A comparison of Latin America inequality data sets

By

Delfina Rossi

The University of Texas Inequality Project

Lyndon B. Johnson School of Public Affairs

The University of Texas at Austin

May 3, 2016

UTIP Working Paper 72

Abstract

This paper compares major inequality data sets for Latin America in respect of coverage and values. It first compares the Socio-Economic Database for Latin America and the Caribbean (SEDLAC) with the Estimated Household Income Inequality (EHII) data set of the University of Texas Inequality Project. Then, it presents a comparison of coverage and values for the Latin American region for three additional inequality data sets that are intended to present consistent coefficients that can be compared directly across countries and time: the Luxembourg Income Studies (LIS) summary statistics, the OECD income inequality statistics, and the World Bank's World Development Indicators (WDI).

Key Words: Inequality, Latin America, Gini Coefficient.

I. Introduction

This paper follows Galbraith et al. (2015b) by adding in a comparison between the Estimated Household Income Inequality (EHII) data set of the University of Texas Inequality Project and the Socio-Economic Database for Latin America and the Caribbean (SEDLAC) of the Economic Commission for Latin America and the Caribbean. Section III expands the comparison by including comparisons to the World Bank World Development Indicators (WDI), the summary tables of the Luxembourg Income Studies (LIS) and the Latin American measures presented in the income inequality measures published by the Organization for Economic Cooperation and Development (OECD).

II. SEDLAC and EHII comparison

Like LIS, SEDLAC is a non-official regional income inequality data set constructed from household surveys, in this case by the staff of CEDLAS (Universidad Nacional de La Plata) and The World Bank's LAC poverty group (LCSPP). SEDLAC is the only harmonized data set realized at the regional level outside the European Union and the OECD countries.

SEDLAC provides statistics on poverty, inequality and other socio-economic indicators for 24 Latin American and Caribbean countries. The statistics are computed from household survey micro data using homogeneous methodology (data permitting) and they are updated periodically.

SEDLAC reports different measures of inequality. Regarding Gini coefficients, SEDLAC provides with thirteen distinctive Gini coefficients and their variations. SEDLAC Gini1 reports the distribution of a range of household and per capita income variables across a range of geographic units. In order to compare SEDLAC with EHII, I restrict myself to the measure of total household income inequalities at the national level. For example, for Argentina SEDLAC provides data on inequality in Buenos Aires area, 15 main cities and 28 cities, and then nationally; here we use only the national coefficient. Finally, when SEDLAC provides Gini coefficients per semester or per quarter, I calculated the average to obtain an annual Gini coefficient.

EHII data set is a panel of estimated annual Gini coefficients at the national level derived from measured industrial pay inequality across manufacturing industries (UTIP-UNIDO) and the share of the population employed in the manufacturing sector. EHII was introduced by Galbraith and Kum (2005) and updated and expanded by Shams in 2014. EHII covers 149 countries from 1963-2008. It is a gross household income inequality concept. In Latin America, as Table 1 shows, SEDLAC has 288 observations for 23 countries beginning in 1974. EHII has 536 observations for the same countries and time period, and 618 observations for the same countries over the full period from 1963 to 2008. There are 109 observations that overlap exactly with respect to both country and year.

SEDLAC coverage of 23 countries includes Guyana and Panama, which are not covered by EHII. EHII covers 26 countries, including Barbados, Bahamas, Puerto Rico, Cuba, and

Trinidad and Tobago, which are not covered by SEDLAC. Figures 1 and 2 in Appendix I show the coverage of each data set.

Table 1: SEDLAC coverage

Data set	Total Observations	Countries Covered	Years Covered	Observations covered by both data sets	EHII Observations Matched by Countries and Years Covered	EHII Observations Matched by Countries, 1963-2008
SEDLAC	288	23	1974-2013	109	536	618

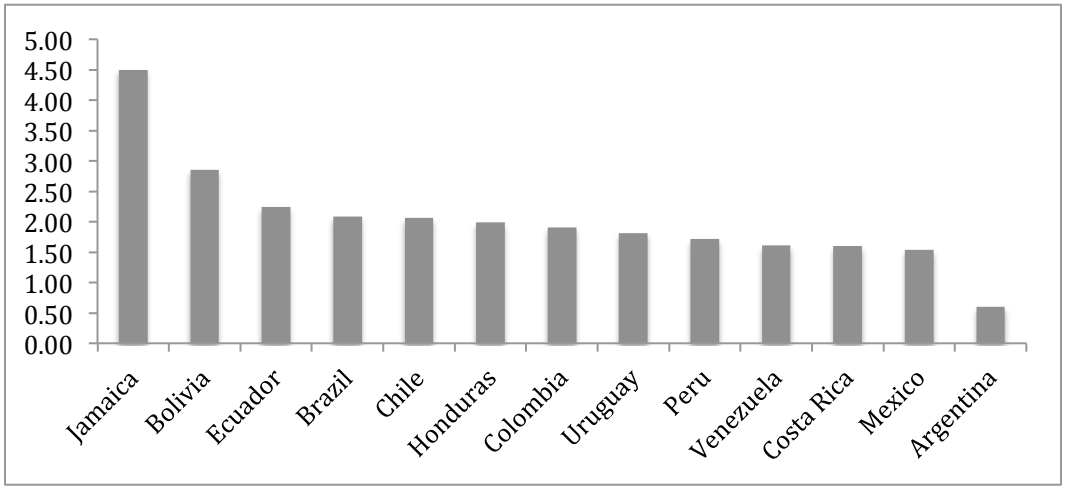
Comparing Gini values for data sets that overlap only partially is a delicate exercise. Table 2 summarizes the divergences of Gini coefficients between SEDLAC and EHII for the observations that directly overlap. The standard deviation of divergence measures the joint standard deviation across all the 109 overlapping observations. The average standard deviation of country divergence from EHII measures the mean of the dispersion of the differences within each country. EHII and SEDLAC differ in average by 4.86 Gini points with a standard deviation close to 2 Gini points.

