invested capital (graded).		
										Per cent.
$2,000–$2,500 [1]	2,157	$4,727,045	$36,111	$144,716						
$2,000–$2,500	2,124	4,824,395	234,837	166,782	$18,700		$81		$18,781	0.39
$2,500–$3,000 [1]	503	1,337,422	18,727	114,108						
$2,500–$3,000	1,941	5,299,612	57,630	77,589	24,086				24,086	.45
$3,000–$4,000 [1]	173	577,359	11,976	215,776						
$3,000–$4,000	2,309	7,931,616	108,263	378,752	49,464		773	$400	50,637	.64
$4,000–$5,000 [1]	60	259,394	5,339	168,923						
$4,000–$5,000	1,186	5,270,265	88,857	243,002	16,629		146		16,775	.32
$5,000–$6,000 [1]	6	32,613	2,166	23,472						
$5,000–$6,000	653	3,561,945	58,671	285,612	47,045	$2,066	1,488		50,599	1.42
$6,000–$7,000	398	2,573,101	42,730	314,408	12,342	5,155	12,236	141	29,874	1.16
$7,000–$8,000	270	2,024,757	36,791	310,298	8,399	5,970	8,373		22,742	1.12
$8,000–$9,000	203	1,721,870	30,368	280,923	6,686	7,298	6,758	1,047	21,789	1.27
$9,000–$10,000	169	1,600,450	25,087	336,101	10,543	8,873	8,253		27,669	1.73
$10,000–$11,000	131	1,375,096	23,557	215,932	12,504	10,082	8,813	1,634	33,033	2.40
$11,000–$12,000	91	1,039,530	29,935	168,382	23,127	8,658	11,592	523	43,900	4.22
$12,000–$13,000	77	959,740	17,858	234,201	20,309	9,869	8,956	52	39,186	4.08
$13,000–$14,000	60	797,223	12,065	205,959	16,697	9,937	6,132	149	32,915	4.13
$14,000–$15,000	56	813,473	15,363	198,020	9,219	9,655	8,157		27,031	3.32
$15,000–$20,000	157	2,719,868	45,506	586,444	8,870	46,499	32,296	5,182	92,847	3.41
$20,000–$25,000	80	1,748,518	32,997	524,974	32,619	40,013	16,897	10,943	100,472	5.74
$25,000–$30,000	41	1,091,135	17,309	404,969	20,066	33,319	13,395	4,542	71,322	6.53
$30,000–$40,000	57	1,933,465	39,870	644,770	12,960	64,637	34,529	20,843	132,969	6.88
$40,000–$50,000	35	1,545,044	25,630	660,389	30,304	68,698	19,377	9,523	127,902	8.27
$50,000–$60,000	16	890,477	31,507	305,175	17,358	48,532	13,988	12,060	91,938	10.32
$60,000–$70,000	9	588,431	13,909	229,029	12,991	36,684	3,229		52,904	8.99
$70,000–$80,000	3	218,332	3,445	13,715	5,557	22,095	2,304	24,823	54,779	25.09
$80,000–$90,000	1	(2)	(2)	(2)	(2)	(2)			12,353	14.17
$90,000–$100,000	5	(2)	(2)	(2)	(2)	(2)	(2)	(2)	62,955	13.06
$100,000–$150,000	1	(2)	(2)	(2)	(2)	(2)	(2)		19,416	18.74
$150,000–$200,000										
$200,000–$250,000	3	(2)	(2)	(2)	(2)	(2)	(2)		148,451	22.12
$250,000–$300,000	1	(2)		(2)	(2)	(2)	(2)		54,840	18.69
$300,000–$400,000										
$400,000–$500,000	1	(2)	(2)	(2)	(2)	(2)	(2)		148,650	36.31
$500,000–$750,000	1	(2)	(2)		(2)	(2)	(2)		325,396	46.47
$750,000–$1,000,000										
$1,000,000–$1,500,000										
$1,500,000–$2,000,000										
$2,000,000–$3,000,000										
$3,000,000–$4,000,000										
$4,000,000–$5,000,000										
5,000,000 and over										
Classes grouped [2]		2,747,386	17,711	649,292	72,385	592,289	104,500	2,887		
Grand total	12,978	60,209,562	1,084,215	8,101,713	488,860	1,030,329	322,273	91,749	1,936,211	3.22

Typically, the state income tables annually released by the IRS display, for each state and for a series of income brackets, the number of tax returns, the dollar amount of adjusted gross income, and some sources of income. While the IRS tables have been available in an electronic format since 1997, the tables published earlier are paper-based. All paper-based tables were scanned and tabulated in Excel worksheets. They record the first two variables (number of returns, and income amounts).

Arraying these income classes from the lowest income to the highest, it is possible to derive cumulative totals of household population and income, and to draw partition lines splitting the distribution evenly in a series of fractiles. Fractiles divide a given set of observations into equal portions, e.g. into two halves with the median, into four quarters with quartiles (0-25 percent, 25-50 percent, 50-75 percent, and 75-100 percent), into ten equal shares with deciles (0-10 percent, 10-20 percent, etc.), one hundred shares with percentiles, etc.

What is called ‘top decile’ refers to the income group that represents the richest 10 percent of the population. Only the richest 10 percent of the population can be examined consistently from 1913 to 2003. Why is it relevant to reduce the whole distribution to the top decile only? Before 1944, large exemption levels reduced the proportion of individual filing returns to about 10-15 percent (versus the vast majority after WWII). Between 1913 and 1916, this share drops as low as 1 percent.

3.2 Methodology: Pareto Estimation of Income Levels in Upper Decile

Piketty (2001) pointed out the strong heterogeneity of the top decile. Piketty's methodology is applied here, and is fully described in Les hauts revenus en France au XXème siècle[1]. Piketty's book deals with the upper income shares over the past century, and focuses on France's individual tax returns in the 20th century. It includes voluminous appendices detailing step by step the 'why's and 'why not's of the methodology used. The methodology can be broken into two steps: 1) the definition of the fractile tables that need to be created for each state and for each year between 1913 and 2003, and 2) the interpolation method used to estimate each of them.

3.2.1 17 Fractiles to be Estimated

There are 17 fractiles to be estimated: 6 threshold incomes, 6 average incomes, and 5 inter-fractiles. The following table describes what they correspond to.

[1] The Upper Shares of Income in France in the 20th Century.

Table 3.2 Definition and Notation of the 17 Fractiles to be Estimated

6 Threshold Incomes						
Percentile of tax units by income level from low to high	**90%**	**95%**	**99%**	**99.50%**	**99.90%**	**99.99%**
Threshold income required to reach each percentile	$74,838	$98,292	$196,683	$285,144	$666,101	$2,218,587
Threshold income required to reach each percentile - Notation	**TI90**	**TI95**	**TI99**	**TI99.50**	**TI99.90**	**TI99.99**

6 Average Incomes						
Percentile of tax units by income level from low to high	**90-100%**	**95-100%**	**99-100%**	**99.50-100%**	**99.90-100%**	**99.99-100%**
Cumulative percentage of tax units from top down	10%	5%	1%	0.5%	0.1%	0.01%
Average income of tax units in each cumulative range	$137,646	$189,694	$434,720	$633,256	$1,504,101	$4,961,690
Average income of tax units in each cumulative range - Notation	**AI90-100**	**AI95-100**	**AI99-100**	**AI99.50-100**	**AI99.90-100**	**AI99.99-100**

5 Inter-Fractiles						
Percentile of tax units by income level from low to high	**90-95%**	**95-99%**	**99-99.50%**	**99.50-99.90%**	**99.9-99.99%**	**99.99-100%**
Percentage of tax units within each interval	5%	4%	0.50%	0.40%	0.09%	0.01%
Average income of tax units in each interval	$85,599	$128,437	$236,185	$415,544	$1,119,925	$4,961,690
Average income of tax units in each interval - Notation	**AI90-95**	**AI95-99**	**AI99.00-99.50**	**AI99.50-99.90**	**AI99.90-99.99**	**AI99.99-100**

The threshold income labeled TI90 refers to the income level above which a tax unit belongs to the top decile and below which it does not. Six income thresholds were calculated: TI90, TI95, TI99, TI99.5, TI99.9, and TI99.99.[2] All of the income

[2] The TI notation here corresponds to the P notation in Piketty's publications.

amounts are expressed in 2003 dollars, yet they do not take into account the cost of living differentials from one state to another.[3]

The dollar amount in the AI90-100 category refers to the average income of the top decile. For instance, the richest 10 percent of Alabama's tax payers earned about $137,646 on average in 2003 (current dollars).[4]

This upper income class displays enough disparities to subdivide the top decile into 6 inter-fractiles: AI90-95, AI95-99, AI99-99.5, AI99.5-99.9, AI99.90-99.99, and AI99.99-100[5]. Still considering the case of Alabama in 2003, the richest 0.01 percent earned about $10,128,674 that year. Such a high income could not be revealed by an income of $137,646 for the AI90-100 fractile.

The example of year 1940 reveals how sharp the differences can be from one state to another. For the households living in Arkansas that year, the threshold income TI90 is $9,354, while the threshold income of TI99.99 is $474,840. Meanwhile, $11,546,827 is the threshold of the richest 0.01 percent in Delaware. The same year, an income of about $47,000 makes a tax filer part of the wealthiest 1 percent in Mississippi, but part of the wealthiest 10 percent in the District of Columbia.

---

[3] It would be very difficult to compute cost of living differences both across states and over time. State-level CPI data have the same base year value of 100, so these data do not reflect cost of living differences in the base year, only relative rates of inflation beginning from the same base value.

[4] 'AI90-100' here is equivalent to Piketty's average income accruing to fractile 'P90-100'.

[5] As Atkinson pointed it out, the top fractile AI99.99-100 needs to be deflated due to the lack of income brackets at the top. Therefore, a 5% decrease was applied to the years displaying 12 income classes or less, i.e. between 1988 and 2002.

In the whole sample (except Alaska), the District of Columbia has the highest TI90 threshold from 1917 to 1942, and again from 1944 to 1947, 1953-55, 1958, (except in 1918 with a third position after South Dakota and Nebraska, and in 1925 ranking second after Florida).

The year 2000 is an illustration of the peak in income levels reached by the top percentile that year. Among all the states, the highest income level of the AI99-99.5 fractile is recorded in Connecticut ($665,798). Still, that income level of the bottom half of the top percentile (AI99-99.5 in Connecticut) needs to be multiplied by more than 3.1 to reach the lowest values (among the states) of the highest fractile (AI99.99-100), which was recorded in West Virginia the same year ($665,798 * 3.1 < $2,096,212). Another example is that of the year 1982, when earnings of about $200,000 placed a household in to the AI99-99.5 fractile in Connecticut[6] but to the AI99.99-100 fractile in South Dakota[7].

Fractiles can be determined graphically using the cumulative curve. The cumulative distribution function plots the number of tax units (i.e. the number of households who filed a tax return) against the different income thresholds. For instance, the median interval simply corresponds to that bracket of income that contains 50% of the statistical population. However, to get a numerical estimation rather than a broad interval,

---

[6] To be more precise, the AI99-99.5 fractile (i.e. the average income of the bottom half of top percentile) in Connecticut, 1982 is $199,214.

[7] To be more precise, the AI99.99-100 fractile (i.e. the average income of the top 0.01 percent) in South Dakota, 1982 is $200,293.

income statisticians use the fact that typically, income cumulative curves fit the Pareto distribution fairly well at the upper tail.

3.2.2 Fractiles are Estimated by the Pareto Interpolation Method

The Pareto interpolation is a method of estimating either the median, or the deciles, or the percentiles of a population that follow a Pareto distribution.

- The Pareto Distribution

Studying the income distributions of various countries, the Italian economist Vilfredo Pareto noticed a specific pattern of the income allocation among individuals: Whenever the amount of wealth doubles, the number of people falls by a constant factor. In the theoretical literature, this constant factor is usually called the Pareto coefficient and is labeled b. This factor may vary from country to country, but the pattern remains basically the same. Buchanan (2002) comments further on the Pareto distribution:

> Unlike a standard bell curve distribution, in which great deviations from the average are very rare, Pareto's so-called fat-tailed distribution starts very high at the low end, has no bulge in the middle at all, and falls off relatively slowly at the high end, indicating that some number of extremely wealthy people hold the lion's share of a country's riches. In the United States, for example, something like 80% of the wealth is held by only 20% of the people. But this particular 80-20 split is not really the point; in some other country, the precise numbers might be 90-20 or 95-10 or something else. The important point is that the distribution (at the wealthy end, at least) follows a strikingly simple mathematical curve illustrating that a small fraction of people always owns a large fraction of the wealth.

Visually, the Pareto distribution can be represented by the probability density function (PDF) taking the power form $f(y) = \alpha\, k^{\alpha} / y^{(1+\alpha)}$, where y stands for income, α is a scalar, and k, the minimum level of income in the distribution. It captures well the 80-20 percent rule as f(y), the "probability" or fraction of the population that earns a small amount of income per person (y), is rather high for low-income levels, then decreases steadily as income y increases: The higher the values of α, the wider the gap of inequality.[8]

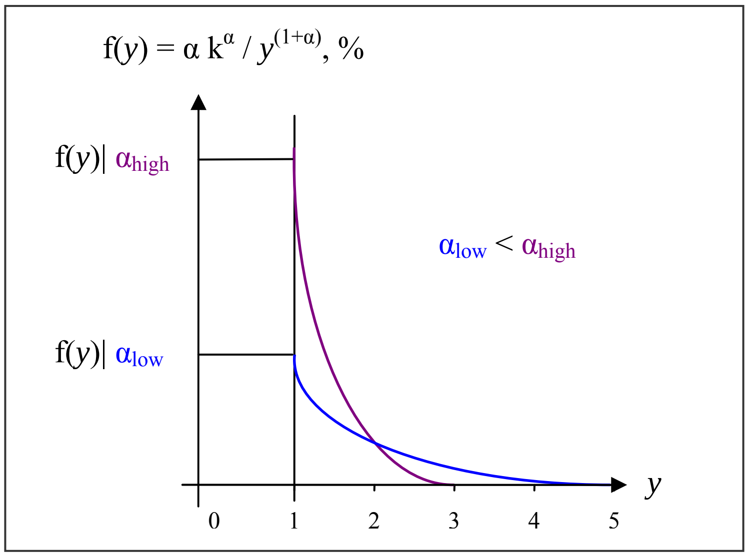


Figure 3.1 Pareto Probability Density Function

Justifying that an income allocation fits Pareto distribution is the first step towards the calculation of dispersion indicators, such as the 17 fractiles to be estimated. This estimation can be performed using the Pareto interpolation technique (where the curved portion of the cumulative curve is approximated to a straight segment).

---

[8] The scalar α will be used again in the analysis on convergence (Chapter 6). In that chapter, the hypothesis that the top income fractiles are well approximated by the Pareto distribution will be extended to the functional form of the Lorenz curve, without making further assumptions. The Lorenz curve will be used to derive in all states the lower fractiles of the distribution (down to the lower decile).

- The Pareto Interpolation Method

The Pareto interpolation method suggests a specific formula to estimate fractiles. In the case of the top decile TI90, the formula is written as follows:[9]

(3.1) $$TI90 = k\,(10\%)^{1/\alpha}$$

The estimation of the parameters k and α start with the determination of s_i, the lower bound of the income interval $[s_i ; s_{i+1})$ displayed in the IRS tables. For instance, the IRS table for Alabama in 2003 (listed in the table below), displays N_i and Y_i for each income bracket $[s_i ; s_{i+1})$.

Table 3.3 All the Steps from IRS Tables to TI90

Income brackets	s_i lower bound	N_i	N_i^*	Y_i in \$1,000	Y_i^* in \$1,000	$y_i = Y_i^* / N_i^*$ in \$
[1 ; 30,000)	1	1,088,495	1,883,765	14,335,594	74,842,664	
[30,000 ; 50,000)	30,000	332,057	795,270	12,931,107	60,507,070	76,084
[50,000 ; 75,000)	50,000	229,168	463,213	14,052,671	47,575,963	102,709
[75,000 ; 100,000)	75,000	117,246	234,045	10,063,119	33,523,292	143,234
[100,000 ; 200,000)	100,000	92,505	116,799	12,021,949	23,460,173	200,859
200,000 or more	200,000	24,294	24,294	11,438,224	11,438,224	470,825
Total		1,883,765		74,842,665		

[9] To calculate the top percentile instead, then the formula becomes: $TI99 = k\,(1\%)^{1/\alpha}$.

Table 3.3 continued

Income brackets in 1,000 $	**y_i =** Y_i^* / N_i^* in $	**b_i =** y_i / s_i	**α_i =** $b_i / (b_i - 1)$	**p_i % =** N_i^* / N^*	**k_i =** $s_i \cdot [\, p_i^{(1/\alpha i)} \,]$	Min $\lvert p_i - 10\% \rvert$ i = 1 to 5	**TI90** $k_i / [\, 0.1^{(1/\alpha i)} \,]$
0.001-30							
30-50	76,084	2.54	1.65	34.10	15,634.98	24.10	
50-75	102,709	2.05	1.95	19.86	21,812.81	9.86	
75-100	143,234	1.91	2.10	10.03	25,084.20	0.03	75,125
100-200	200,859	2.01	1.99	5.01	22,235.44	4.99	
200 or more	470,825	2.35	1.74	1.04	14,480.79	8.96	

Columns N_i^* and Y_i^* simply correspond to the cumulative sums of N_i and Y_i, respectively. The variable labeled y_i represents the average income earned by the individuals lying in the $[s_i ; s_{i+1})$ income interval. The Pareto coefficient $b_i = y_i / s_i$ divides the average income earned in the income class $[s_i ; s_{i+1})$ by the minimum income of that class. It is then straightforward to derive the parameter $\alpha_i = b_i / (b_i - 1)$. Both α_i and p_i, the fraction of tax units earning more than s_i are used in the estimation of $k = s_i [\, p_i^{(1/\alpha i)} \,]$.

Next, the lower bound s_i chosen in the calculation of threshold income TI90 is such that the fraction p_i of tax units with income above s_i is as close as possible to the fractile TI90. This way of choosing s_i is one of Piketty's (2001) contributions. Finally, the Pareto formula mentioned above is applied here: estimated decile TI90 = $k / 0.1^{(1/\alpha)}$. Similarly, the formula estimating the top percentile is TI99 = $k / 0.01^{(1/\alpha)}$, and so on for the other threshold-income fractiles.

Getting average-income fractiles requires nothing more than multiplying the Pareto coefficient b_i with the corresponding threshold income. Using the same example of the top decile as previously, average income level is defined as follows:

(3.2) $$AI90\text{-}100 = b_i \,.\, TI90$$

Then, inter-fractiles are deduced from simple subtractions summarized in the table below.

Table 3.4 Inter-Fractiles Calculation

AI90 - AI95	2 (AI90-100) - (AI95-100)
AI95 - AI99	[5 (AI95-100) - (AI99-100)] / 4
AI99 - 99.5	2 (AI99-100) - (AI99.5-100)
AI99.5 - 99.9	[5 (AI99.5-100) - (AI99.9-100)] / 4
AI99.90 - 99.99	[10 (AI99.9-100) - (AI99.99-100)] / 9

After the computation of fractiles for each state and each year from 1913 to 2003, more adjustments need to be made. Section 3.3 describes the main corrections added to the time-series (and the appendices describe all other adjustments not discussed in the text).

3.3 The Main Adjustments

To ensure the homogeneity of the time-series, various adjustments had to be performed. Below are the three corrections that most affected the data-set: 1) the exclusion of capital gains, 2) the treatment of the 1944 change in the definition of income, and 3) the upward adjustment of the count of tax units, smaller in size than households. The interested reader will find in Appendix 2 the description of all other adjustments.

3.3.1 Excluding Capital Gains

Among the different sources of income (wages, dividends, interests, business incomes, etc.) the capital gains declared are realized, and only a fraction of realized gains is reported on the tax returns. Moreover, capital gains feature an extreme volatility in the short-term, and keeping their aberrant values in the sample skews the curve of the overall trend. Unequivocal comparisons therefore require the exclusion of all capital gains from the upper income time-series. Piketty and Saez (2004) note (p. 33):

> From 1913 to 1933, 100% of capital gains were included in net income (there was no capital gains exclusion); from 1934 to 1937, 70% of capital gains were included in net income (i.e. 30% of capital gains were excluded); from 1938 to 1941, 60% of capital gains were included in net income (i.e. 40% of capital gains were excluded); from 1942 to 1978, 50% of capital gains were included in net income (1942-1943) or in AGI (1944-1978) (i.e. 50% of capital gains were excluded); from 1979 to 1986, 40% of capital gains were included in AGI (i.e. 60% of capital gains were excluded); from 1987 on, 100% of capital gains were included in AGI (there was again no capital gains exclusion).

After withdrawing the share of declared capital gains from total income, one problem occurs. The subtraction of capital gains in each fractile series leads to a significant reduction, if not the elimination, of the highest incomes. This is the case for these tax units filing returns with capital gains as a main source of income. In other words, the removal of capital gains generates a new series whose ranking is inconsistent with the initial income-class intervals as displayed in the IRS publications. The rank reversal is an issue that needs to be corrected. According to Piketty and Saez, the top fractile AI99.99-100 needs to be increased by 40 percent, the fractile AI99.9-99.99 by 2 percent, AI99-99.5 and AI99.5-99.9 by 1 percent to preserve the initial income ranking.

AI90-95 and AI95-99 are left unaffected. These corrections are applied here to the fractile series uniformly across states. It seems reasonable to apply the treatment of capital gains uniformly across states because splitting capital gains into 51 different areas also divides the biasing effect by the same number.

3.3.2 Net Income and Adjusted Gross Income (from 1944 on)

The IRS definition of income has varied over time. The IRS used the term 'net income' until 1943, and 'adjusted gross income' from 1944 on. In the net income definition, the various deductions (charitable gifts, interest paid, local taxes, etc.) taken into account underestimated the 1913-1943 series compared to the 1944-2003 series. In consequence, the final series from 1913 to 1943 had to be adjusted upward. The sharpest correction concerned the top percentile, inflated by up to 20 percent. To a lesser extent, the 1944-2003 series also had to be adjusted upward (for the same reasons as explained by Piketty and Saez, 2004, p. 33, iii).

As in Piketty and Saez (2004), income is here defined as the annual gross income reported on tax returns, excluding capital gains and government transfers (such as social security pensions, unemployment benefits, welfare payments, etc.) and before individual income taxes and employees' payroll taxes.

3.3.3 Households versus Tax Units

In this paper, the terms 'household' or 'tax unit' are used similarly, one as being the substitute for the other. However, there is a difference between 'tax unit' and 'household'. The tax-unit series was constructed by Piketty and Saez (2004, p. 4) according to the following definition. "A tax unit is defined as a married couple living together (with dependents) or a single adult (with dependents), as in the current tax law." Due to the limited availability of data at the state level, the tax-unit series could not be identically constructed at the state level, hence, the resort to the household population as a proxy variable for tax units. The household series comes from the Bureau of the Census who defines the term 'household' as follows:

> A household includes all the persons who occupy a housing unit. A housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied (…). The occupants may be a single family, one person living alone, two or more families living together, or any other group of related or unrelated persons who share living arrangements.

Therefore, to believe that households approximate well tax units is a delicate question because more than one tax unit may be included in one household (e.g. spouses filling returns separately, students co-renting the same housing, etc.) In other words, the number of tax units is usually greater than the number of households. Whether income per tax unit or income per household is chosen, the latter exceeds income per tax unit (by 30 percent on average, according to Piketty and Saez, 2004, p. 30).

This is the reason why the number of households had to be adjusted upward so that the national aggregate of the time-series I collected fits Piketty and Saez's series of tax units, leaving the initial state proportions intact. In other words, I calculated the percent share that state *i* represents in the U.S. number of households using BEA state data. Then, the same percentages (state / U.S.) were applied to Piketty and Saez's denominators to get a new series by state. This is the most significant adjustment made for the N* variable.

3.4 51 Ratios of Top Incomes over State Mean: Within-State Inequality

The within-state inequality ratios are defined as the average income per household accruing to each fractile divided by state *i*'s average income per household. State *i*'s average income data at the denominator cannot come from the IRS publications because the number of income tax returns does not represent more than the richest 10 percent of the population in the early years of the sample. In other words, IRS data cannot represent the state average income consistently from 1913 to 2003, hence the resort to another data source for a comparison of top incomes (IRS) with the mean income (BEA).

Unlike the IRS's Statistics of Income, the BEA's State Personal Income data are not displayed by income classes, and both differ in definition.[10] As mentioned in the introduction chapter, adjustments needed to be made on the BEA income series so that the BEA definition fits the IRS income (itself modified for time-consistency purposes as described in Section 3.3.2). The details of the adjustments made on the BEA personal income series are presented in Appendix A.1.2. Overall, to divide the top income levels by the state average yields an indicator measuring within-state inequality.

---

[10] "In addition to non-taxable government transfers, non-taxable personal income includes imputed rent; interest and dividends received by pension plans, life insurance carriers and non-profit institutions; non-taxable employer and employee contributions to pension plans, health insurance, day care, etc.; capital and inventory adjustments (NIPA capital consumption is generally smaller than IRS capital consumption, so that NIPA entrepreneurial income is generally larger than IRS entrepreneurial income); etc. See Park (2000) for a detailed description of the differences between NIPA personal income and individual tax return income." (Piketty and Saez, 2004, p. 28, note 49.)

3.5 51 Ratios of Top Incomes over U.S. Mean: Between-State Inequality

There are 51 inter-state inequality ratios, one for each state and one for the District of Columbia. These ratios divide the per-household income of state *i*'s top fractiles by the national average of per-household income. Converted in 2003 dollars, the national income used from 1913 to 2003 comes from Piketty and Saez's series, as it is already corrected with all necessary adjustments. As well as an indicator of inter-state inequality, this 'convergence ratio' parallels Piketty and Saez's income shares whose fluctuation interval ranges from 0 to 1. Piketty and Saez's income shares represent the percent weight of top incomes in total income.[1] Here, however, the top income per household is divided by the national income per household (not total income), and therefore most of the ratios exceed unity. The few that lie within unity are the states where the richest per household incomes did not go past the level of the national average; this particular case typically occurred in the early years of the time-period 1913-2003.

3.6 Income Shares by State

Income shares measure the weight of top incomes in total income at the state level. They are also part of the panel database.

---

[1] See Les Hauts Revenus en France, Appendix B, p. 619, footnote 1, and "Income Inequality in the United States", long version, p. 38.)

Chapter 4

CROSS-SECTIONAL AND TREND-OVER-TIME ANALYSES

The results can be analyzed over time, by cross-section, or both at the same time. First, the income levels are presented with the charts of fractiles for the full time period and by region, as defined by the U.S. Census Bureau: West, Midwest, South, and Northeast.

Table 4.1 The U.S. Census Regions

West	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.
Midwest	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin.
South	Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia.
Northeast	Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.

Second, the ratios dividing the top income levels by the state average income are mapped in selected years. Third, the indicator capturing the departure of top income levels from the U.S. mean is mapped annually from 1913 to 2003 and animated in a short video file that any standard media player can read.

4.1 Income Levels over Time

The figures below show the regional trend of fractiles AI90-95, AI95-99, and AI99-100 for the entire time period. The regional data were obtained by taking the average values of the states in that region. The state recording the highest level of income for fractile AI99.99-100 is also displayed. Throughout the 91 years of observations, the peak income of the top 0.01 percent (AI99.99-100) exceeds by 23 times the maximum value of the top percentile (AI99-100) in the Midwest, and 37 times in the West.

4.1.1 West

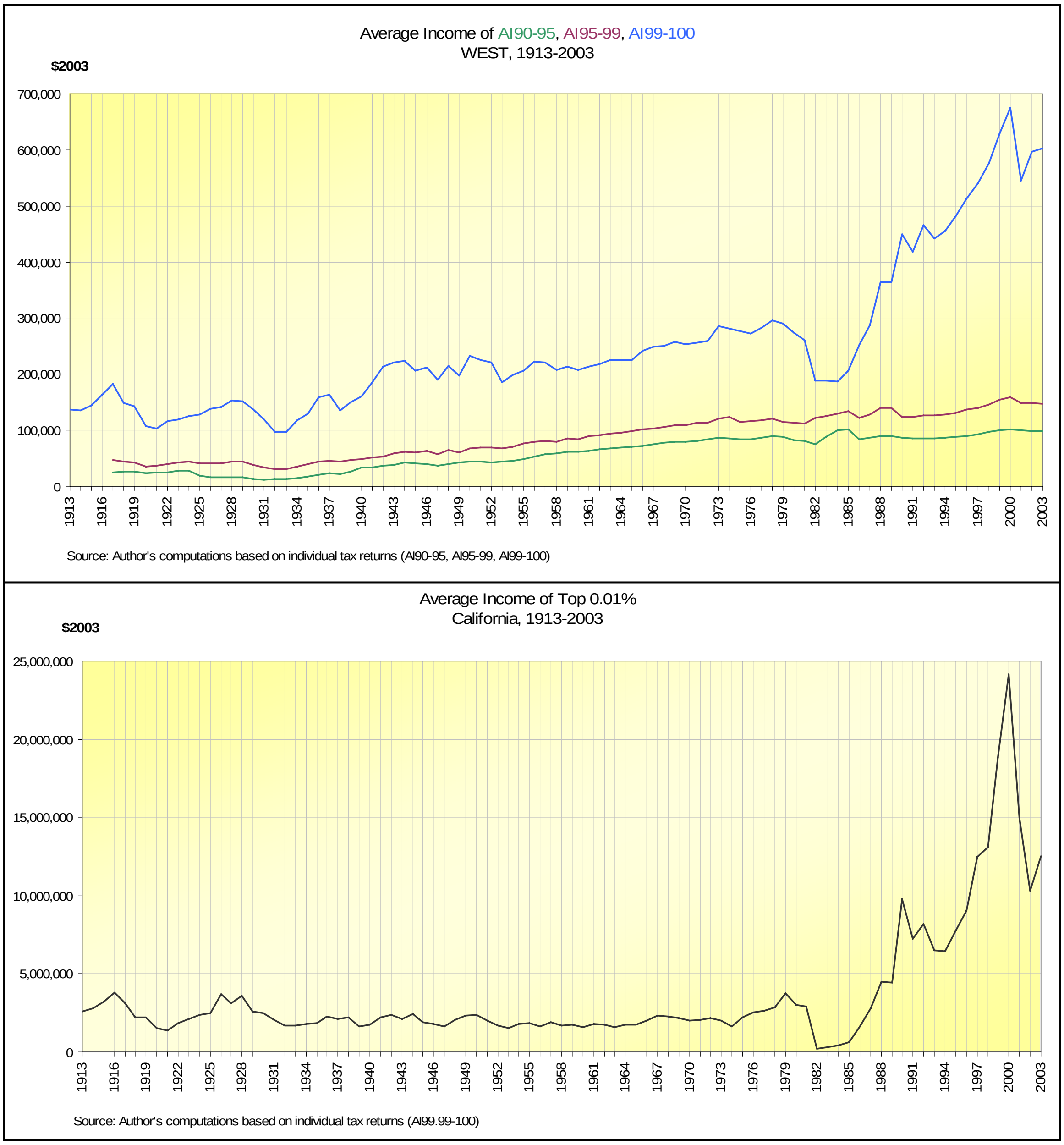


Figure 4.1 Average Income Levels in the West

The striking feature of this figure is the dramatic rise in income level of the top 0.01 percent at the end of the 1990 decade in California.

4.1.2 Midwest

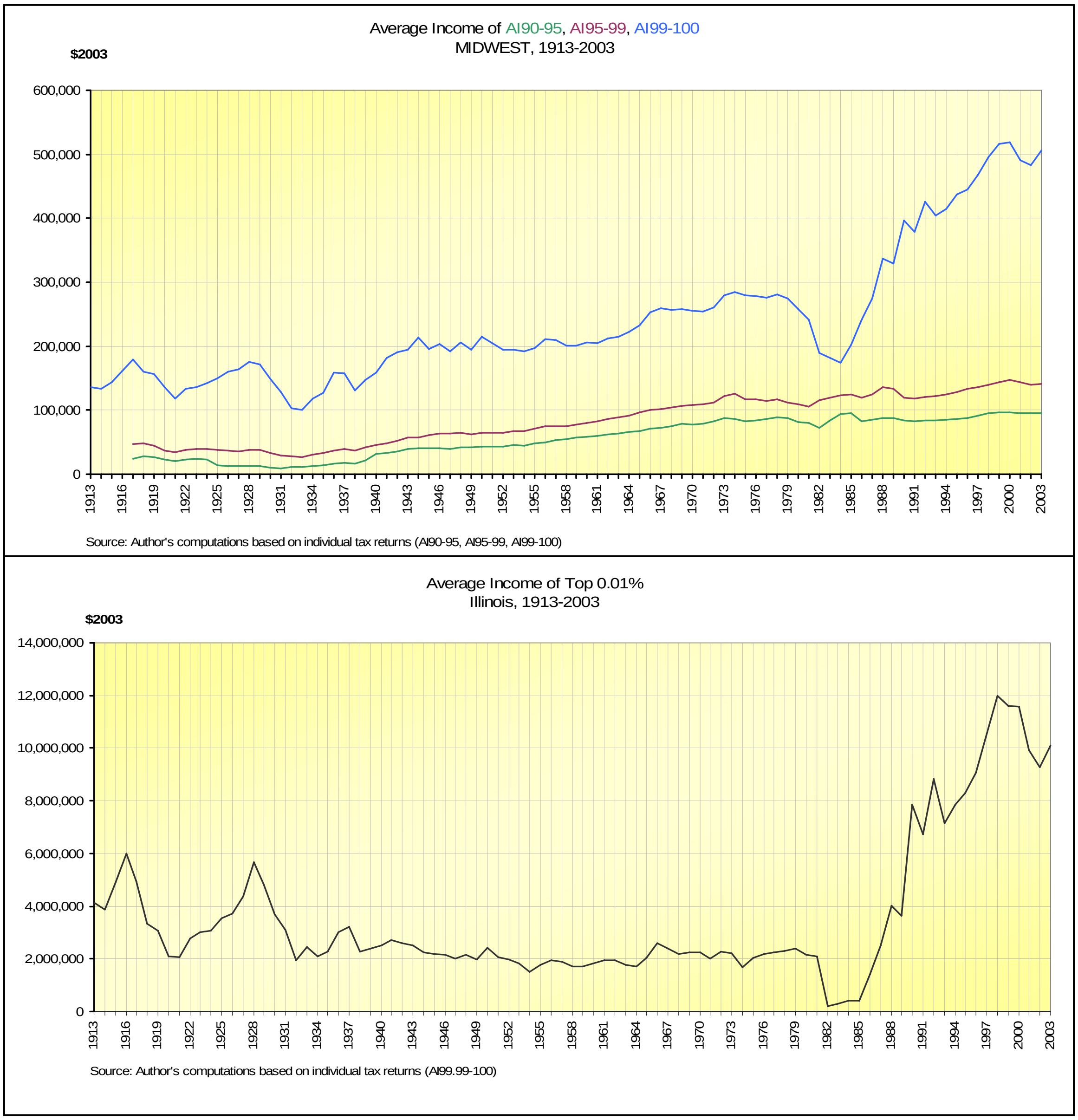


Figure 4.2 Average Income Levels in the Midwest

Occurring in 1982, the sharp drop in income of the top 0.01 percent is hard to explain. Even though the accuracy of the IRS data were affected by a very restricted number of income brackets in the SOI tables that year, the following years still confirm a very pronounced downward trend.

4.1.3 South

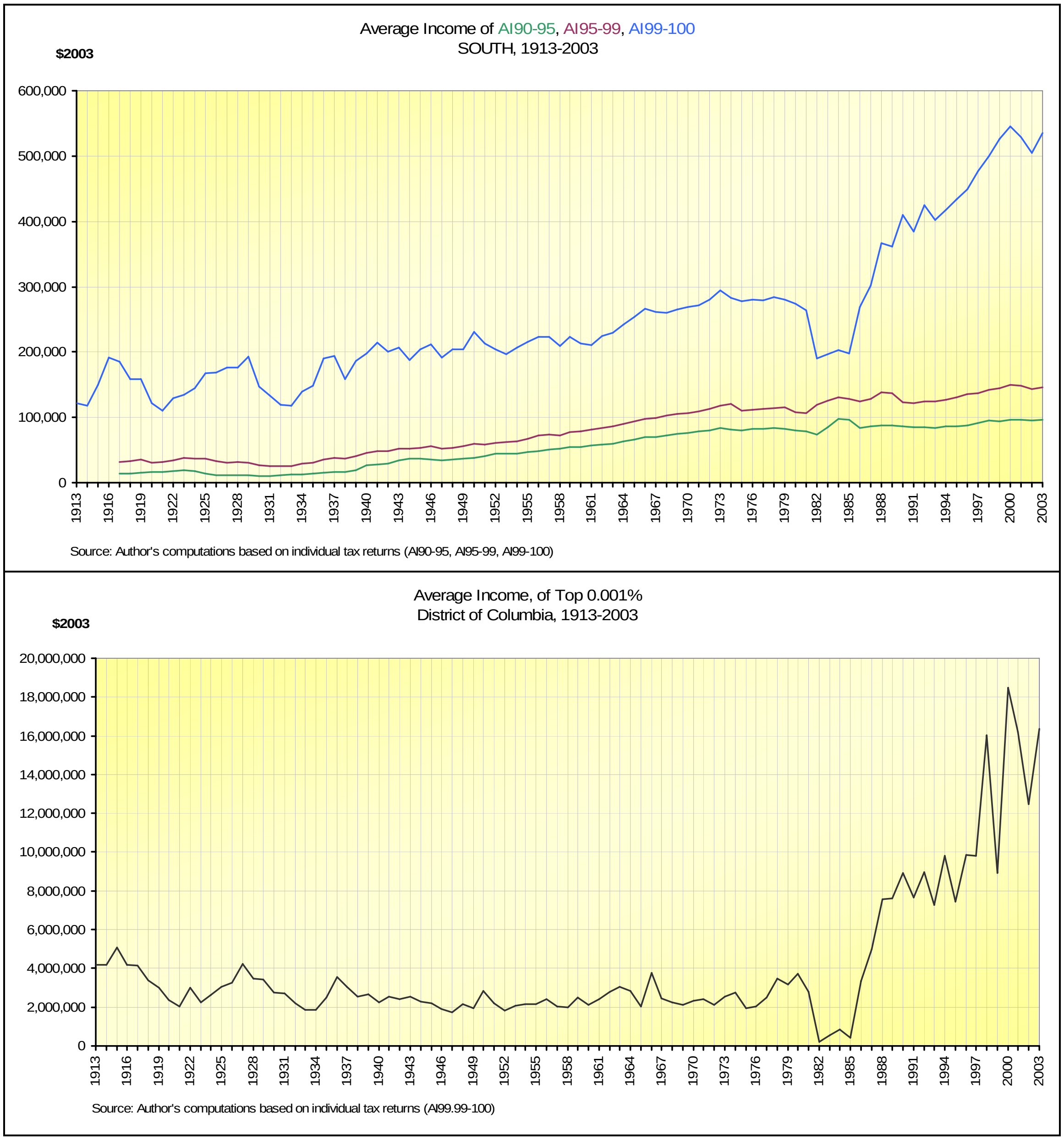


Figure 4.3 Average Income Levels in the South

Again, the top 0.01 percent seem to reach all-time high levels at the end of the sample period, far above the ranges of the beginning of the sample 1913-2003.

4.1.4 Northeast

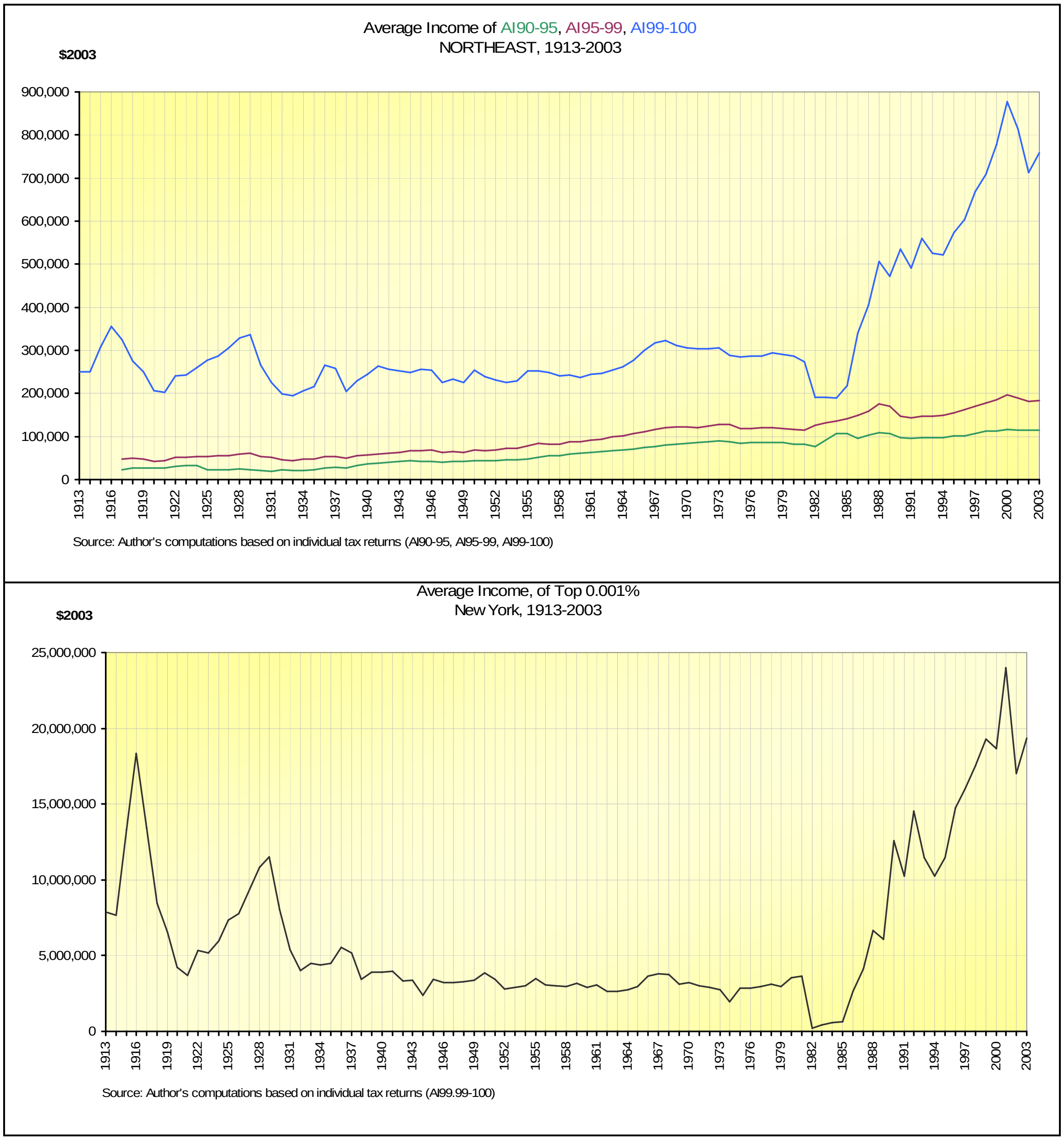


Figure 4.4 Average Income Levels in the Northeast

For the four regions, the 4 percent between the 95th and 99th percentiles of the distribution grow slowly and at the same rate over time. For instance, the AI95-99 incomes all start around $40,000 in 1917, and reach the threshold of $100,000 (constant dollars) in 1966 for the West and Midwest, 1968 for the South, and 1964 for the Northeast. Barro and Sala-i-Martin (1995) conclude that states grew roughly at the same rate over the past century: this is true only for the fractiles AI90-95 and AI95-99.

4.2 Within-State Inequality ($ytop_{i,t}$ / $ybar_{i,t}$)

4.2.1 Maps

In regard to the upper percentile relative to the state mean, only the results based on the top percentile (AI99-100) are displayed below. Between 1913 and 2003, the values of the intra-state inequality ratio range from 1.4 to 47. For the map of year t to be comparable to the map of any other year $n \neq t$, a ranking of the values in four equal intervals is applied to the entire time period (as opposed to the same partition applied to one particular year). 5.9 is the threshold of the first quartile, 7.6 is that of the median, and 9.7 of the third quartile. Hence the same legend for all maps:

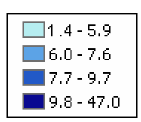


Figure 4.5 Legend

Here, the magnitude of the third quartile (75-100 percent) is 37.3 (= 47 – 9.7) and refers to the top percentile. It lies between 3 and 1,833, its equivalent values based on fractiles AI90-95 and AI99.99-100, respectively.

The years selected below (1929-1933, 1939-1945, 1949-1979, 1990, and 2000) represent the beginning or the end of historical events that affected the U.S. economy more or less severely.

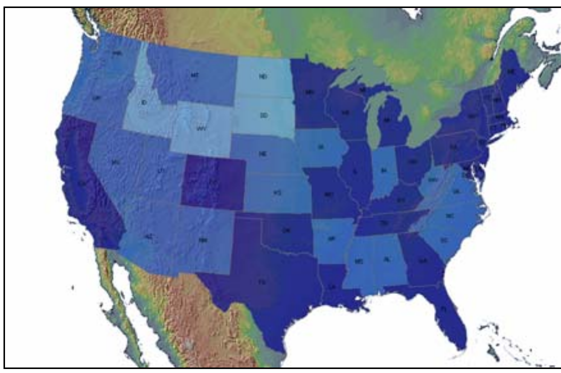


Figure 4.6 1929

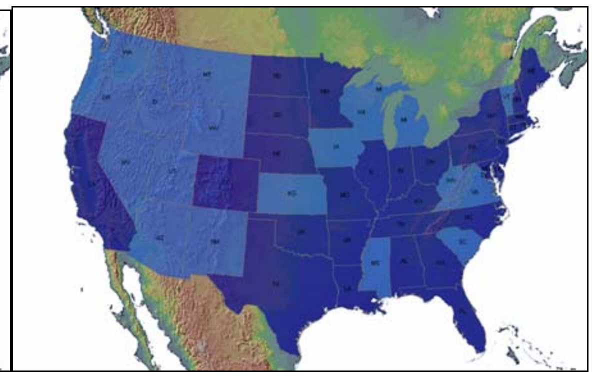


Figure 4.7 1933

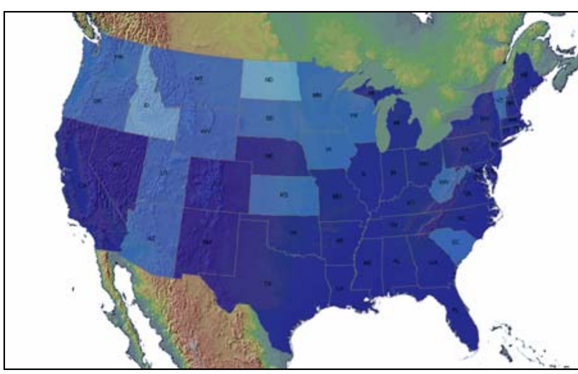


Figure 4.8 1939

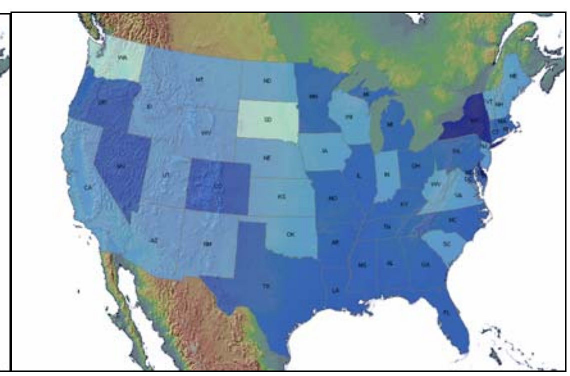


Figure 4.9 1945

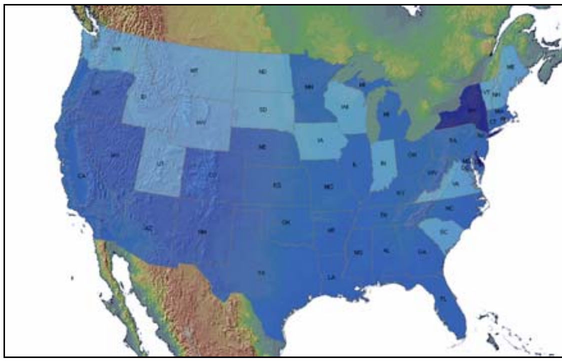
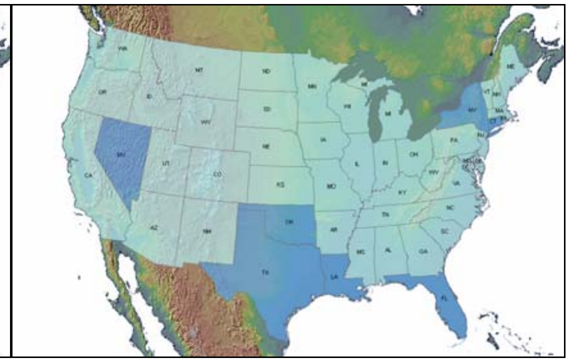



Figure 4.10 1949 **Figure 4.11 1979**

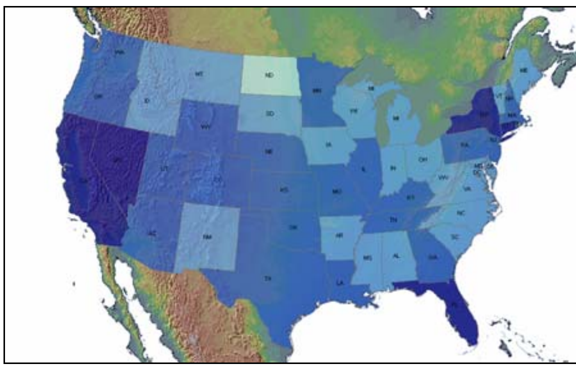
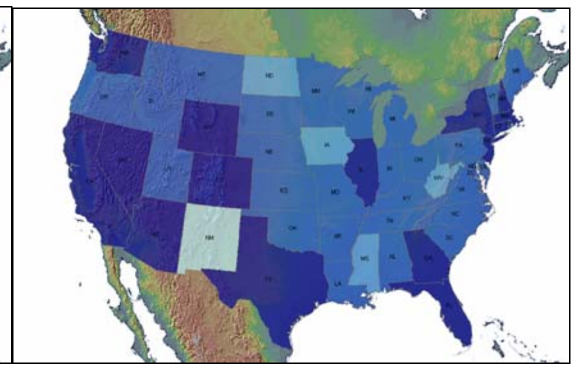



Figure 4.12 1989 **Figure 4.13 2000**

4.2.2 Trend over Time

Regressions of intra-state inequality on time and state by state yield time coefficients that are summarized in the three plots below.

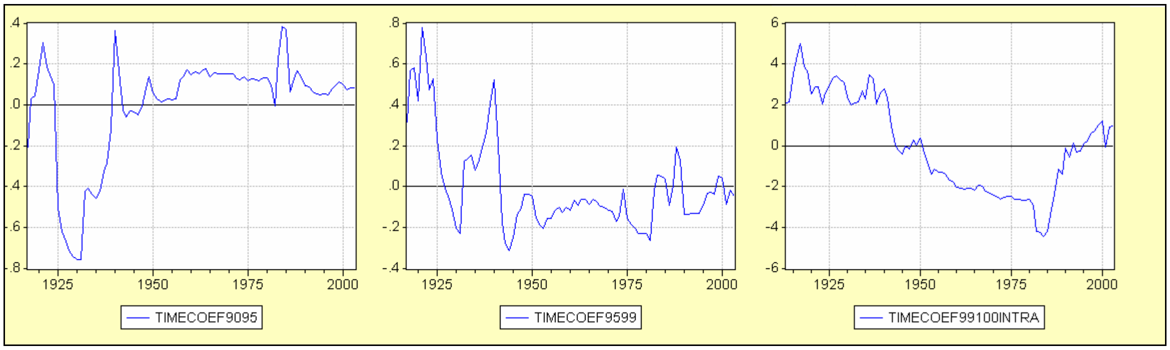


Figure 4.14 Time Coefficients of Intra-State Inequality Regressions

An alternative way of presenting the results is to resort to the coefficient of variation (CV). By definition, the coefficient of variation divides the standard deviation by the mean, measures the volatility of a series with respect to its mean. The coefficient of variation $CV_t(ytop_{i,t} / ybar_{i,t})$ is here regressed on time to observe the variability of within-state inequality over time. The table below shows that the annual income gaps within states have significantly declined in magnitude from 1913 to 2003.

Table 4.2 Variability in Within-State Inequality

Dependent variable: $CV_t\,(ytop_{i,t} / ybar_{i,t})$	AI 90-95%	AI 95-99%	Top 1%
Regressor coefficient: Time	-0.0035*	-0.0022*	-0.0015*

* Statistically significant at the 1% confidence level.

A visual representation of the coefficients of variation illustrates the same result, in a lesser extent for the top percentile (in the right panel):

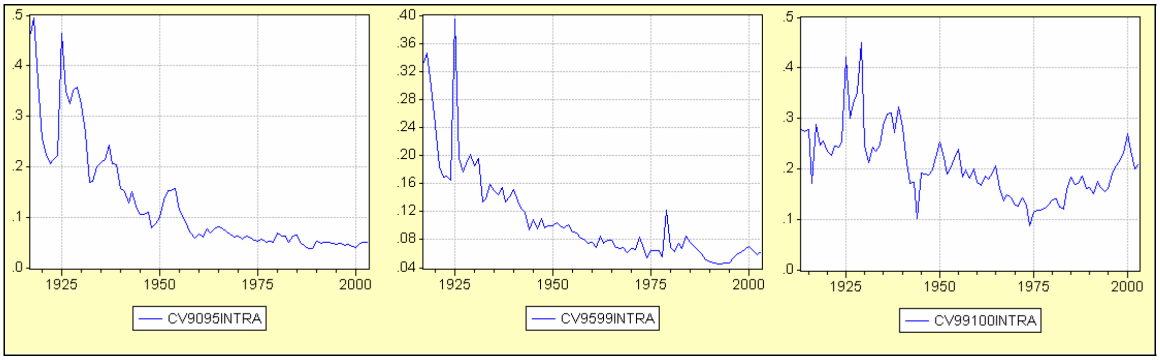
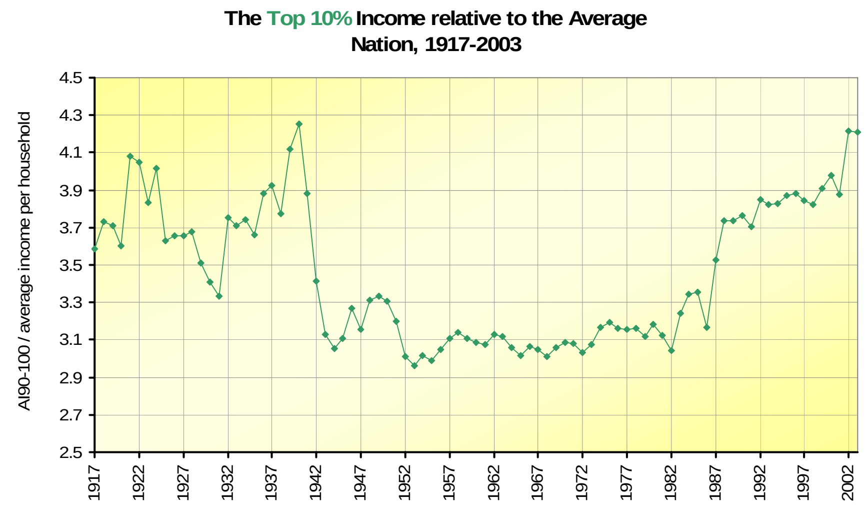


Figure 4.15 Coefficients of Variations of Intra-State Inequality

4.3 <u>Inter-State Inequality</u> ($ytop_{i,t}$ / $ybar_{US,t}$)

The estimation method applied at the state level is the same as Piketty and Saez's at the national level. Therefore, the unweighted addition of top incomes across states (top panel) should match Piketty and Saez's top income figures at the national level (bottom panel).



Source: Sommeiller's computations

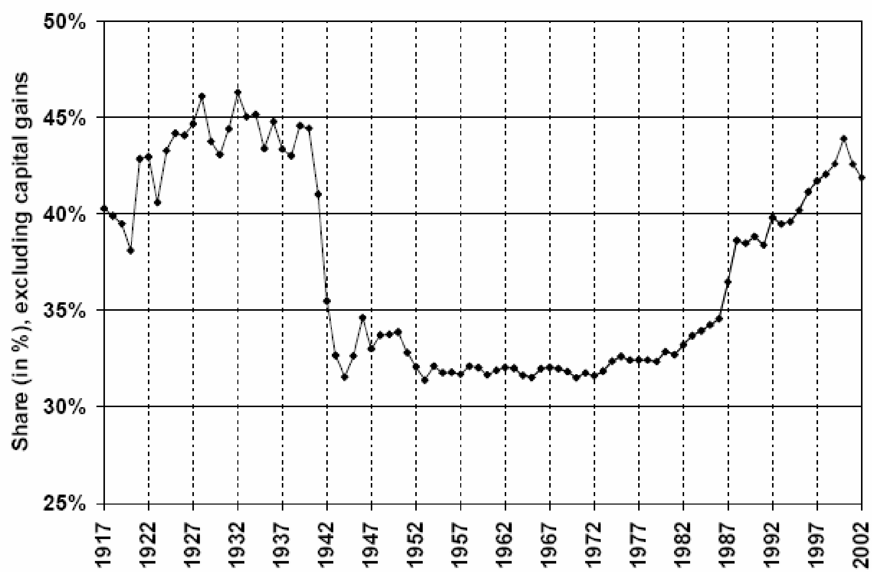


FIGURE 1
The Top Decile Income Share, 1917-2002

Source: Piketty & Saez (2004, p. 47)

Figure 4.16 Top Decile at National Level: Here versus Piketty and Saez (2004)

The reason why this calculation is done is to check the accuracy of the *state* estimates calculated here. Basic regressions of one against the other yield a (positive) R^2 of 0.8. The unexplained part of the correlation may be linked to several factors. One of them is that the two curves above are not identically defined. The former divides the top decile income by *per household* income (and typically exceeds unity) while the latter divides the top decile income by *total* income (and is smaller than one). This should not affect the similarity in their respective trend, however. Another explanation refers to the unit income as measured by Piketty and Saez, who constructed a tax-unit series to calculate income per tax unit, while we refer here to the income per household. Indeed, the availability of the data at the state level did not allow the state and national series to be identically constructed, despite all the adjustments already performed on the household series.[1]

One may keep in mind that if the incomes of the top 10 percent do not exceed 4.5 times the national average during the entire time period, the top 0.01 percent reaches a peak of 380 times in 1916, and always lies above 150 times the national average, except for the 1940-1990 time interval.

4.3.1 Animated Mapping

Below is the media file showing the temporal evolution of the ratio dividing the top percentile by the U.S. average personal income. More precisely, it consists in putting together all the annual maps of the ratio, and the juxtaposition of all maps

---

[1] More explanations are provided in Section 3.3.3, and in appendices A.1.3.

chronologically creates an animated effect recorded in a video file. Ranging from 3.1 to 73.4 (in all years and states), the values of the inter-state ratio are displayed with 2 visual effects. The first one is a 3-dimensional projection that gives height to the state polygons, the height being equal to the state value of the ratio. The second is the partition of the values into 4 equal groups of colors, shading from light blue (low values) to dark blue (top values), breaking at the 3 quartile thresholds (Q25 = 6.6, Q50 = 8.4 and Q75 = 11).

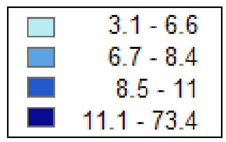


Figure 4.17 Legend for the Video File

The animation file lasts about 1 minute and 30 seconds. Click here to play the video. For readers using the paper-version of this thesis, below is one map extracted from the video.

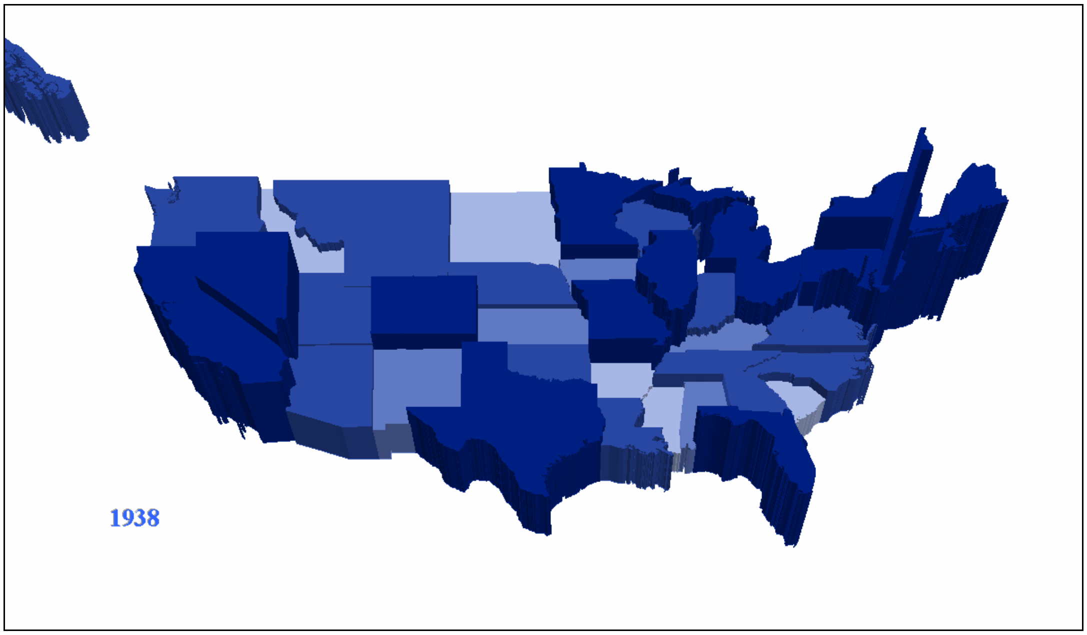


Figure 4.18 Inter-State Inequality in 1938

The evolution of the top income with respect to the national average is referred here as 'inter-state convergence', and can be summarized in a trend analysis.

4.3.2 Trend over Time

Inter-state convergence regressions produce the following time estimates.

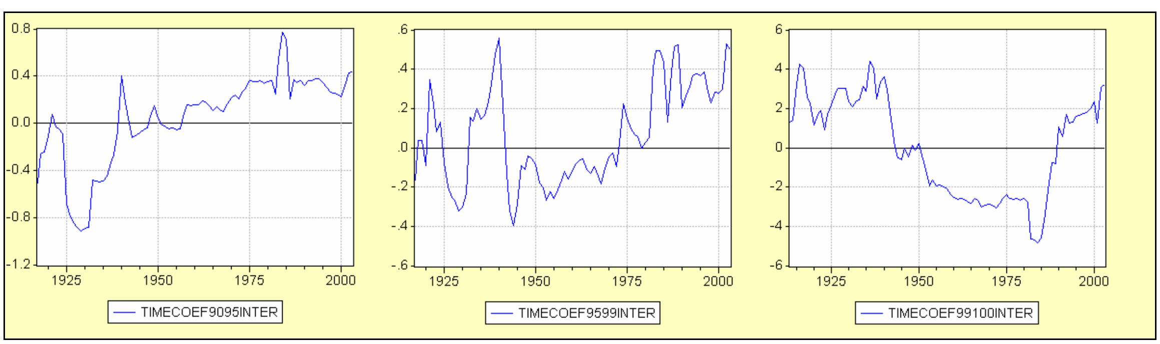


Figure 4.19 Time Coefficients from Inter-State Inequality Regressions

There is a clear difference between the fractiles of 90-99 percent and the top percentile. While the increasing trend of the former reveals a further deviation from the national mean after the 1940s, the latter sharply converges to the national mean after the Second World War until the mid 1980s.

Chapter 5

CONVERGENCE 1:

INCOME INEQUALITY AND ECONOMIC GROWTH

Chapter 5 deals with the evolution of income inequality compared to the growth rate of state average income. As mentioned in Chapter 3, state average income refers to the BEA series of per household personal income by state adjusted to Piketty and Saez's series, and income inequality is measured with various ratios of the top percentile.

This chapter aims at assessing the sensitivity of the results to two different approaches. The first one is borrowed from Kuznets (1958); the second one is taken from Barro and Sala-i-Martin (1995). Therefore, this chapter starts with a Kuznets-type analysis based on the ranking of states according to their respective initial average income (BEA data). The analysis continues with a formal econometric analysis based on the generalized method of moments to address the issue of the β convergence. Next, the concept of the σ convergence is applied to both average and top incomes, using different measures of their respective dispersion across states.

5.1 Kuznets-Type Analysis Based on the Ranking of States

The Kuznets-type analysis leads to distinguish three stylized facts: 1) the persistence of income differentials among states over time, 2) the high-income states display the widest income gaps, and 3) the relationship between low income levels and fast growth rates does not hold during recessions. In all three cases, whether a state is considered rich or poor is based on the ranking of the state average income series.

5.1.1 Persistence of Income Differentials Among States over Time

Overall, the rank order of the states with respect to state average income has remained strikingly stable over time. The richest third (i.e. the top 17 states) in 2003 includes 13 of the 1913 wealthiest states. In addition to listing the high-income states, Table 5.1 displays the corresponding income per household expressed in 2003 dollars. The state names that do not appear in both the columns have been grayed out.

Table 5.1 High-Income States in 1913 and 2003

1913		2003	
Alaska	35,317	Connecticut	87,634
California	34,176	New Jersey	84,721
District of Columbia	31,799	District of Columbia	84,310
New York	26,209	Massachusetts	79,210
Connecticut	24,962	Maryland	78,063
New Jersey	24,562	California	77,580
Illinois	21,728	New York	74,655
Michigan	21,464	Alaska	71,553
Delaware	20,905	Hawaii	70,804
Massachusetts	20,002	Illinois	70,417
Rhode Island	19,588	New Hampshire	69,837
Hawaii	18,544	Virginia	68,643
Arizona	18,517	Colorado	68,167
Pennsylvania	17,710	Minnesota	68,153
Maryland	17,405	Delaware	68,045
Ohio	16,964	Washington	65,019
Washington	16,653	Rhode Island	64,903

This secular persistence of income discrepancies across states generalizes what Perloff, Dunn, Lampard, and Muth (1967, p. 502) found for per capita income by state from 1920 to 1955. The authors noted that “there has been remarkably little change in income differences among states over these years. While it is true that some states have had more rapid increases in per capita income than others and there has been a tendency toward equalization of differentials in relative terms, the narrowing of the differentials appears to be related primarily to short-run fluctuations and year-to-year random movements.” The authors point out two economic upheavals featuring the time-period

1920-1955: 1) the Great Depression, and 2) the inflationary spiral that occurred in the aftermath of World War II.

5.1.2 The High-Income States Display the Widest Income Gaps

The main idea is to oppose two series: 1) the level of state average income per household, and 2) the size of the state inequality gap. More precisely, the state average income per household corresponds, as discussed above, to the BEA state personal income adjusted to Piketty and Saez's series. The inequality gap is measured by the inter-state ratio of the top percentile (derived from IRS tables), i.e. the departure of the state top incomes (top 1 percent) from the national average. The first series is ordered in descending sequence, and then partitioned into 3 groups of 17 states each: high-, medium-, and low-income states. Similarly, the second series partitions states of wide, moderate, and narrow income gaps. How many of the poorest states are featured by high inequality? Kuznets predicted the reverse correlation.

Figure 5.1 illustrate the results for selected years, with negative correlation (left panel) and positive correlation (right panel):

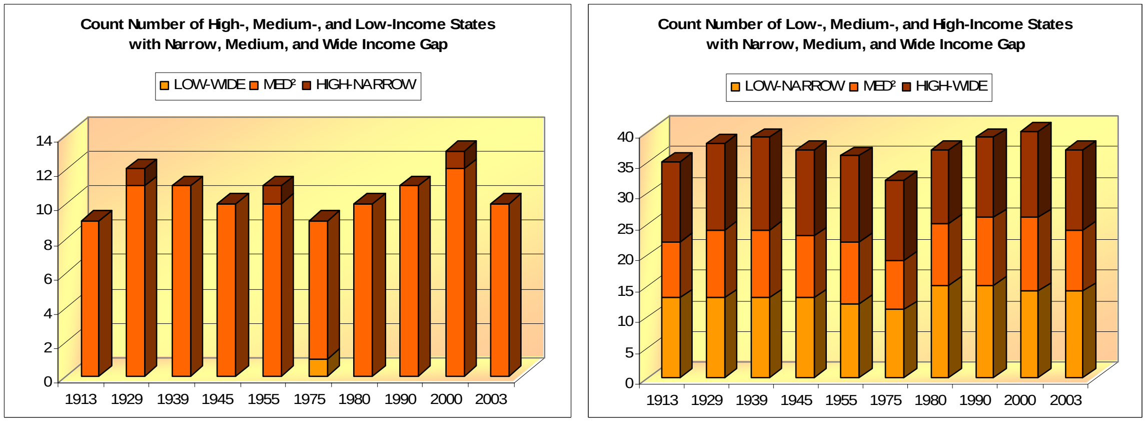


Figure 5.1 The Richest States Record the Widest Inequality Gaps

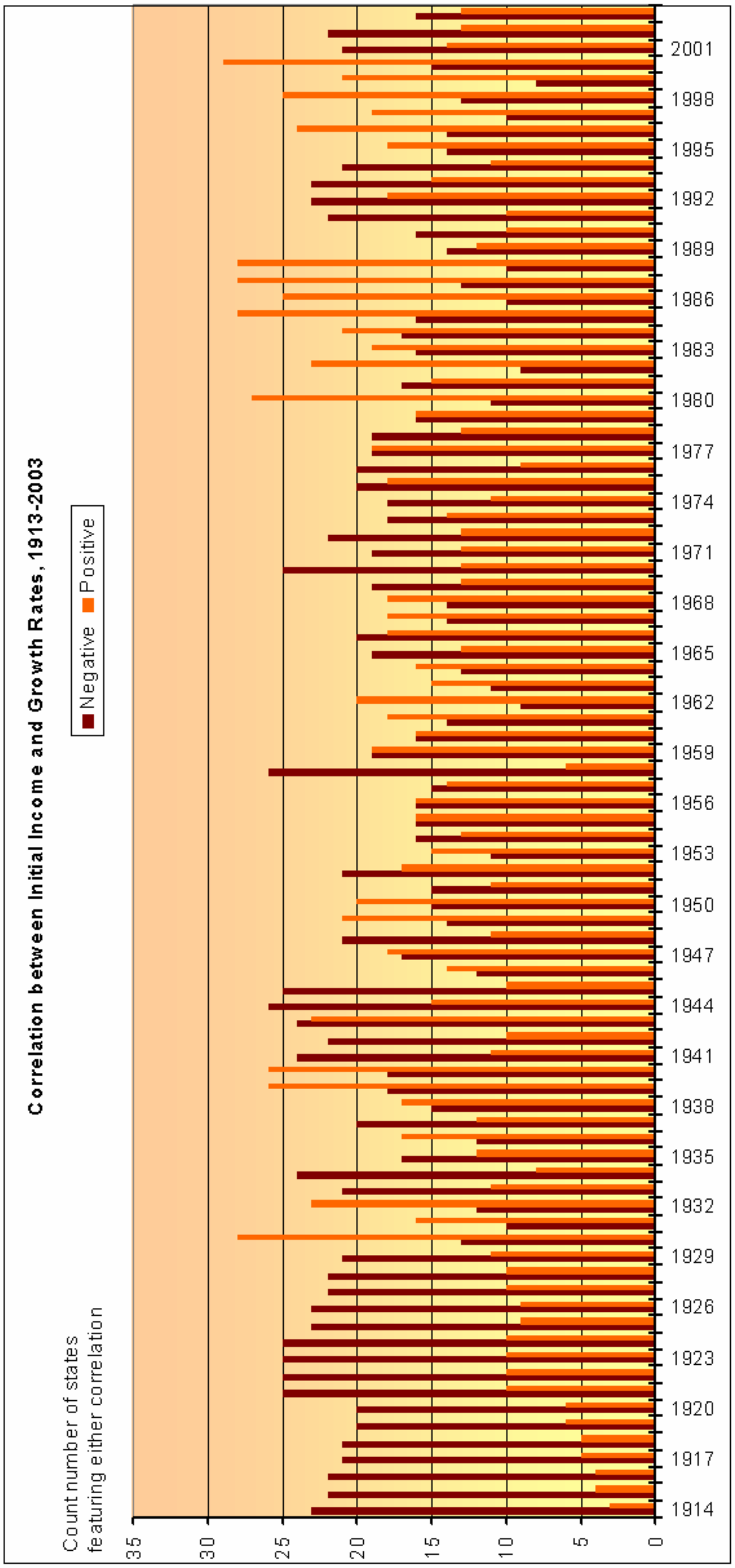
It is evident in Figure 5.1 that there are many more states where a narrow income gap is associated with a low level of income (and sharp inequality with high income levels) than states recording a low income average along with a wide income gap (and vice-versa for 'high-narrow'). Moreover, the state count in either case barely varies over time. The standard deviation for the negative and positive correlations is 1.9 and 3.7, respectively. This result contradicts Kuznets hypothesis, but prior to drawing such a conclusion, we need to focus on how state growth rates evolved with inter-state inequality, measured in the next section by the ratio of the income earned by the top 1 percent relative to the national average income.

5.1.3 The Relation Low Income - Fast Growth Holds in the Long-Run

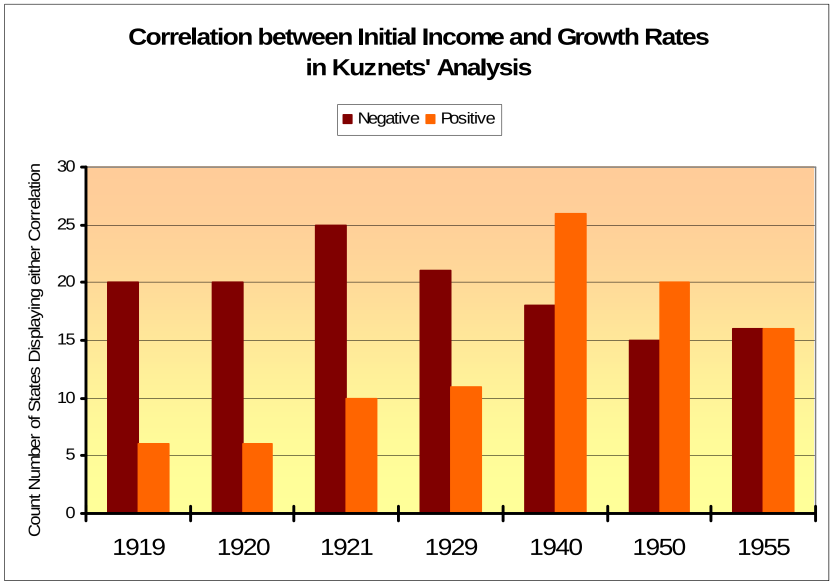
In his study of industrial distribution of income by states, Kuznets (1958, p. 42) writes that "the states with the highest rates of growth in per capita income had low per capita incomes in the initial year; and those with lowest rates of growth in per capita income had initially high per capita incomes… In other words, in the country's growth during recent decades interstate inequalities in per capita income were reduced."

However, Kuznets analyzed only five years or periods (1919-21, 1929, 1940, 1950, and 1955) to draw his conclusion on. Using the series of the state average of per household income, we compare here the number of low-income states with fast growth rates to the number of low-income states with slow rates of growth.[1] Again, the data shed light on this point. The two panels of the figure below provide a visualization of the results. The first one utilizes the full size of the sample (1913-2003), and the second duplicates the former for selected years only (1919-21, 1929, 1940, 1950, and 1955 from Kuznets' perspective).

[1] Alternatively, we could have compared the number of high-income states with slow growth rates to the number of high-income states with fast rates of growth.


Correlation between Initial Income and Growth Rates, 1913-2003
Negative
Positive
Count number of states featuring either correlation
0
5
10
15
20
25
30
35
1914
1917
1920
1923
1926
1929
1932
1935
1938
1941
1944
1947
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001

(a)



(b)

Figure 5.2 Correlation Ambiguities between Initial Income and Growth Rates

Figure 5.2 suggests convergence because the low-income states recorded fast growth rates 52 times out of the 90 years considered. This number exceeds the number of years when low-income states recorded slow growth rates (32 years out of 90). The remaining 6 years featured equality between income levels and growth rates. It is not surprising that both Kuznets and Barro concluded in favor of convergence because the years they considered displayed more states where initial incomes were negatively correlated to growth rates.

Figure 5.2 reveals that the negative correlation switched to a positive correlation in recessionary years: the aftermath of 1929's black Friday, the early 1950s, early 1960s, almost the full decade of the 1980s, and the second half of the 1990s. At these points in time, there were more rich states growing fast than rich states growing slow.[2] This remark suggests that the β convergence does not hold anymore in the downward phase of the Juglar's cycle, whose complete occurrence usually takes eight to ten years.

However, their conclusions were based on different measures of income. Kuznets used per capita income (and not per household income), thereby making comparisons less unequivocal. Note also that Kuznets' main argument was not to demonstrate the links between growth rates and income differentials, but to inform on the industrial distribution of income among the sectors of agriculture (A), manufacturing (M) and services (S), at a very refined degree of disaggregation. For 1919-1921, 1929, 1940, 1950, and 1955, he arrayed the percentage increase in states per capita income in descending order, and partitioned them into six groups of 8 states each. He performed both cross-section and trend over time analyses, and concluded:[3]

> The share of the A sector declined in all six groups of states; but the decline was greater in the states in which growth in per capita income was greatest. (…) Thus the results for the share of the A sector conform to the expectations derived from the cross-section analysis.
>
> The changes in the share of the M sector also agree with our expectations. The share rose in all six groups, but it rose more in the groups in which per capita income grew at a higher rate.
>
> The share of the S sector also rose fairly substantially in all six groups; but the association between the rate of growth of per capita income and the increase in the share is not as clear-cut as for the A and M sectors.

---

[2] Similarly, there were more poor states growing slow than poor states growing fast.

[3] Kuznets (1958, p. 42).

The industrial distribution of income across states is a topic where boundaries lie beyond the scope of the present work. The next step is to address the same issue of inequality and growth, through the traditional concept of the β convergence.

5.2 β Convergence and Inequality Regressions

The β convergence refers to the negative correlation between the initial income level and the growth rate at two different points in time. The initial income level is measure here by the state average income derived from the BEA series. Inequality equations, however, address the convergence issue from a different perspective and link the top incomes to the growth rate of state average income.

5.2.1 β Convergence: Growth and Initial Income Level

Testing the Solow model of growth[4], the traditional regressions found in the empirical literature use a cross-country sample to perform a linear regression of growth rates on a constant term and the initial level of per-capita income. With *i* indexing U.S. states, the main equation is

(5.1) $$(1/T)\ .\ \log\left[y_{i,(t0+T)}/y_{i,t0}\right] = \alpha + \beta \log\left[y_{i,t0}\right] + \gamma X_i + \varepsilon_{it},$$

where:

- T is the amplitude of the time-interval considered,

[4] According to the Solow growth model, two economies with similar technology, savings rates, and population growth rates converge to the same capital-labor ratio and level of per-capita income in the long run. In other words, if poor countries tend to grow at faster rates (per capita) than rich countries, their per-capita income levels are getting closer together over time.

- t_0 is 1917 for the top decile, and 1913 for the top 1 percent,
- log $[y_{i,(t0+T)}/y_{i,t0}]$ is the overall growth rate of the BEA personal income per household in state *i* between year t_0 and t_0+T,
- (1/T) . log$[y_{i,(t0+T)}/y_{i,t0}]$ is the annual growth rate of the BEA personal income per household in state i between year t_0 and t_0+T,
- $y_{i,t0}$ is the initial income level,
- X_i represents one or more additional variables that may affect growth,
- ε_i is a random error term, and
- α, β, and γ are parameters to be estimated.

Initial income is entered in log form because that allows the coefficient β to be interpreted as the marginal effect of a one-percent increase in initial income on the growth rate. A negative value of β provides evidence supporting the convergence hypothesis of the Solow model. [5]

As Crain (2003) pointed it out, the definition of the growth rate (of income) is an issue. Whether the growth rate is defined one way or another, the estimated coefficients of an equation may change in sign. To illustrate his point, Crain compared the 'continuously compounded' growth rate (as used in traditional convergence regressions) to the 'least squares' growth rate (that Crain derives from least squares trend regressions). In regressions based on the 'least squares' growth rates, the β coefficient is no longer statistically significant (even though it keeps the same sign). Crain extended the

---

[5] Barro and Sala-i-Martin define β in a different way and expect a positive β for a conclusion favoring convergence.

analysis from income per capita to income per worker and concluded again on the sensitivity of the results to the growth rate definition, as shown in the table below.

Table 5.2 Barro-Type Test for State Income Convergence, 1969-1999

TABLE 2.1. Barro-Type Test for State Income Convergence, 1969–99

	Growth in Income per Capita		Growth in Income per Worker	
Independent Variables	Continuously Compounded[a]	Least Squares[b]	Continuously Compounded[a]	Least Squares[b]
ln (Initial Income)	−0.009	−0.005	−0.007	−0.004
	(−3.85)**	(−1.38)	(−2.08)*	(−0.94)
Constant	0.104	0.063	0.080	0.049
	(4.61)**	(1.83)	(2.26)*	(1.04)
R-squared	0.27	0.05	0.10	0.02
F-statistic	14.71**	1.92	4.31*	0.88
Number of observations	50	50	50	50

Note: t-statistics are shown in parentheses.

[a]The continuously compounded growth rate is computed as $[\ln(X_{1999}/X_{1969})]$ / 30, where ln is the natural logarithm, X_{1999} is real income in 1999, X_{1969} is real income in 1969, and 30 is the number of years in the sample.

[b]The least squares growth rate is computed by regressing the natural logarithm of income in each state on a linear time trend as follows:

$$\ln(\text{Real Income per Capita}_t) = \text{Constant} + \beta_{ypc}(\text{Time Trend}_{1969-99}) + u_t,$$
$$\ln(\text{Real Income per Worker}_t) = \text{Constant} + \beta_{ypw}(\text{Time Trend}_{1969-99}) + u_t,$$

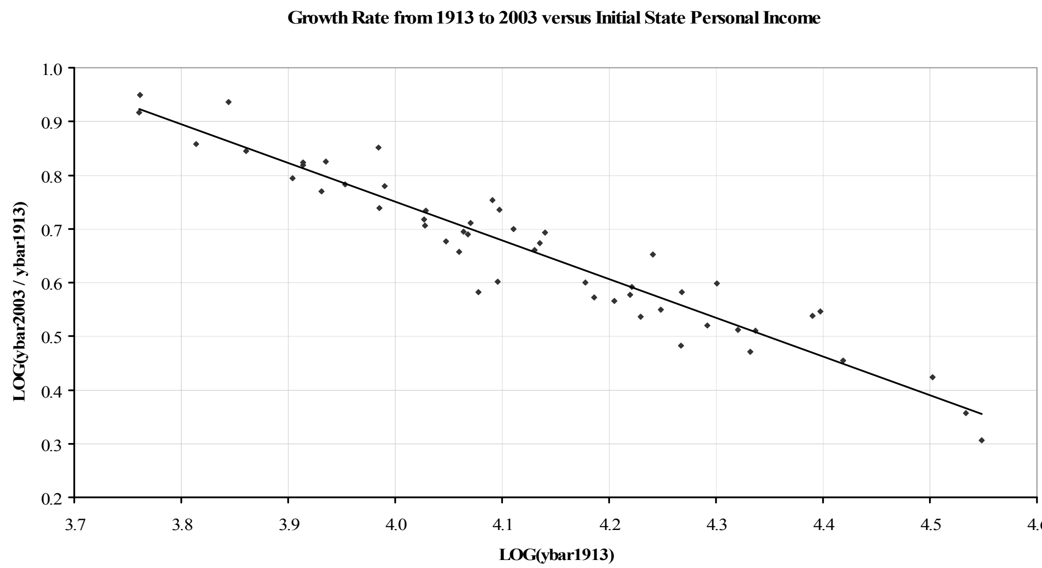
where ln refers to the natural logarithm, the subscript t refers to the value in each year, and u_t is the random error term. In this specification the estimated coefficients for β_{ypc} and β_{ypw} yield the annual growth rates.

* Indicates significance at the 5 percent level for a two-tailed test. ** Indicates significance at the 1 percent level for a two-tailed test.

Source: Crain (2003, p. 28).

Taking this detail into account, the next figure distinguishes two definitions of growth rates, both being based on state average income per household. The first one is equal to the logarithm of the ratio of income level in year 2003 over the income level in year 1913: $\log(ybar_{2003}/ybar_{1913})$. The second one is defined as the average of annual growth rates between 1913 and 2003. The results appear in panels (a) and (b) of the figure below, where each point represents a state.

(a)



(b)

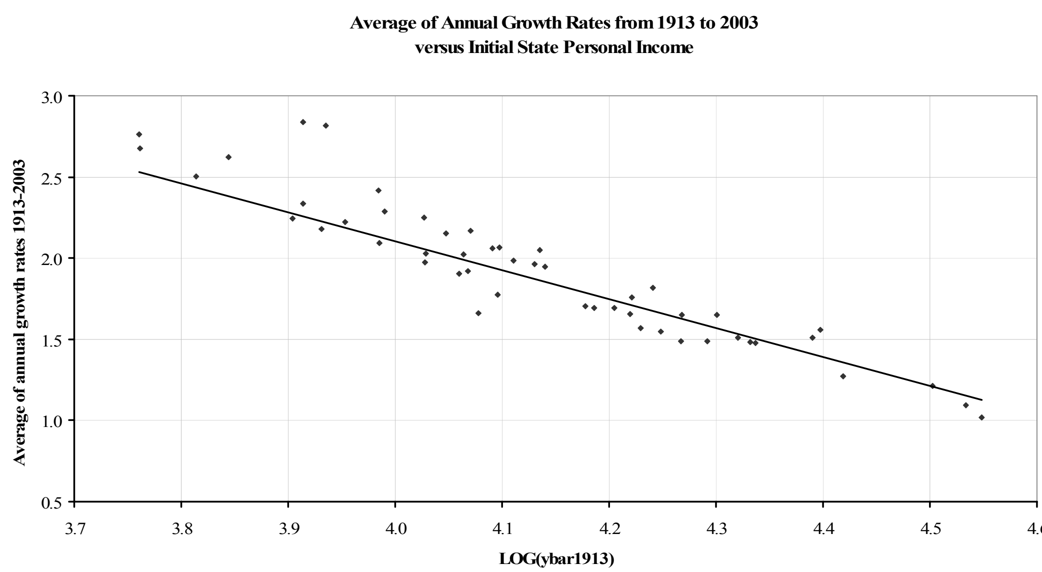


Figure 5.3 Growth Rates versus Initial Income Level

The declining trend clearly coincides with Barro and Sala-i-Martin's results. To visualize where the states are located on the scatter plots, we aggregated the data at the regional level and what was expected clearly appeared in the figure below.

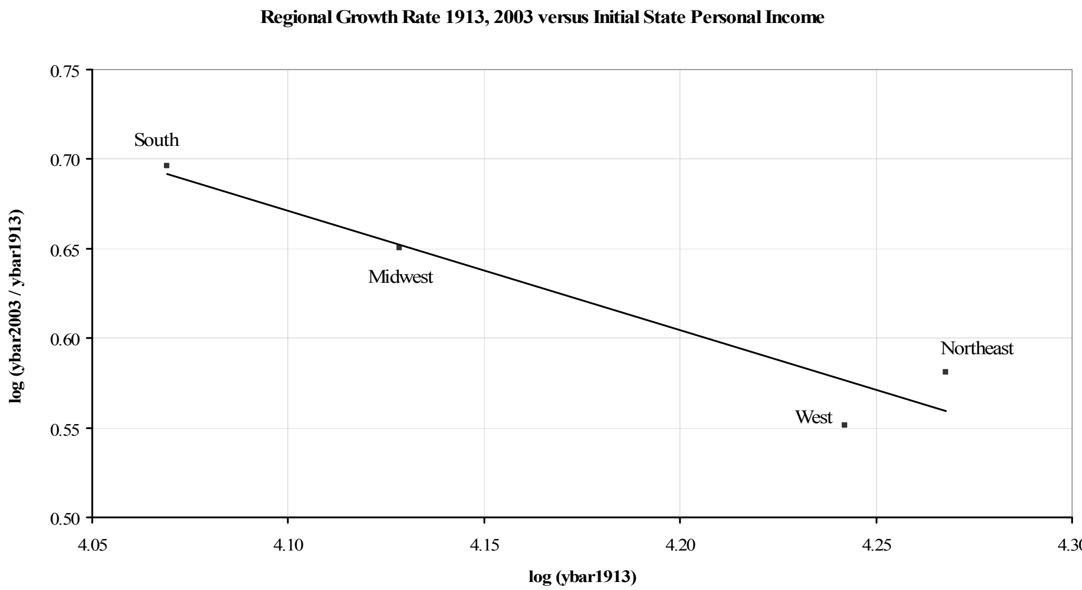


Figure 5.4 Regional Growth Rates versus Initial Income Level

As mentioned earlier, the data depicted in the two figures above are not derived from the IRS tables, but from the BEA state personal income data. The next section deals with the β convergence of the IRS incomes within the top 10 percent.

5.2.2 β Convergence Within the Top Decile

Convergence seems to occur within the top decile. Consider fractiles AI90-95, AI95-99, and AI99-100, (with all income levels expressed in constant dollars of 2003). The beginning of the time-period is 1917 for the first two, and 1913 for the third one. In the figure below, the horizontal axis is the one percent increase in the initial income level of the fractile considered, and the vertical axis is the growth rate of $ytop_{i,}$ between t_0 and (t_0 + T):

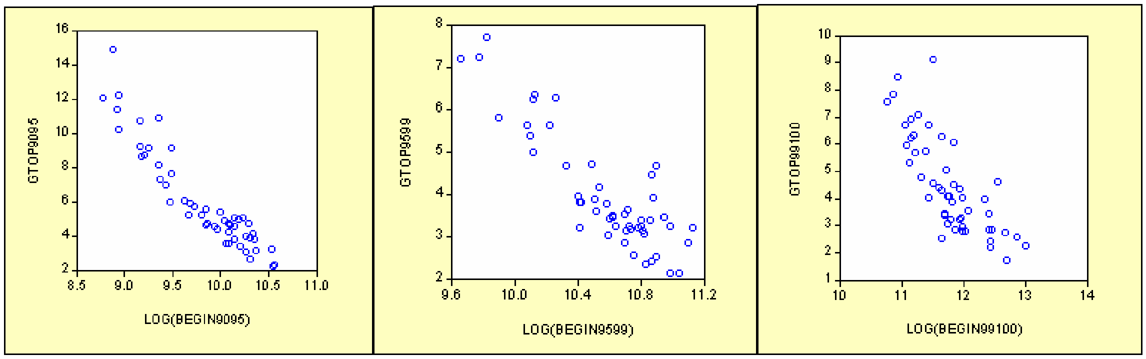


Figure 5.5 Convergence Within the Top 10 Percent

The three scatter plots actually correspond to equation

(5.2) $$(1/T) \,.\, \log [ytop_{i,(t0+T)}/ytop_{i,t0}] = \alpha + \beta \log [ytop_{i,t0}] + \varepsilon_{it}.$$

5.2.3 Inequality Regressions

Before estimating any regressions, one needs to consider the best method of estimation, given the panel data at our disposal.

- Generalized Method of Moments (GMM) as the Method of Estimation

Panel data lend themselves to dynamic models. The chosen method of estimation is a dynamic panel data model where the general approach relies on instrumental variables estimators and on a generalized method of moments estimator. Several reasons justify this choice. First, performing a Hausman test on the growth equation leads to the rejection of the assumption that the regressors are uncorrelated with the error terms. In other words, ordinary least squares estimates cannot be consistent. An alternative method to least squares estimation is the instrumental variable technique. In other words, one needs to introduce instrumental variables to the dynamic panel data specification which is given by Equation (5.3)

$$y_{it} = \alpha_i + \beta\, x_{it} + \delta\, y_{i,t\text{-}1} + \varepsilon_{it}. \qquad (5.3)$$

Second, the time-span 1913-2003 is long enough to use the lagged growth rates ($y_{i,t\text{-}2} - y_{i,t\text{-}3}$) as instrumental variables for ($y_{i,t\text{-}1} - y_{i,t\text{-}2}$). The first differences transformation applied to the dynamic model corrects for the cross-section effects and produces an equation of the form[6]:

$$y_{it} - y_{i,t\text{-}1} \;=\; \beta\,(x_{it} - x_{i,t\text{-}1}) \;+\; \delta\,(y_{i,t\text{-}1} - y_{i,t\text{-}2}) \;+\; (\varepsilon_{it} - \varepsilon_{i,t\text{-}1}). \qquad (5.4)$$

Third, the dynamic model adopted here corrects for time-period fixed effects with dummy variables and corrects for heteroscedasticity with a White diagonal instrument weighting matrix.

- Inequality and Initial Income Level: Significant Positive Correlation

The table below shows a strong positive correlation between inter-state inequality ($AI99\text{-}100_i$ / $ybar_{US}$) and the state average income at the beginning of each period.

[6] See Greene (2000, p. 583).

Table 5.3 Inter-State Inequality Increases with the State Mean Income

Dependent Variable: INTER99100			
Sample	Variable	Coefficient	Prob.
1913-1928	INTER99100(-1)	0.2629	0.0000
	YBAR	-9.89E-06	0.9799
1929-1939	INTER99100(-1)	0.0258	0.7042
	YBAR	0.0019	0.0032
1944-1979	INTER99100(-1)	-0.0233	0.6084
	YBAR	0.0001	0.0000
1980-1989	INTER99100(-1)	-0.1192	0.0001
	YBAR	0.0003	0.0001
1990-1999	INTER99100(-1)	-0.4013	0.0000
	YBAR	0.0007	0.0000
1999-2003	INTER99100(-1)	-0.6120	0.0000
	YBAR	0.0009	0.0000

Coefficients do not vary much when AI99.99-100 is used instead. Do the results change after the addition of another regressor in the GMM regressions? Reflective of the industrial composition of income in each state, this variable added to the X_i matrix is taken from Barro and Sala-i-Martin (1991, p. 117). It breaks down "state *i*'s personal income into nine standard sectors: agriculture; mining, construction; manufacturing; transportation; wholesale and retail trade; finance, insurance, real estate; services, and government." The data relate to the personal income accruing to each sector in 1930, 1940, 1950, 1960, 1970, and 1980 at the national level. The two authors first calculated the national growth rate of each sector from one decade to the other, and then weighted the national growth rates by the share of each sector in state *i*. The variable they compute is a single number per state summing all sectoral (weighted) growth rates together. The authors also computed the agricultural share of personal income for 1920-1930, which is included in the vector of the sectoral regressors.

Table 5.4 Adding Sectoral Variables to the Basic Equation

Sample	Variable	Coefficient	Prob.
1920-1980	INTER9999100(-1)	0.3856	0.0032
	YBAR	0.0048	0.0213
	SECTOR	40.8959	0.6890

The sectoral variables added to the basic equation fail to improve the results. They also affect the number of observations, dropping the sample size from 4,539 to 288. This is due to the decennial nature of the sectoral variable, measuring growth rates of a sector in a state income from one decade to the next (and not from one year to the next) during the time-period 1920-1988.

The correlation between the growth rate of the average income and inequality across states is the focus of the next section.

- Growth and Inter-State Inequality: Significant Negative Correlation

In this section, the main variable is no longer the initial income average, but the initial income in the top decile of the distribution. To what extent does the initial income of the wealthiest tax filers affect the growth rate of the state average income? The main equation is written as

(5.5) $$(1/T)\ .\ \log [y_{i,(t0+T)}/y_{i,t0}] = \alpha + \beta \log [ytop_{i,t0}] + \gamma X_i + \varepsilon_{it},$$

as opposed to

(5.6) $$(1/T)\ .\ \log [y_{i,(t0+T)}/y_{i,t0}] = \alpha + \beta \log [y_{i,t0}] + \gamma X_i + \varepsilon_{it},$$

where the new items are:

- $ytop_{i,t0}$ is the initial per-tax-unit income in the top decile,
- log [$ytop_{i,t0}$] is a one-percent increase in initial income in the top decile.[7]

The variables of the table below show the correlation coefficients (and their corresponding probability of being insignificant) between G, the annual growth, and the inter-state inequality indicator, labeled INTER99100, named after the 51 ratios dividing each year the top 1 percent fractile on the national average income.

Table 5.5 Correlation between Growth and Inequality

Dependent Variable: G			
Sample	Variable	Coefficient	Prob.
1913-1928	G(-1)	0.829	0.000
	INTER99100	-0.007	0.000
	TIME	6.702	0.000
1929-1939	G(-1)	-0.298	0.000
	INTER99100	-1.474	0.000
	TIME	-3.080	0.000
1940-1979	G(-1)	0.122	0.022
	INTER99100	0.496	0.426
	TIME	1.397	0.438
1980-1989	G(-1)	-0.081	0.000
	INTER99100	-0.355	0.000
	TIME	-0.585	0.001
1990-1999	G(-1)	-0.154	0.004
	INTER99100	0.472	0.029
	TIME	3.901	0.000
2000-2003	G(-1)	0.061	0.394
	INTER99100	0.012	0.954
	TIME	3.036	0.000

[7] The β coefficient is here interpreted as the marginal effect on the growth rate, and a negative sign is expected for a conclusion of convergence. However, Barro and Sala-i-Martin define β in a different way and expect a positive β for a conclusion favoring convergence.

What appears from the table above is unequivocal: whenever the inequality coefficient is negative, it is statistically significant. In other words, a decrease in the growth rate occurs along with a wider inequality gap, and vice-versa, but the two variables hardly move in the same direction at the same time. These results need to be taken with caution, as correlation does not mean causality. Granger tests of causality show that growth does not Granger cause inequality, and inequality does not Granger cause growth.

With the dependent variable being the annual growth rate of personal income in state *i*, the results of the GMM regressions including the industrial composition of income are summarized in the table below.

Table 5.6 GMM Regressions with the Sectoral Variable

Variable	Coefficient	t-Statistic	Prob.
G(-1)	-0.5204	-1.2333	0.2185
INTER99100	-0.9112	-1.4141	0.1584
SECTOR	36.5199	1.1143	0.2661
1920-1930	11.4669	1.6569	0.0987
1930-1940	34.9368	5.4154	0.0000
1940-1950	-15.0182	-8.3573	0.0000
1950-1960	-8.7370	-1.9994	0.0465
1960-1970	-0.3046	-0.6538	0.5138
1970-1980	-5.4314	-1.6213	0.1061

The R^2 jumps from 0.59 to 0.77, and the inter-state inequality indicator keeps its negative sign but loses its statistical significance. Note again that the sample size shrank drastically from 4,539 to 288, as the sectoral variable is available for 6 years and 48 states only.

5.2.4 Spatial Auto-Correlation

The purpose of this section is to test the robustness of the previous results against the test of spatial-autocorrelation of the panel data set. A traditional measure of spatial auto-correlation is Moran's *I*. It is used extensively in economic geography (Anselin, 1988, Janikas and Rey, 2005), and more and more by regional economists (Dall'erba and Le Gallo, 2005). Moran's *I* is a correlation coefficient (and therefore varies from negative one to positive one) weighted by the state-to-state distance matrix. More precisely, Moran's *I* is defined as follows:

(5.7) $$I = \frac{n}{S_0} \sum_i \sum_j \frac{w_{ij}(x_i - \mu)(x_j - \mu)}{\sum_i (x_i - \mu)^2}$$

where n is the number of observations, w_{ij} is the spatial weight matrix based on state-to-state distances, S_0 is the sum of all the elements of the spatial weight matrix, x_i and x_j stand for the error-term vector derived from the GMM regressions, and μ is the arithmetic mean of x. Usually, the weight matrix takes the inverse of the distance between state i and j, so that a high value of Moran's *I* indicates a cluster effect and a low value the independence of nearby points. "The Moran's *I* value that indicates spatial independence of values (or the lack of spatial autocorrelation) is a negative number close to zero" (Hare, 2001). We intentionally increase the differences in state-to-state distances by raising their values to power (minus) two:

(5.8) $$w_{ij}^* = \{d_{ij}\}^{-2}$$

where $\{d\}$ represents the distance between regions i and j. It is also common in the spatial data literature to normalize the rows of the weight matrix so that each row sums to unity, and so that S_0 equals n in the formula defining Moran's I.[8]

(5.9) $$w_{ij} = \frac{w_{ij}^*}{\sum_j w_{ij}^*}$$

Applied to the present data set, Moran's I takes the value of -0.2341, which tends to justify the fact that the attractive forces of economic poles spread locally, but not globally. In other words, polarization is a self-reinforcing phenomenon that pulls further apart high and low-income areas: high-income states exert attraction on neighboring areas (and these peripheral areas bring more income revenues to the core) but not to the extent of reaching rural remote areas where poverty falls into the trap of hysteresis, where low-income states stayed with low-income levels simply because it was the case in the previous time period.

5.3 σ Convergence

The σ convergence refers to the dispersion of income across regions. In Barro and Sala-i-Martin (1995), σ convergence is measured with the unweighted cross-sectional standard deviations of state per capita income. In addition to the standard-deviation, we consider here the coefficient of variation for a closer look at the data. The

---

[8] See Fingleton, Igliori, and Moore (2003) to learn more on the row-normalization of the weight matrix.

purpose of this section is to compare the dispersion of average incomes to the dispersion of top incomes.

5.3.1 Average Income and Top Percentile Income Decline in Dispersion

The point here is to compare the income dispersion of the top percentile (AI99-100) with the dispersion of average incomes (σ-convergence). Both dispersion indicators are measured by the unweighted cross-sectional standard deviations of income over time. Barro and Sala-i-Martin's $\sigma(\log y_{i,t})$ is based on the log of per capita personal income. Here $\sigma(\log ytop_{i,t})$ is based on the logarithm of the per household income accruing to the top percentile. The figure below shows two things: 1) similar trends of income dispersion; 2) the income dispersion displays sharper volatility for the top 1 percent income than for the average personal income.

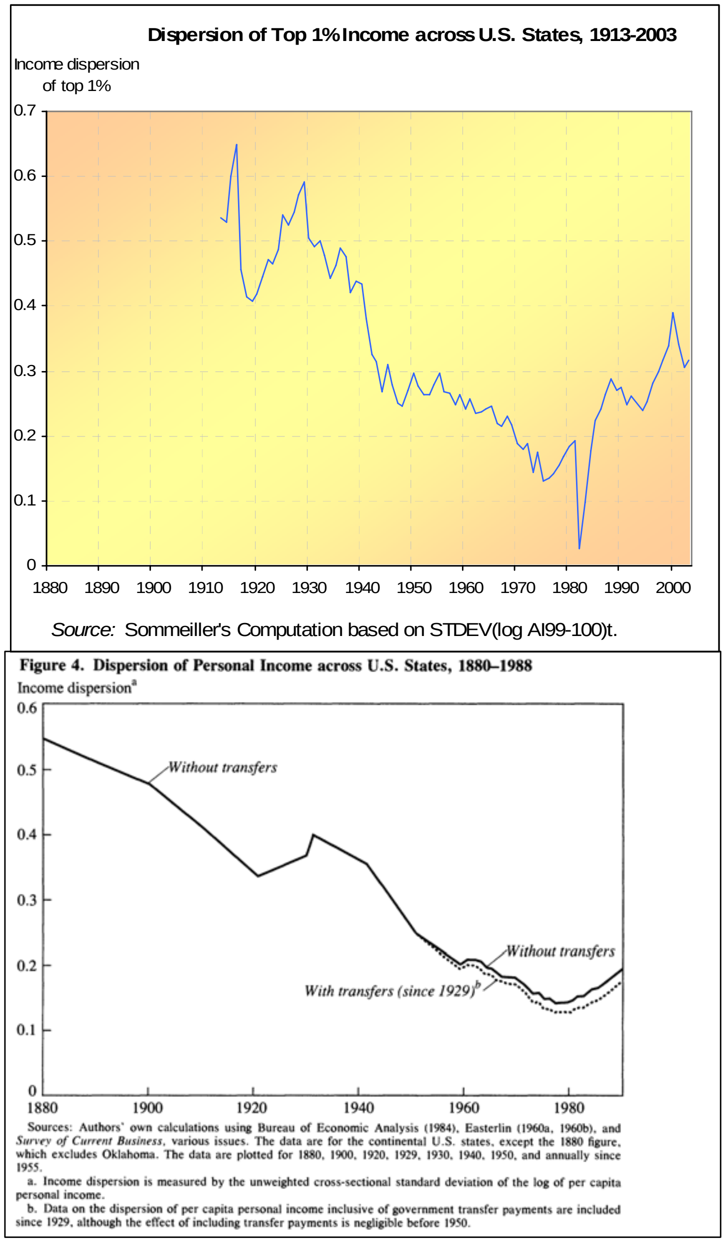


Figure 5.6 Dispersion between Average and Top Incomes are Alike in Trend

Very sensitive to measurement errors, the issue of income inequality ought to be considered in various ways, and we turn next to the measure of the income dispersion of both average income and the incomes of the 90-99 percent (instead of the 99-100 percent).

5.3.2 Incomes of Top 90-95 and 95-99 Percent Record Increasing Dispersion

The incomes of the 90-95 and 95-99 percent revealed an *increase* in dispersion across states over time. This result contrasts with the previous section where average income and income of top 1 percent both recorded a sharp *decrease* in dispersion across states over time (except after the mid 1980s). This difference is suggested twice: once with the coefficient of variation regressed on time (table below), and another time with graphs of the standard deviation (figure below).[9]

Table 5.7 Inequality Across States over the Past Century

Dependent variable: $CV_t(ytop_{i,t})$	AI 90-95%	AI 95-99%	Top 1%
Time coefficients	1,167.4 *	1,550.6*	-0.0047 *

* Statistically significant at the 1% confidence level.

---

[9] By definition, the coefficient of variation (CV) divides the standard deviation by the mean, measures the volatility of a series with respect to its mean, and is an index without unit. Why resorting to the CV ratio and thereby complicating the overall economic interpretation of its coefficient? While the standard deviation is sensitive to the addition of a constant, the coefficient of variation is not. For instance, assume a hypothetical standard deviation of income of $2,000 associated with a mean of $3,000 in a given year, thereby depicting a high degree of inequality. Had everyone earned a bonus of $100,000,000 the following year, the standard deviation would drop, suggesting a decline in inequality. Unlike the standard deviation, the coefficient of variation is scale invariant as the constant term cancels out by appearing on both the numerator and the denominator. Therefore, the coefficient of variation is often considered a better measure of income inequality, especially for series displaying different group means across sections or over time.

The figure below displays income dispersion within the top decile, measured by the coefficient of variation, and the decomposition of its ratio (standard deviations and means are expressed in 2003 dollars).

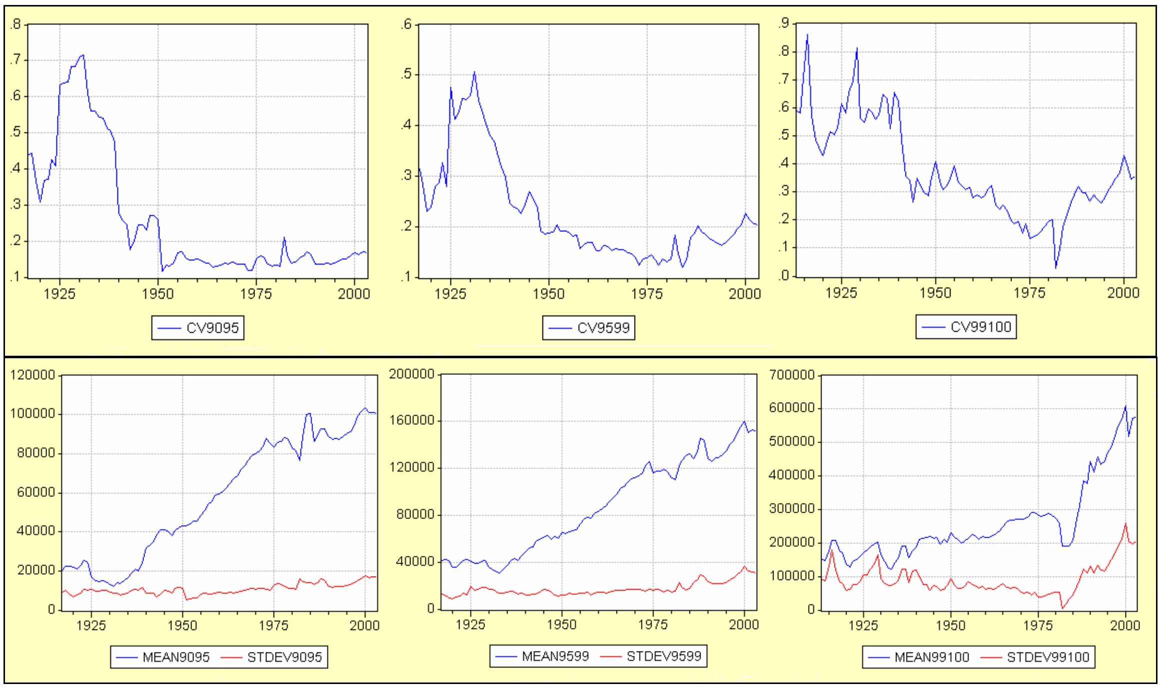


Figure 5.7 CV, Mean and Standard Deviation Within the Top Decile

A common point to all three groups is that the increasing portion of the CV curve after the mid-1980s is accounted for by a rise in both the mean and the standard deviation. However, the differences among them lie on the opposition mentioned earlier: Considering the coefficient of variation of the two fractiles of the 90-99 percent interval, one may notice their respective decrease between the mid-1930s and the late 1970s. This downward trend occurred along with a sharp increase in their mean income, but with a slow and smooth rise in their standard deviation. As for the top percentile CV, it experiences a drop between 1939 and 1980 due to opposite trends in the standard deviation (decreasing) and the mean (increasing), the same way the dispersion of average income did (bottom panel of Figure 5.6).

Chapter 6

CONVERGENCE 2:

THE FULL DISTRIBUTION OF INCOME

What has been discussed so far concerns the top 10 percent of the income distribution, with no attention paid to the remaining 90 percent (other than the state average income). This incompleteness obviously limits the relevance of the conclusions drawn previously. This chapter discusses the task of estimating the full income distribution, including the lower quartiles, the bottom decile, and the Gini coefficient. This is a challenging task to perform because it implies the generalization of the Pareto approximation to the entire distribution, which, undeniably, is a strong assumption. Therefore, the estimates obtained are far from being perfect, probably need some improvement, but are certainly worth further investigation.

The estimation of the full income distribution can be done for each state, but not for the entire time-period 1913-2003. The latter has to be restricted to 1965-2003. This is because prior to 1965, the number of individual tax returns does not approximate the household population well.

What follows provides a basic understanding on the Lorenz curve, and explains how the assumption made about the Lorenz curve and its functional form obtains the three quartiles, the lower decile, and the Gini coefficient.

6.1 The Lorenz Curve and the Gini Coefficient

6.1.1 General Considerations

The Lorenz curve plots the cumulative percentage of the population number (horizontal axis) against the cumulative percentage of total income (vertical axis). Because both axes sum up to 100 percent, the Lorenz curve is represented in a square box that scales from 0 to 1. In the following hypothetical example, income inequality is highlighted by reading the graph this way: 75 percent of the population earns less than 50 percent of the total income, and 50 percent of the population less than 25 percent of total income.

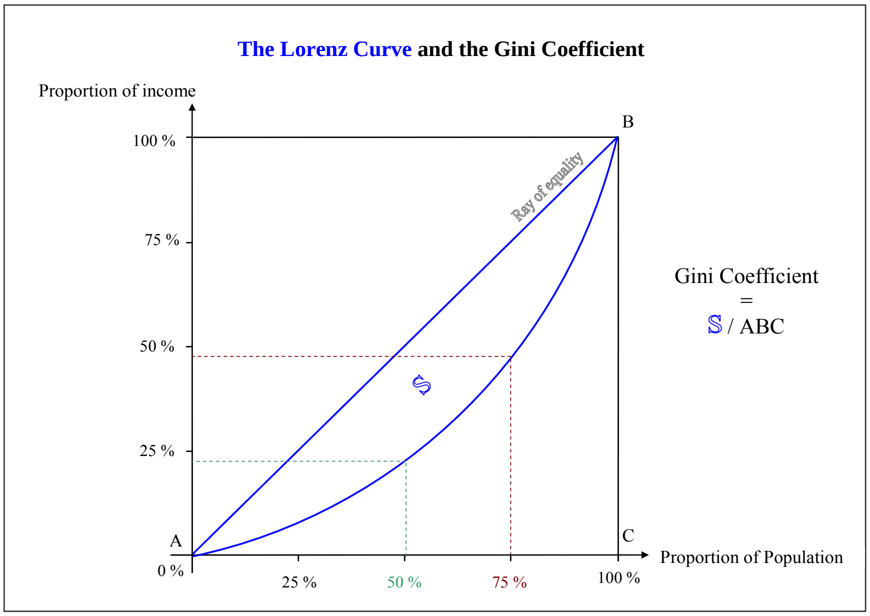


Figure 6.1 Lorenz Curve and Gini Coefficient

The so-called 'concentration area' corresponds to the surface 𝕊, bounded with the diagonal and the arc formed by the Lorenz curve (departing from the equality ray). Then, the Gini coefficient is defined as the ratio of two areas: concentration area 𝕊 over triangle ABC:

(6.1) $$0 < G_{ini} = \mathbb{S} / ABC < 1.$$

The Gini coefficient and its lack of a unit of measurement conveniently allows for temporal and spatial comparisons, as it is a *relative* measure of inequality; hence, the extensive use of the Gini coefficient. However, to depict a socio-economic phenomenon as complex as that of income inequality by a unique index is always restrictive. This is a limitation to bear in mind.

6.1.2 The Case of Crossing Lorenz Curves

Comparison problems arise when two Lorenz curves intersect such as depicted in the graph below (which is also hypothetical).

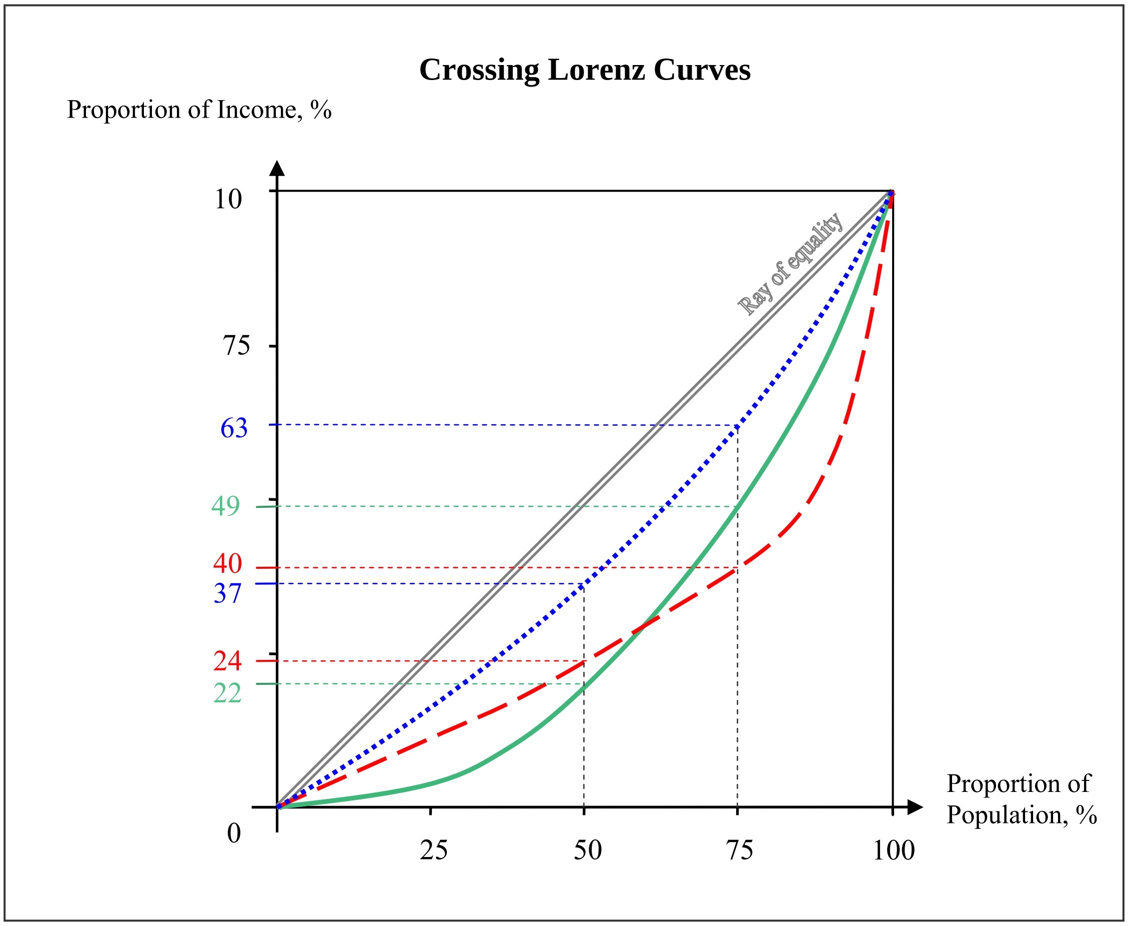


Figure 6.2 Crossing Lorenz Curves

Suppose that the initial income distribution is represented by the dashed line, so that 50 percent of the population earns 24 percent of the national income, and that 75 percent of the households earn 40 percent of total income. If a redistributive policy is implemented to flatten the dashed curve closer to the equality line, two different situations may arise. The first one is the dotted Lorenz curve that respects the principle of transfers,[1] that is, the transfer of income from rich to poor households reduces income inequality and shifts the initial Lorenz curve closer to the diagonal of equity. However, the second outcome might be the solid curve, which violates this principle. On the one hand, looking at the median[2] suggests that the initial distribution (50 percent of the population earning 24 percent of total income) is more equal than post-transfer distribution (50 percent of the households earning 22 percent of overall income). On the other hand, looking at the upper quartile suggests the contrary conclusion (75 percent of the households enjoy a higher income after the tax-transfers). In other words, crossing Lorenz curves introduces an unequivocal bias in any attempt of ranking inequality distributions.

Policy implications are at stake. Non-crossing Lorenz curves are used to justify redistributive policies transferring income from rich to poor individuals. Ideally, such top-to-bottom transfers shift the Lorenz curve inwards, towards the line of equality. However, the case of intersecting Lorenz curves clearly imposes a limit to the 'rich-to-poor reasoning' because a consistent ordering of income distributions in terms of inequality is no longer possible. This reveals the complexities of the income inequality phenomena.

---

[1] Creedy (1998, pp. 14-16).

[2] In the case of the Lorenz curve, the term 'median' truly means the share of median income in total income.

With these considerations in mind, we now need to choose which functional form may best approximate the Lorenz curve, before proceeding to the computation of the distribution itself.

6.2 <u>The Assumption of Pareto Distributions</u>

This section explains why the Pareto distributions are believed to be a relatively good fit for the Lorenz curve, and how the Gini coefficient is defined in that case.

6.2.1 The Functional Form of the Lorenz Curve

The Lorenz curve literature suggests different functional forms of the Lorenz curve: exponential function, power function, log-normal function, etc., all involving the estimation of a certain number of parameters. Among all these parametric forms, which one is relevant here? None is fully satisfactory, but the one chosen here fits the Pareto distribution of income, as assumed and used earlier with the estimations of fractiles within the top decile. Applying the Pareto distribution to the functional form of the Lorenz curve is an old and accepted tradition in the empirical literature.[3] Moreover, the Pareto functional form of the Lorenz curve limits to one the number of parameters to be estimated. Multiplying the number of parameters in the model specification, however

[3] This remark pertains to the case of Australian households living in urban areas, with data from the survey of consumer expenditures 1967-1968. See Kakwani and Podder (1973).

appealing it can be, raises complications in terms of interpretation of the results and the computation of the estimates.

With x being the cumulative proportion of the household population, and y being the cumulative percentage of income earned, the parametric function of the Lorenz curve is supposed to be of the form:

$$(6.2) \qquad \begin{cases} L(\delta): \quad y(x) = 1-(1-x)^{\delta}, & 0<\delta<1 \\ \delta = 1-\dfrac{1}{\alpha}, & \alpha>1 \end{cases}$$

How best to represent the Pareto form of the Lorenz curve? The probability density function associated with the Pareto distribution is the same as the probability density function (PDF) associated with the Pareto interpolation technique described in Chapter 3:

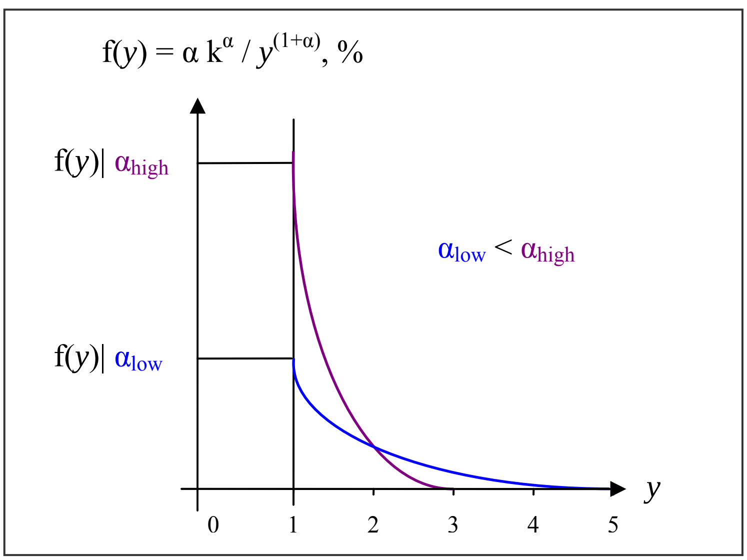


Figure 6.3 Pareto Probability Density Function

With y standing for income, α a scalar, and k, the minimum level of income in the distribution, $f(y) = \alpha\, k^{\alpha} / y^{(1+\alpha)}$, the Pareto PDF, captures well the 80-20 percent rule. $f(y)$, the "probability" or fraction of the population that earns a small amount of income per person (y), is rather high for low-income levels, then decreases steadily as income y increases: The higher the values of α, the wider the gap of inequality.

The parameter α corresponds to the coefficient calculated earlier in the estimation of fractiles of the upper decile, based on the Pareto interpolation method. There is one α coefficient attributed to each income bracket displayed in the *Statistics of Income* tables annually and at the state level.

Note that $L(\delta): y(x)$ when $x = 90$ does not correspond to AI90-100, the average income of the top decile. The former is a fraction expressed in percent, the latter is the income of the top 10 percent measured in constant dollars of 2003. $L(\delta): y(90)$ does not correspond either to S90-100, the share of fractile 90-100 in total income. It is the income 'share' of the threshold income TI90 that $L(\delta): y(90)$ corresponds to. Both are identically calculated.

In the case of France in the time-period 1900-1910, Piketty[4] used a Pareto coefficient of 2.6, which means an α coefficient of 1.625 as α = b / (b – 1). In the case of our panel data on the United States, the α coefficients calculated in Chapter 3 average to

[4] Piketty (2001, p. 619, note 2).

2.1 for all states and years, and display a very low standard deviation (0.57), as expected.[5]

6.2.2 The Gini Coefficient in the Case of the Pareto Distribution

The corresponding Gini coefficient, labeled G, is the concentration ratio applied to income distribution and is defined as:

$$G = 1 - 2 \cdot \int_0^1 [1-(1-x)^{\delta}]dx \tag{6.3}$$

Re-arranging and simplifying the terms yields a coefficient depending only on the parameter δ:

$$G = \frac{1-\delta}{1+\delta} \tag{6.4}$$

Clearly, the α coefficients will be used again to compute the Gini coefficients as $\delta = 1 - 1/\alpha$. As there are as many α coefficients as income brackets, we are able to average a value of α for each state in each year.

We now turn to the application to the United States from 1965 to 2003.

---

[5] This holds after exclusion of aberrantly high values of the Pareto coefficient associated with the AI99.99-100 fractile.

6.3 Applied to the Panel Data of the United States, 1965-2003

6.3.1 Lorenz Curves Skewed to the Right

Below are the Lorenz curves computed for Alabama and California in year 2000. Skewed to the right, these Lorenz curves depict the weight of top incomes in the distribution.

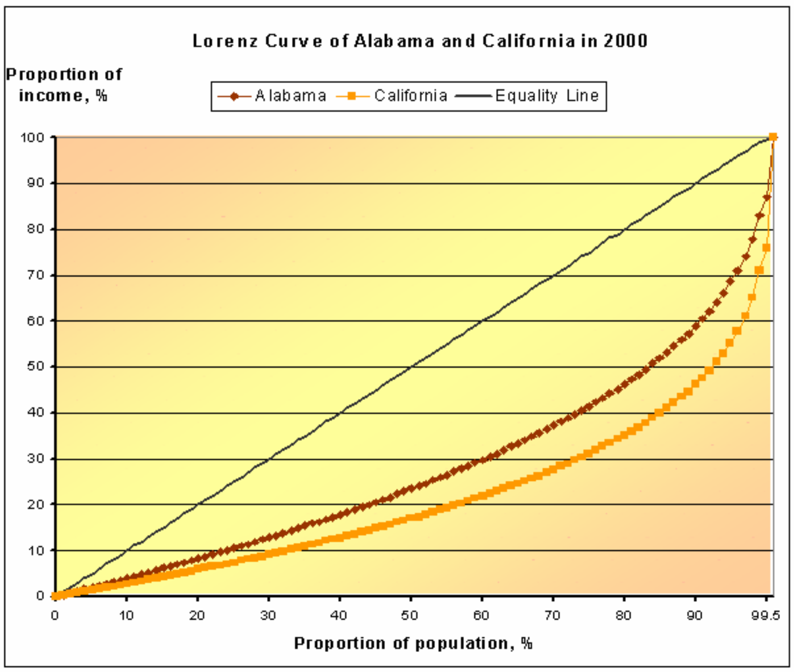


Figure 6.4 An Example of Lorenz Curves

The Gini coefficient is 0.443 for Alabama and 0.577 for California.

6.3.2 Inequality Regressions on Gini Coefficients

As earlier, we use GMM estimation to assess the correlation link between inequality and economic growth. Like Chapter 5, economic growth is measured by the growth rate of state average income per household. Unlike in Chapter 5, the inequality indicator is no longer a fractile but the Gini matrix. The regressions lead, once again, to a negative correlation between the two variables.[6]

However, the introduction of time fixed effects in the inequality equation leads to regression coefficients that display neither consistency in sign, nor strong significance, as shown in the figure below.

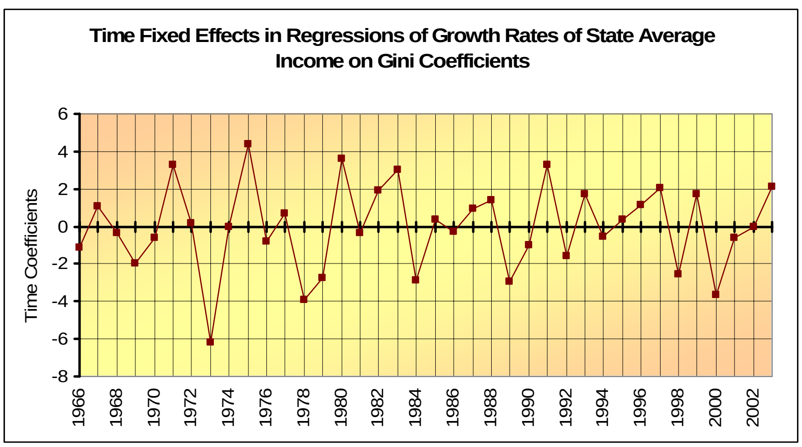


Figure 6.5 Time Fixed Effects in Growth Regressions with Gini Coefficients

An alternative way of examining the whole distribution is to consider 'opposite' percentiles, such as P10 vs. P90, P25 vs. P75, take their percentage share in

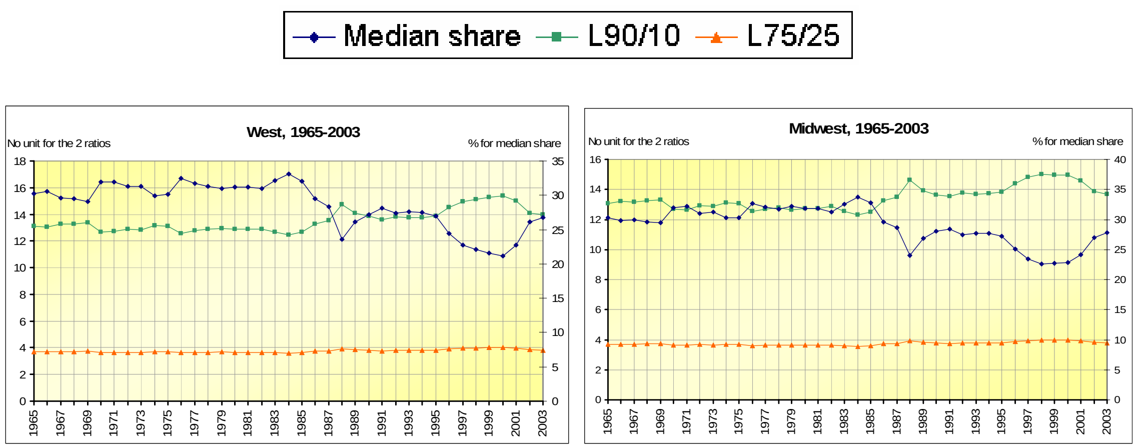
---

[6] The negative coefficient is statistically significant at the 10 percent confidence interval.

total income as displayed on the Lorenz curve (L10, L90, and L25, and L75), and calculate their respective ratios.[7] More details are available in the next section.

6.3.3 The Share of Median Income in Total Income, L90/10, and L75/25

The figure below summarizes the estimates of three indicators by region and annually from 1965 to 2003: 1) the percent share of median income in total income, 2) L90/10 or the ratio of the income share earned by 90 percent of the population over the income share of the bottom 10 percent of the population, and 3) L75/25, the ratio of the income share held by 75 percent of the population over the income share of the bottom 25 percent. The same legend applies to each region *i*'s curves.



---

[7] P as in percentile. L as in Lorenz.

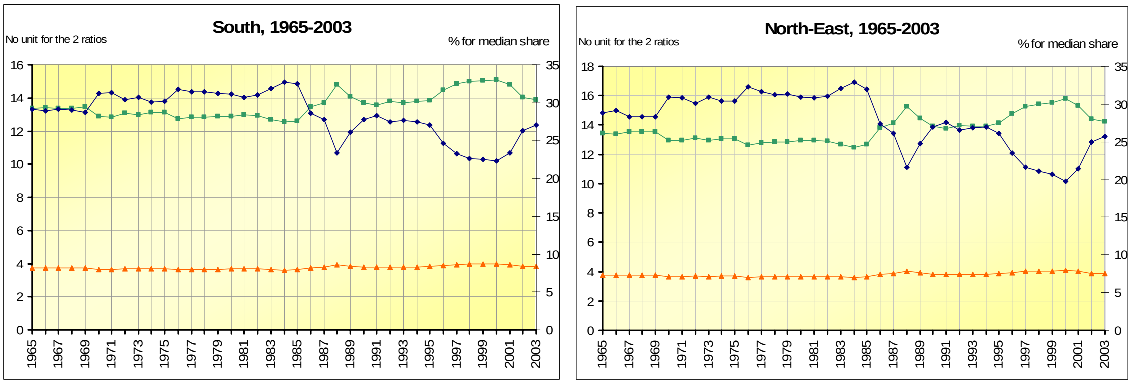


Figure 6.6 Median Share, L90/10, and L75/25 in the Census Regions

The primary axis has no units as it represents the two ratios L90/10 and L75/25, whereas the secondary axis (on the right of each graph) displays percentages for the median share in total income.

For all four regions, the two ratios remain fairly constant from 1965 to the mid 1980s, reach a sudden peak in 1988, and then display a very smooth cycle with a downward trend from 1988 to the mid 1990s, and an upward trend culminating in year 2000. The trends of L90/10 and L75/25 are clearly alike, suggesting that similar inequality dynamics occur both in income classes of low- and wide-inequality gaps (L75/25 and L90/10, respectively). However, this conclusion is questionable because the Pareto assumption on the functional form of the Lorenz curve is applied to the entire income distribution.

Among all states and years, the L75/25 ratio varies between 3.57 and 4.08, suggesting that the variable fluctuates within a very narrow band; likewise for L90/10, in a lesser extent. This apparent stability can be misleading. The actual fluctuations mean drastic changes in income shares, as suggested by Figure 6.7, showing the L75/25 curve in the case of the Northeast region.

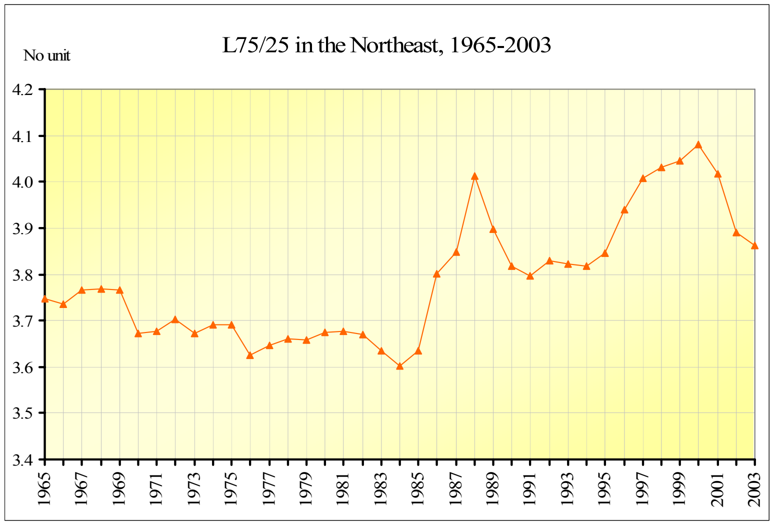


Figure 6.7 Fluctuations of Inequality Ratio L75/25, Northeast

Consider one state in the Northeast region, say Massachusetts. According to the Massachusetts' Lorenz curve, the wealthiest 25 percent of the household population (labeled L0-25) earned about 14.5 percent of the total state income in 1978. Meanwhile, the wealthiest 75 percent of the household (labeled L0-75) earned 52.9 percent of the state income, which means that ratio L75/25 was of 3.7 that year. The same variables in 2000 recorded the values of 7.4 percent of income share for L0-25, 30.8 for L0-75, and therefore 4.2 for ratio L75/25. With L75/25 increasing by 0.5 (= 4.2 – 3.7) 'only', the share of the wealthiest 25 percent almost shrank by a half, from 14.5 percent in 1978 to 7.4 percent in 2000, which is a drastic change. Meanwhile, the income share of the wealthiest 75 percent of the household population did not drop as sharply as the L25 share (from 52.9 percent of income in 1978 to 30.8 in 2000, which is far above 26.5 percent, or 52.9 / 2).

The same type of comments applies to the L90/10 ratio. In 1975, L90/10 coordinate slight exceeded 13.1 in the South. The same ratio in 1988 was 14.8. This increase from 1975 to 1988 does not seem to be much at the first glance. Considering the definition of the ratio (L90/10 = L0-90 / L0-10), it appears that L0-10, the share of the bottom decile in total income, decreased from 5.3 percent in 1975 to 4 percent in 1988. Expressed in dollars of 2003, the total income for all Southern states was about $75,897,000 in 1975, and 5.3 percent of that total represents slightly more than $4,030,000. Similarly, 4 percent of $119,427,000 in 1988 represents about $4,729,000. This means that the poorest 10 percent of the population experienced an overall increase by less than $700,000 from 1975 to 1988. From the top decile perspective, however, L0-90 is 69.6 percent in 1975, and 58.5 percent in 1988. Therefore, the richest 10 percent benefited from an increase by $17,088,000, from $52,824,000 in 1975 to $69,913,000 in 1988. The full calculation process is summarized in Table 6.1.

Table 6.1 Bottom and Top Deciles in the South from 1975 to 1988

South	**L90/10** (a)	**L0-10**, % (b)	**L0-90**, % (c)	**total income**,$ (d)	$**Y0-10** (b) * (d)	$**Y0-90** (c) * (d)
1975	13.12	5.31	69.60	75,896,796	4,030,120	52,824,170
1988	14.82	3.96	58.54	119,426,905	4,729,305	69,912,510
difference	1.70			43,530,108	699,186	17,088,340

In a nutshell, the income of the richest 10 percent increased 24 times faster than that of the poorest 10 percent in the same time[8], and this what an increase by 1.7 in L90/10 means.

Among the former studies released on this topic, one is conducted by the Center on Budget and Policy Priorities and the Economic Policy Institute (2000). The authors of the report used non-annual data from the Census Bureau's March Current Population Survey, and measured the inequality gap with the ratio between the richest quintile and the poorest quintile (Q80/20). The authors focused on two time periods: one short (1988-90 versus 1996-98), and one long (1978-80 versus 1996-98), but do not comment on the intermediary range 1978-80 versus 1988-90. The results shown with the L90/10 ratio estimation do not perfectly fit their short-term and long-term analyses, but the trends toward a wider inequality gap look alike.

---

[8] The number 24 comes from the following division: 17,088,340 / 699,186 = 24.4.

Chapter 7

CONCLUSION

The conclusion briefly summarizes the contents of the dissertation and broadens the subject with several suggestions as for what can be done in future research.

7.1 Summary

The main contribution of this dissertation is the construction of a new homogeneous set of panel data by state cross-sections and annually from 1913 to 2003, using the *Statistics of Income* publications by the U.S. Internal Revenue Service. This database represents well the top 10 percent of the income distribution, but data from other sources are used to account for average income. Meanwhile, the top decile database offers an alternative estimate of average income figures used by Barro and Sala-i-Martin to study the same topic.

In order to address the issue of income convergence across the United States over the long-run, three types of convergence are distinguished: 1) the β convergence of average income in comparison with the β convergence of income in the top decile (growth and inequality regressions), 2) the σ convergence (dispersion of average and top incomes), and 3) the convergence of top incomes towards the lower fractiles of the income distribution.

In the case of β convergence, we found evidence confirming conclusions towards convergence within the top decile, and more mitigated results for convergence among state average incomes. More particularly, β convergence based on average income does not seem to hold in recessionary phases of the business cycle. To compare growth regressions to inequality regressions, where inequality indicators are based on the top income series, the results showed that income inequality is positively correlated to the average income, and negatively correlated to economic growth rates.

In the case of σ convergence, the trend over time compares two groups of income. On the one hand, average income and income of the top percentile both recorded a decline in dispersion across states (except after the mid 1980s). On the other hand, the incomes of fractiles 90-95, and 95-99 percent were featured with a rise in dispersion across states.

Finally, the convergence (or divergence) of the top decile towards (or away from) the bottom decile is examined from 1965 to 2003. Similarly, the top and bottom quartiles are compared as well. The dispersion indicators of the lower layers of the income distribution were estimated by extending the Pareto assumption from the top decile to the full income distribution. What emerges from these estimates is that the top income shares did not grow faster than the low income shares from 1965 to 1984. This trend was totally reversed and reached a peak in 1988 in all states, then decreased again until 1985, and finally reached a local maximum in 2000.

7.2 Suggestions for Future Research

Further investigation may compare regional convergence of income to regional convergence of salaries and wages, as the IRS tables also reveal income composition by state. Section A.4 of the appendices suggests a beginning analysis towards that goal.

On the one hand, average income and income of the top 1 percent both record a sharp decrease in dispersion across states over time (except after the mid 1980s). On the other hand, the incomes of the top 90-95 and 95-99 percent feature an increase in dispersion across states over time. The analysis of such contrasting results remains to be done.

In Chapter 6, the assumption of the Pareto distribution applied to the entire income distribution is questionable. More assumptions on the functional form of the Lorenz curve need to be tested, inasmuch as the hypothesis influences the results and affects their economic interpretation to a certain extent.

Finally, the IRS tables record tax liabilities data consistently from 1913 to the present. Several aspects of the tax policies implemented over the past century could be explained with those data that remain to be computed.

Appendix 1

DATA SOURCES

A.1.1 Internal Revenue Service and *Statistics of Income*

All income levels of the top deciles were derived from the *Statistics of Income* (SOI). Listed below are the various publications in which the SOI tables were released, along with comments on the data availability when necessary.

- 1913-1915: "Annual Report of the Commissioner of Internal Revenue". The 1913, 1914, and 1915 tables provide the number of tax returns by size of income and by state. However, the dollar amount of the state total income accruing to each of the income classes was not released.
- 1916-1973, 1975-1981: "SOI Individual Income Tax Returns" of Tax Year n, Publication 79 (8-83), Basic Tables, Part 4 (State Data), first and last tables.
- 1944: State data are published by composition of income, but not by income-class interval (only U.S. aggregates are available by income class).
- 1974: U.S. Department of the Treasury, IRS "Supplemental Statistics of Income. Small Area Data, Individual Income Tax Returns", 1974. Publication 1008 (12-77). Basic Table 3 'Selected Income and Tax Items by Size of Adjusted Gross Income and by State and County', pp. 60-436.

- Between 1982 and 1987, the IRS interrupted the regular publications of state data by size of adjusted gross income (AGI). This interruption apparently corresponds to a change in methodology from sample estimation to master file computerization. Only the national aggregate by size of AGI is found in the "SOI Bulletin" of Tax Year n + 2, Fall, Publication 1136, "Selected Statistical Series", Table 2 'Individual Income and Tax by State, [Year n]'. The substitute data for these six years were found in the following issues:
    - 1982: "SOI Bulletin", Volume 5, Number 1, Summer 1985, 'Individual Income by Zip Code Area, 1979 and 1982'.
    - 1983: "Statistics of Income, Individual Tax Model for 1983", IRS.
    - 1984-1985: Bureau of the Census annual data extracted from the March Current Population Survey, household count for 21 income-class intervals for each state.
    - 1986-1987: "SOI Bulletin", Volume 10, Number 2, Fall 1990, 'Individual Income Tax Return Data by State, 1986-1988'.
    - 1988-1996: "SOI Bulletin" of Tax Year n + 2, Fall, Publication 1136, 'Selected Historical Data' (yellow), Table 2 'Individual Income and Tax Data by State and Size of AGI, [Year n]'.
    - 1997-2003: Data by state, by income size, and by income sources are available on the IRS website.

A.1.2 Bureau of Economic Analysis and State Personal Income

Unlike the IRS's Statistics of Income, the BEA's State Personal Income data are not displayed by income classes. The State Personal Income from the BEA are used instead when it comes to comparing the income level of the top decile to the average income. The BEA introduces some differences between the national aggregate and the state estimate of personal income:

> The main differences between the national income and product accounts (NIPAs) estimates of personal income and the State estimates of personal income stem from the treatment of the income of U.S. residents who are working abroad and the treatment of the income of foreign residents who are working in the United States. The national total of the State estimates of personal income consists of only the income earned by persons who live within the United States, including foreign residents working in the United States. The measure of personal income in the NIPAs is broader.

From 1929 to 2003, the national and state personal income data come from the Bureau of Economic Analysis. For data prior to 1929, Piketty and Saez[1] completed the personal income series by linking it to the 1913-1929 personal income series published by Kuznets (1941, 1945). The authors made downward adjustments over the time series so that the total personal income does not exceed the total tax return gross income. The contrary would "seem implausible: This would imply that non-filers have higher average incomes than filers."

I used Piketty and Saez's national series from 1913 to 1928 to estimate their state equivalent, with the income share of each state in the national income between 1913 and 1928 assumed to be proportional to the same state income share for year 1929. In

---

[1] Piketty and Saez (2001, p. 38, note 63).

other words, the state personal income estimates maintain the same proportion of the national aggregate[2] from 1913 to 1928 as the state shares of personal income represent in U.S. total income for 1929.

A.1.3 Bureau of the Census and Household Numbers

The number of households by state was published by the U.S. Census Bureau, in "The Demographic Trends in the 20th Century". The table provides us with the number of households by state for each decade of the 20th century (1900, 1910, 1920, and so on to 2000). The intercensal estimates of the household population are produced by the U.S. Bureau of the Census for each year between 1981 and 1989[3] and again between 1991 and 1998.[4]

The intercensal estimates have been retrieved from 1911 to 1979 by assuming for each decade a linear progression of the type $V_n = V_0 * (1 + r)^n$, where $n = 1$ to 9 and $r = (V_{10} / V_0)^{1/10} - 1$. Because the interpolated series assumes constant the magnitude of the annual change, this method of estimation is also called 'straight-line interpolation'. Note that it is the default option of the Bureau of Economic Analysis, when no other data source is available.

---

[2] The national aggregates come from Piketty and Saez's series (2004).

[3] The estimates are consistent with Current Population Report Series P25-1123, issued in October 1994.

[4] Census Bureau's "Estimates of Housing Units, Households, Households by Age of Householder, and Persons per Household of States: Annual Time Series, July 1, 1991 to July 1, 1998". Publication ST-98-51.

Labeled N*, the state estimates of households for 2001, 2002 and 2003 were obtained by inflating the Census state figure for year 2000 with the average growth rate of the household population over the 1990-2000 decade.

Before 1960, the data for Alaska and Hawaii are missing in the publications mentioned above. The missing data were replaced by the number of occupied dwelling units released in the following publications: (1) the 1950 Census of Housing[5] for both Alaska and Hawaii, (2) the 1940 Census of Population for Alaska in 1940, and (3) the 1940 Census of Housing, for Hawaii in 1940, 1930, and 1920.

Between 1913 and 1939, Alaska's number of households is assumed to be proportional to the share of Alaska's households in the U.S. total in 1940 (as 1940 is the earliest year this data is available). Similarly, estimates of N* for Hawaii between 1913 and 1919 are hooked to Hawaii's population share in the national aggregate in 1920.

The main issue relative to the series N* concerns the equation of one return per household. Because most of the time more than one tax return is filed per household, the initial time-series of households was scaled up so that the national aggregates fit Piketty and Saez series, keeping intact the state proportions.

A.1.4 Federal Reserve Bank of Minneapolis

The consumer price index data are published by the Federal Reserve Bank of Minneapolis for the time period 1800-2005.

---

[5] Census of Housing (1950, General Characteristics, Part 7).

Appendix 2

DATA HOMOGENEITY

The adjustments presented below add to the adjustments already mentioned in earlier chapters. All together, all adjustments performed in the database have been covered at some point in the text.

A.2.1 Alaska, Hawaii, D.C., and Other Areas

From 1913 to 1959, Alaska and Hawaii do not record reliable data due to (1) the excessive amount of missing values for early years of the time period, (2) to the lack of Census data in terms of household units prior to 1960 (when Hawaii was not a U.S. state yet, only a U.S. territory).

- Alaska

1921-1938 and 1943-1954: Alaska is included in Washington state. To separate Alaska from Washington state, we calculated *r*, the growth rate of the share of Alaska in the total of the two states between 1942 and 1955, assuming a linear progression of this share annually:

(A.1) $$r = (\text{Alaska share}_{1955} / \text{Alaska share}_{1942})^{1/13} - 1$$

Using this constant rate, it is straightforward to derive the annual values of Alaska's share:

(A.2) Alaska share_n = Alaska share_{1942} * $(1 + r)^n$, with n = 1943 to 1954

Defined as Alaska / (Alaska + Washington), the Alaska share contains only one unknown at this point: the numerator. After Alaska data are deduced, Washington is just the difference between the two states taken together and Alaska just estimated. Likewise for time period 1921-1938.

- Hawaii

The island became a United States territory in 1900, and was admitted as the 50th state in 1959.

1913-1959: Alaska and Hawaii do not record reliable data due to (1) the excessive amount of missing values for early years of the time period, (2) to the lack of Census data in terms of household units prior to 1960 (when Hawaii was not a U.S. state yet, a U.S. territory only). It is therefore possible to encounter unexplained results like the one for Alaska in 1957, with a drop in TI99.5 below its minimum level.

- District of Columbia

1961 and 1962: District of Columbia is included in Maryland for these two years. The 1960 and 1963 data are used to proxy the two variables (number of returns and adjusted gross income) for 1961 and 1962. However, the income class [$10,000-$15,000) in 1960 does not provide information for the intermediate income classes [$10,000-$11,000), [$11,000-$12,000), [$12,000-$13,000), [$13,000-$14,000), and [$14,000-$15,000) as 1963 does. The missing classes for 1960 where estimated by applying to 1960 the same percentage that each intermediate income class represents in the [$10,000-$15,000) total of 1963.

(A.3) $$V_{1963} = V_{1960} (1 + r)^3 \text{, where } r = (V_{1963} / V_{1960})^{1/3} - 1.$$

- Puerto Rico and the Virgin Islands

The United States took possession of Puerto Rico and the Virgin Islands in 1898, and 1917, respectively. Unlike the Virgin Islands, Puerto Rico became a Commonwealth on July 25, 1952 (Source: National Park Service, U.S. Department of the Interior).

1913-1951: Puerto Rico is included in Maryland.

1917-1951: Puerto Rico and Virgin Islands are included in Maryland.

1952: Puerto Rico is included in both Maryland (1st half of 1952) and New York (2nd half of 1952). Nothing is mentioned about the Virgin Islands.

1953: Puerto Rico and Virgin Islands are included in New York.

1954: For the first time, Puerto Rico is a separate entity in the State and Territory classification.

1954 on: Puerto Rico and Virgin Islands are grouped together and they either appeared separately or were added to the 'Other Areas' category.

- Panama Canal Zone

The United States Protectorate in the Republic of Panama started with the Treaty of 1903.

1913-1954: Panama Canal Zone is included in Florida.

A.2.2 Classes Grouped for Disclosure Purposes, 1917-1937

As the IRS tables display income by income brackets, it happened for 1917-1937 that several income classes, typically those at the very top of the ranking, were displayed with an income grouped together and tagged with a note specifying that is was done to conceal the identity of taxpayer. The number of returns in every income class was left visible, however. To retrieve the income data removed for disclosure purposes, the estimates were approximated based on a mid-point approach. Breaking it down, the estimation method takes the midpoint of each income class that fell under disclosure. Expressed in dollars, each midpoint was weighted by the number of returns as they appear in each income class, and then added together. The final estimate therefore represents in the classes-grouped total (which is given in the tables) the same share as one midpoint represents in the sum of midpoints.

Consider the following example made of hypothetical numbers. The table below shows an income of $20,000,000 as the total of two income classes grouped together.

Table A2.1 An Example of Classes Grouped

Net income classes in current $	Number of returns	Net income in $	Midpoint of the income class weighted by number of returns, in $	Estimated net income, in $
2,000,000 under 3,000,000	5	D	12,500,000	11,627,907
3,000,000 under 4,000,000	3	13,000,000	No need of …	… estimation
4,000,000 under 5,000,000	2	D	9,000,000	8,372,093
5,000,000 and over	1	7,000,000	No need of …	… estimation
Total classes grouped		20,000,000	21,500,000	20,000,000

The income class [$2,000,000 - $3,000,000) records five returns while that of $4,000,000 - $5,000,000, only two. Both classes fall under the protection of the taxpayer's identity. For those two income classes, we weighted the midpoint income by the number of returns ($12,500,000 = $2,500,000 * 5, and $9,000,000 = $4,500,000 * 2).

Relative to $20,000,000, which is the classes-grouped total displayed, the net income estimates should keep the same proportions as the midpoint figure in the total of $21,500,000 (obtained by adding $12,500,000 to $9,000,000). In other words, if the midpoint of the [$2,000,000 - $3,000,000) income class represents 58% of the midpoints total (58% = 100 * 12,500 / 21,500), then the net income estimate should also represent 58% of the classes-grouped total, namely: 58% * 20,000,000 = $11,627,907. Likewise for estimating the net income of the [$4,000,000 - $5,000,000) income class ($20,000,000 * $9,000,000 / $21,500,000 = $8,372,093). As this method refers to 'relative midpoint' for estimation, the sum of all estimates always equals the given classes-grouped total.

In a minority of cases[6], the income estimate calculated with the 'relative midpoint' approach did not lie within the interval of the corresponding income class. Whenever an estimates fell below the lower bound of the income class, it was replaced with the lower bound weighted by the number of returns. Whenever the estimate exceeded the upper bound, it was replaced with the upper bound weighted by the number of returns. Either way introduces a deviation of the sum of the estimates from the given classes-grouped total. One could have used instead the 'absolute midpoint' approach. However, the sum of the income estimates would have deviated further apart from the classes-grouped total.

The highest income class does not have an upper bound, e.g. [$5,000,000 and over). In that case, we used the 'relative mid-point' approach with an assumption on

[6] The minority of cases lies around 8% of all income classes displayed in the IRS tables from 1917 to 1937.

the upper bound, but replaced the estimate of the highest income class by the residual term. Doing so eliminated the deviation of the sum of the estimates from the classes-grouped total.

Whatever the estimation method is, none of them can correct for the errors coming from the IRS itself. For instance, New Hampshire in 1918 recorded classes-grouped that were displayed with a total income amount that cannot be consistent with the income classes it corresponds to, as shown in the table below.

Table A2.2 A Typo in New Hampshire Table, 1918

New Hampshire, 1918 Net income classes, $	Number of returns	Net income, in $
150,000 under 200,000	1	D
200,000 under 250,000	2	D
Classes grouped	---	739,319

Indeed, the total of classes grouped ($739,319) exceeds the sum of the upper bounds weighted by the number of returns (739,319 > 700,000 = 200,000 + 500,000). Likewise for Nevada 1927 (Table A2.3):

Table A2.3 Another Typo in IRS Tables

Nevada, 1927 Net income classes, $	Number of returns	Net income $
60 under 70	1	D
70 under 80		
80 under 90	2	D
90 under 100	1	D
100 under 150	1	D
150 under 200		
200 under 250	1	D
Classes grouped	---	820,937

The highest estimate, that is, the sum of the upper bounds weighted by the number of returns (750,000 = 70,000 + 180,000 + 100,000 + 150,000 + 250,000), lies beneath the level of classes grouped ($820,937).

Overall, 92 out of 1,071 classes grouped (51 states times 21 years between 1917 and 1937), had to be 'midpoint' adjusted. This represents around 8% of all income classes during this period.

There is no net income data by size of net income and by state for years 1913, 1914, and 1915. Only the number of returns is available. The net income estimates were derived from the same estimation approach as the classes grouped estimates were based on.

Appendix 3

NOTES ABOUT THE IRS TABLES

Returns with adjusted gross income can be both taxable and non-taxable. Returns with no adjusted gross income are non-taxable. Adjusted gross income means gross income minus a series of allowable deductions (e.g. trade and business deductions, expenses of travel and lodging in connection with employment, losses from sales or exchanges of property, etc.) Should these allowable deductions exceed the gross income, there is an adjusted gross deficit.

1917-1928, 1943 and 1945-1954: Returns with no net income (1917-1928, and 1943) or no adjusted gross income (1945-1954) were not included in the state totals, but were in the national aggregate. Why? According to the IRS, returns with adjusted gross deficit were too few and the sample variability too great, to permit their presentation on a state basis. For consistency purposes, I subtracted them from the national aggregate for both variables: number of returns and income amount.[7]

[7] The operation is a substraction for the number of returns, but an addition for income amounts, as a deficit is a negative number.

1913-1916: No net income data are available at the state level. Income brackets start at [$3,000 - $4,000) except for 1913, with [$2,000 - $3,333.33) being the lowest income class.

1946-1952: Non-taxable returns of AGI of $4,000 and over are all included in the [$4,000 - 5,000) income class.

1943 and 1945: Non-taxable returns of AGI of $1,000 and over are all included in the [$1,000 - 2,000) income class.

1942: The 1942 table does not detail Forms 1040A by income classes and displays the state aggregate only. Therefore, the 1040A forms cannot be used for the calculation of fractiles, even though their value appears in the state totals.

From 1943 on, money amounts for net income (1943) and AGI (1944-2003) are in $1,000.

1962: Delaware data appear in the national aggregate, but not separately.

Appendix 4

GROWTH RATES AND WAGES

This section briefly exposes how inter-state inequality co-varied with top wages in the recent years. The idea is to see whether the regional convergence that sometimes occurs with income also occurs with salaries and wages.

A.4.1 Composition of Income in IRS Tables

The IRS tables also provide information on income composition at the state level. The IRS adjusted gross income (AGI) includes the following items: salaries and wages, interests (taxable and non-taxable), dividends, entrepreneurial income (i.e. business and farm incomes), pensions and annuities, and realized capital gains. Using the data from 1997 to 2001, the average incomes (in 2003 dollars) accruing to AI90-95, AI95-99, and to the top percentile, AI99-100, were broken down into each income source.

A.4.2 Initial Wage Levels versus Growth Rates

Six states were selected so that both urbanized and rural states are represented: California, Texas and New York vs. Kansas, Iowa and Nebraska.

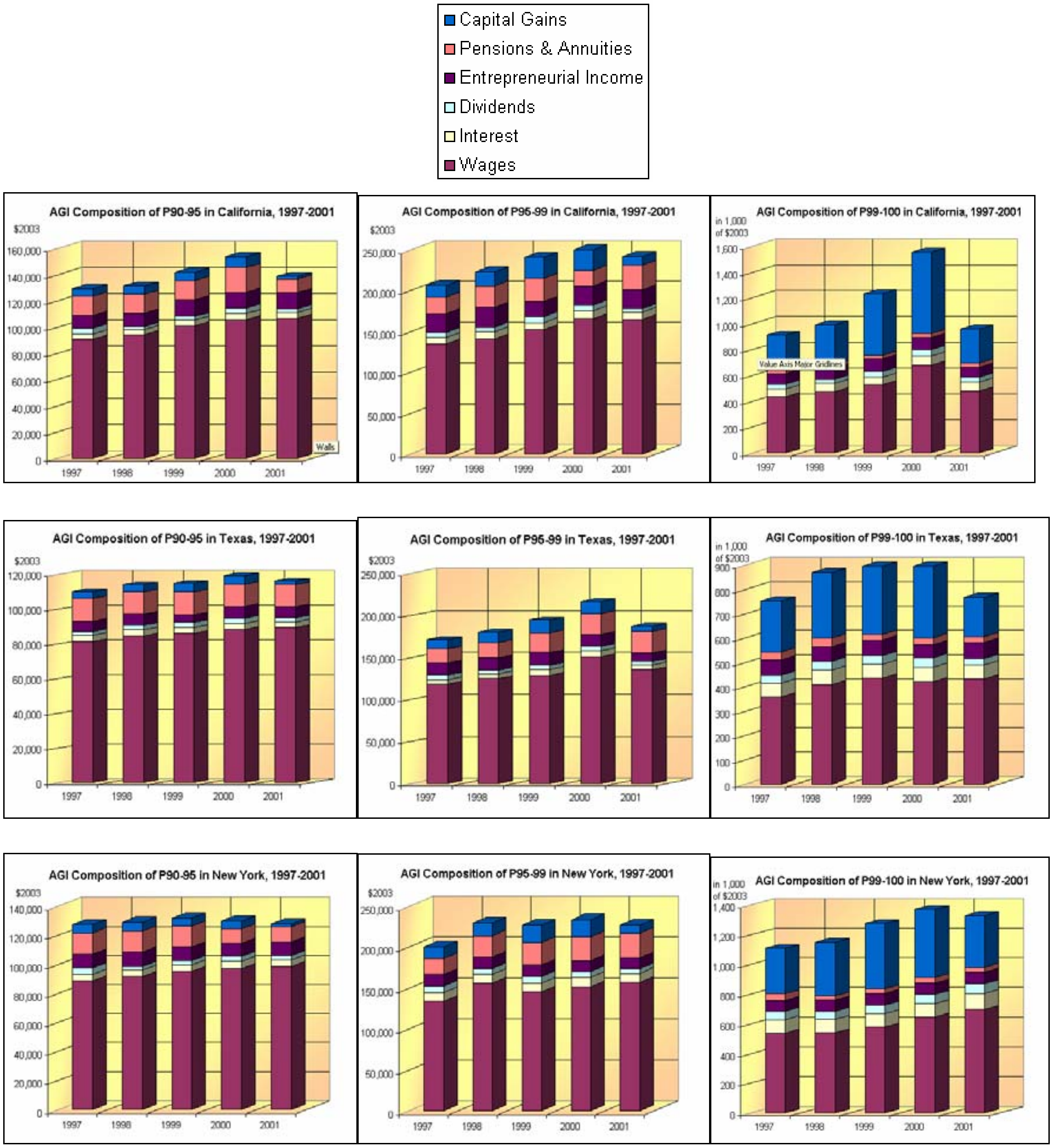


Figure A4.1 Income Composition in California, Texas and New York

In all six states, the top percentile clearly distinguishes itself from the bottom two fractiles. In the top 1 percent, wages represent a lesser share of total income than in the other two layers of the distribution, and capital gains a more substantial share. While capital gains fluctuations seem to be consistent with those of total income, wages do not necessarily vary in the same direction as total income does.

Even though there are differences in levels between the ‘rural’ states and the ‘richer ones’, wages appear to be more stable than total income, and much more stable than capital gains in all of them.

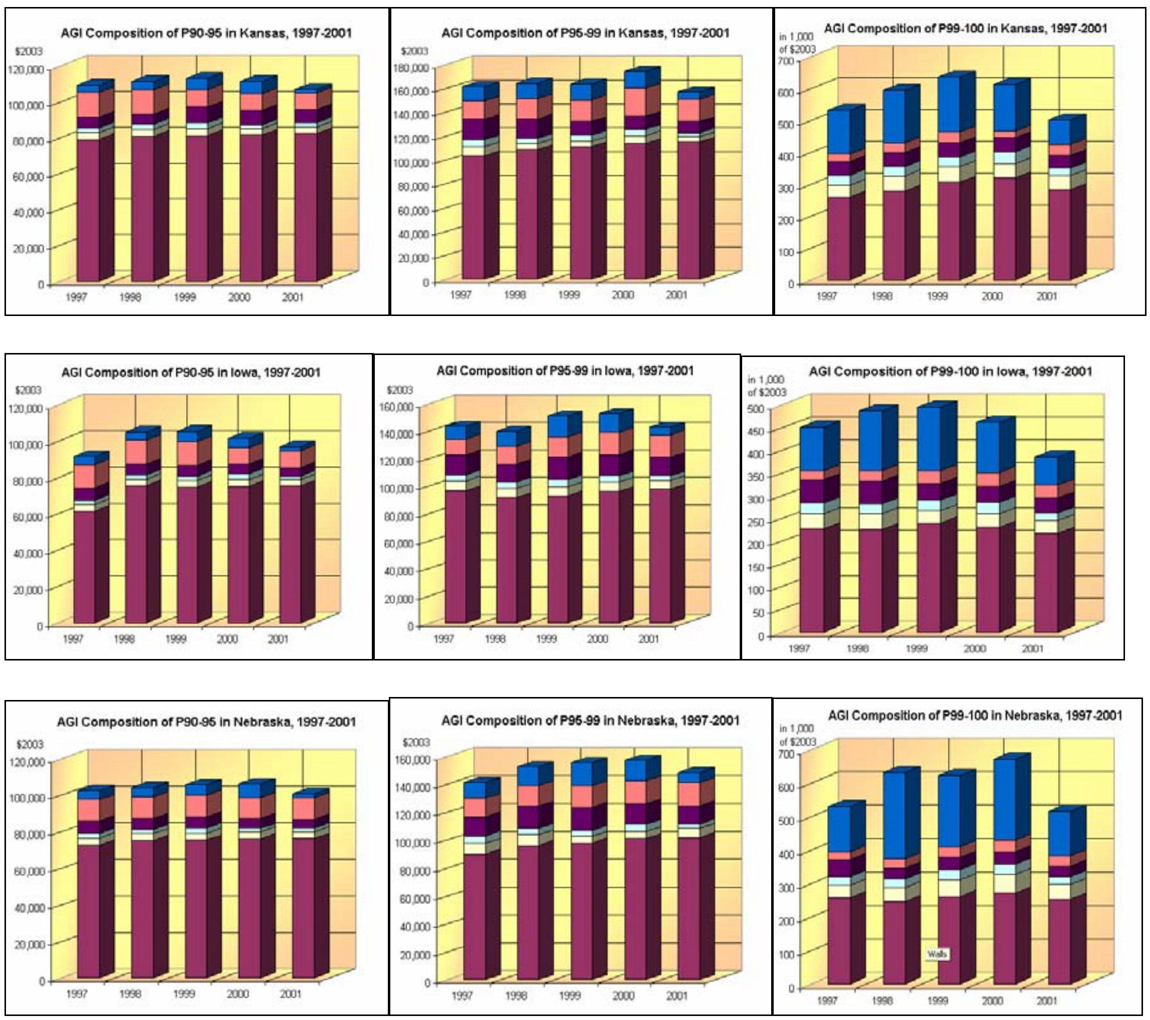


Figure A4.2 Income Composition in Kansas, Iowa and Nebraska

In sum, wages seem to reach higher levels in states displaying high inequality, and follow more or less closely the fluctuations of capital gains in poorer states than in richer ones.

REFERENCES

Amos Orley M, Jr. “Unbalanced regional growth and regional income inequality in the latter stages of development.” *Regional Sciences Urban Economics*. Issue 18. pp. 549-566. 1988.

Amos Orley M, Jr. “An Inquiry into the Causes of Increasing Regional Income Inequality in the United States”. *The Review of Regional Studies*. Vol. 19.No. 2. pp. 1-12. Spring 1989.

Anselin Luc. Spatial Econometrics. Kluwer. 1988.

Atkinson Anthony B. “Is Rising Income Inequality Inevitable? A Critique of the Transatlantic Consensus”. WIDER Annual Lecture. 1999.

Atkinson Anthony B. “Top Incomes in the United Kingdom over the Twentieth Century.” University of Oxford Discussion Paper in *Economic and Social History*. No. 43. 2002.

Atkinson Anthony B. and Brandolini Andrea. “The Panel-of-Countries Approach to Explaining Income Inequality: An Interdisciplinary Research Agenda.” Mannheim 2003. April 2003.

Barro Robert J. “Inequality and Growth in a Panel of Countries”. *Journal of Economic Growth*. Vol. 5. pp. 5-32. March 2000.

Barro Robert J., Sala-i-Martin Xavier. Economic Growth. McGraw-Hill. 1995.

Barro Robert J., Sala-i-Martin Xavier. “Economic Growth and Convergence Across the United States.” *National Bureau of Economic Research*. Working Paper 3419. August 1990.

Barro Robert J., Sala-i-Martin Xavier, Blanchard Olivier, Hall Robert E. “Convergence Across States and Regions.” *Brookings Papers on Economic Activity*. Vol. 1991. No. 1. pp. 107-182. 1991.

Braun D. “Income Inequality and Economic Development: Geographic Divergence”. *Social Sciences Quarterly*. Vol. 72. pp. 520-536. 1991.

Buchanan Mark. “Wealth Distribution and the Role of Networks.” Mimeo. 2002.

Center on Budget and Policy Priorities and Economic Policy Institute. Pulling Apart. A State-by-State Analysis of Income Trends. January 2000.

Champernowne D.G., and Cowell F.A. Economic Inequality and Income Distribution. Cambridge University Press. 1998.

Chotikapanich Duangkamon, Griffiths William E. “Averaging Lorenz Curves”. Not published. Posterior to 2003.

Chusseau Nathalie. “Les inégalités de salaire et de revenus dans l’Union européenne : où en sommes-nous ?” Conférence internationale Médée et Pôle Jean Monnet. *The Enlargement of the European Union : What are the Stakes and Potential Effects?* November 2004.

Coughlin and Mandelbaum. “Why have state per capita income diverge recently? ” *Review*, The Federal Bank of Saint-Louis. Vol. 70(5). pp. 24-36. 1988.

Crain, Mark W. “Institutions, Policy, and the Performance of American State Economies.” Chapter 2 in Volatile States. The University of Michigan Press. 2003.

Dall’erba Sandy, Le Gallo Julie. “Dynamique du processus de convergence régionale en Europe.” *Région et Développement*. Issue 21. 2005.

Danziger and Gottschalk. Introduction. In: Danziger and Gottschalk (eds) Uneven Tides: Rising Inequality in America. Russell Sage Foundation, New York. 1993.

Fan Cindy C., Casetti Emilio. “The spatial and temporal dynamics of U.S. regional income inequality, 1950-1989.” *The Annals of Regional Science*. Vol. 28. No. 2. pp. 177-196. 1994.

Feenberg Daniel R., Poterba James M. “Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns.” Tax Policy and the Economy edited by J. Poterba. NBER. MIT Press. Vol. 7. pp. 145-177. 1993.

Feenberg Daniel R., Poterba James M. “The Income and Tax Share of Very High Income Households, 1960-1995.” *National Bureau of Economic Research*. Working Paper 7525. February 2000.

Goldsmith Selma, Jaszi George, Kaitz Hyman, Liebenberg Maurice. "Size Distribution of Income Since the Mid-Thirties." *Review of Economics and Statistics*. Vol. 36(1). 1954.

Goolsbee Austan. "What Happens When You Tax the Rich? Evidence from Executive Compensation." *Journal of Political Economy*. Vol. 108(2). pp. 352-378. 2000.

Hare Timothy S. Political Economy, Spatial Analysis, and Postclassic States in the Yautepec Valley, Mexico. Doctoral dissertation. Albany, State University of New York. Department of Anthropology. 2001.

Haining Robert. Spatial Data Analysis, Theory and Practice. Cambridge. 2003.

Kakwani N.C. and Podder N. "On the Estimation of Lorenz Curves from Grouped Observations". *International Economic Review*. Vol. 14, No. 2. pp. 278-292. June 1973.

Krugman Paul. Development, Geography and Economic Theory. London: MIT Press. 1995.

Krugman Paul. "Urban Concentration: the role of increasing returns and transport costs". *International Regional Science Review*. Issue 19. pp. 5-30. 1996

Krugman Paul. "What's new about the New Economic Geography?" *Oxford Review of Economic Policy*. 14. 7-17. 1998.

Kuznets Simon. National Income and Its Composition, 1919-1938. National Bureau of Economic Research. Volumes 1 and 2. New York. 1941.

Kuznets Simon. Shares of Upper Income Groups in Income and Savings. National Bureau of Economic Research. 1953.

Kuznets Simon. "Economic Growth and Income Inequality". *American Economic Review*. Vol. 45. No 1. pp. 1-28. March 1955.

Kuznets Simon. "Quantitative Aspects of the Economic Growth of Nations: III. Industrial Distribution of Income and Labor Force by State, United States, 1919-1921 to 1955". *Economic Development and Cultural Change*. Vol. 6. No 4. Part 2. pp. 1-128. July 1958.

Lampman R.J. "The Share of Top-Wealth Holders in National Wealth, 1922-1956." *National Bureau of Economic Research*. Princeton University Press. 1962.

Lindert Peter. "Three Centuries of Inequality on Britain and America." Handbook of Income Distribution. Atkinson A. and Bourguignon F. ed. North-Holland. pp. 167-216. 2000.

Morrisson Christian. “Historical Perspectives on Income Distribution: the Case of Europe”. Handbook of Income Distribution. Atkinson A. and Bourguignon F. ed. North-Holland. pp. 216-260. 2000.

Park, Thae S. “Comparison of BEA Estimates of Personal Income and IRS Estimates of Adjusted Gross Income.” *Survey of Current Business*. November 2000. pp. 7-13

Peach James. “Regional Income Inequality Revisited: Lessons from the 100 Lowest-Income Counties in the United States”. Chapter 15 in Inequality, Racial Institutionalist Views on Race, Gender, Class, and Nation. 1996.

Perloff Harvey S., Dunn Edgar S., Jr., Lampard Eric E., Muth Richard F. Regions, Resources, and Economic Growth. Bison Book 319. 1967.

Piketty Thomas. Les hauts revenus en France au XXème siècle. Inégalités et redistributions 1901-1998. Grasset. 2001.

Piketty Thomas. “Income inequality in France, 1901-1998”. *Journal of Political Economy*. Vol. 111, pp. 1004-1042. 2003.

Piketty Thomas, Saez Emmanuel. “Income Inequality in the United States, 1913-1998”. *Quarterly Journal of Economics*. Vol. 118, pp. 1-39. 2003.

Piketty Thomas, Saez Emmanuel. “Income Inequality in the United States, 1913-2002”. November 2004. Longer updated version, forthcoming in A.B. Atkinson and T. Piketty eds., Oxford University Press.

Rasche R.H., Gaffney J., Koo A.Y.C., and Obst N. “Functional Forms for Estimating the Lorenz Curve”. *Econometrica*. Vol. 48. No. 4. pp. 1061-1062. May 1980.

Rowley, Redman and Angle. The Rapid Rise in Income Inequality in the 1980’s: Sources and Prospects. U.S. Department of Agriculture. Staff Report AGES 9104. 1991.

Sandretto René. Rémunérations et Répartition des Revenus. Hachette. 1994.

Slemrod Joel. “High Income Families and the Tax Changes of the 1980s: the Anatomy of Behavioral Response.” Empirical Foundations of Households Taxation edited by M. Feldstein and J. Poterba. University of Chicago. 1996.

Slemrod Joel, Bakija Jon. “Does Growing Inequality Reduce Tax Progressivity? Should it?” *National Bureau of Economic Research*. Working Paper 7576. 2000.

Jesper B. Sørensen. “The Use and Misuse of the Coefficient of Variation in Organizational Demography Research”. Unpublished version posterior to 2000.

US Department of Commerce, Census Bureau. *March Current Population Survey, Annual Social Economic Supplement*. Non-published. 1984 and 1985 data.

US Department of Commerce, Bureau of Economic Analysis. "Regional Economics Accounts, State Personal Income". http://www.bea.gov/bea/regional/statelocal.htm

US Department of the Treasury, Internal Revenue Service. *Annual Report of the Commissioner of Internal Revenue.* Annual Publication. 1913-1915.

US Department of the Treasury, Internal Revenue Service. *Statistics of Income, Individual Income Tax Returns.* Annual Publication. 1916-1973, 1975-1981.

US Department of the Treasury, Internal Revenue Service. *Supplemental Statistics of Income. Small Area Data, Individual Income Tax Returns*. Publication 1008 (12-77). 1974.

US Department of the Treasury, Internal Revenue Service. *Statistics of Income Bulletin, Individual Income by Zip Code Area, 1979 and 1982.* Volume 5, Number 1, Summer 1985.

US Department of the Treasury, Internal Revenue Service. *Statistics of Income, Individual Tax Model for 1983*.

US Department of the Treasury, Internal Revenue Service. *Statistics of Income Bulletin*. Tax Year n + 2, Fall, Publication 1136, "Selected Statistical Series". 1988-1998.

US Department of the Treasury, Internal Revenue Service. *SOI Tax Stats - Individual Tax Statistics - State Income*. 1997-2003. Mimeo

Williamson Jeffrey G. "Regional Inequality and the Process of National Development: A Description of the Patterns". *Economic Development and Cultural Change*. Vol. 13. No 4. Part 2. pp. 1-84. July 1965.

Williamson Jeffrey G. Lindert Peter. American Inequality – A Macroeconomic History. Academic Press. (362 p.) 1980

Wolff Edward. Top Heavy – The Increasing Inequality of Wealth in America. The Twentieth Century Fund. 1994.