Table 2: SEDLAC-EHII differences

Data set	Years covered	Mean Divergence from EHII	Standard deviation of divergence from EHII	Average standard deviation of divergence from EHII
SEDLAC	1974-2013	4.85	2.82	2.04

Figure 1 shows the standard deviations of the differences between EHII and SEDLAC Gini coefficients for each country when there are overlapping observations. Jamaica has the biggest standard deviation of differences from EHII, while Argentina has the lowest. We consider that overall the correspondence between the two data sets is quite good, which lends some confidence to the use of EHII for measures in countries and years when the micro-based SEDLAC data are unavailable.

Figure I: Standard Deviation of Differences between EHII-SEDLAC



III. Comparing different data sets in Latin America

This section compares coverage and values of the major international data sets on Latin America and the Caribbean. Taking the EHII as the baseline, I compare the EHII values with inequality measures from SEDLAC, WDI, LIS and OECD.

LIS covers Brazil, Colombia, Guatemala, Mexico, Peru and Uruguay for the years 2009, 2010 and 2011. The OECD data set reports on Mexico and Chile – the only Latin American members of the OECD – between 1984 and 2012. EHII, WDI and SEDLAC are more comprehensive. The WDI has 23 countries with observations from 1979 until 2012. As explained above, SEDLAC covers 23 countries from 1974 until 2013, while EHII covers 26 countries of the region from 1963 until 2008.

Table 3 summarizes the coverage of each data set. Figure 4 in Appendix II shows graphically the coverage of each data set.

Table 3: Inequality data sets for Latin America - Descriptive statistics

Data set	N. Observations	Mean	Std. Dev.	Min.	Max.
EHII	618	45.61	3.77	30.60	55.37
SEDLAC	288	49.73	5.21	36.5	71.3
LIS	20	47.48	2.19	43	52.3
OECD	10	50.13	2.33	46.30	53.08
WDI	331	51.44	5.44	34.42	69.17

Figure 5 in Appendix II summarizes the number of observations per year of each data set, showing how extend the EHII data set is. EHII is the only data set reporting on Latin American inequalities from 1960s on. Figure II exhibits that the movement of inequalities is similar across the five data sets, and EHII tends to report lower values.

Table 4: Differences with respect to EHII – Overlapping observations

Data set	N. Observations	Mean	Std. Dev.	Min.	Max.
SEDLAC	109	4.86	2.82	0.18	16.41
LIS	11	3.87	2.68	.64	8.75
OECD	4	5.56	2.27	3.70	8.80
WDI	140	6.28	4.00	.01	19.28

Table 4 reports the absolute difference of EHII with respect to the four other data sets. On average EHII reports Gini coefficients that are 5 points lower than the other data sets. The standard deviation of the difference between EHII and WDI shows that WDI is constructed from different sources and thus the volatility is much higher. With respect to the other sources, the standard deviation is similar, most probably because LIS, OECD and SEDLAC are based on household surveys.

IV. Conclusion

Galbraith et al. (2015b) show that EHII is highly consistent with LIS, OECD and the EU-SILC dataset of the European Union. Similarly, I find that EHII is consistent with SEDLAC, despite the different measurement and estimation techniques underlying the two data sets.

When comparing all the data sets available for the Latin American and the Caribbean region, EHII is on average 5 points below the other data sets, a discrepancy mainly attributable to the very high maximum values reported for certain countries and years in the other data sets. Since EHII is based on a projection of industrial wage inequalities using data that are predominantly taken from advanced countries, it is not altogether surprising that survey-based income inequality values in Latin America should often be higher. This factor should be taken into account in using the EHII data for Latin America.

That said, however, the movements of EHII seem broadly to correspond with the movements of inequality observed in other data sets, so that it is reasonable to use the EHII numbers as representations of the trends in inequality in the region, taking advantage of the much larger number of observations available in EHII than in the other measures.

Figure 6 offers country-by-country graphs of the different data sets between 1963-2011. In sum, EHII offers more coverage for countries and years in a consistent manner and therefore can be considered a useful comparative tool to analyze inequality in Latin American and the Caribbean with a deep historical coverage.

V. References

Deininger, Klaus and Lyn Squire. 1996. "A New Data Set Measuring Income Inequality." World Bank Economic Review 10(3): 565-591.

Galbraith, James K. and Hyunsub Kum. "Estimating the Inequality of Household Incomes: A Statistical Approach to the Creation of a Dense and Consistent Global Data Set." Review of Income and Wealth (1): 115-143.

Galbraith, James K. and Hyunsub Kum. 2003. "Inequality and Economic Growth: A Global View Based on Measures of Pay." CESifo Economic Studies 49(4): 527-556.

Galbraith, James K., Amin Shams, Béatrice Halbach, Aleksandra Malinowska and Wenjie Zhang (2014). "The UTIP Global Inequality Data Sets 1963-2008: Updates, Revisions and Quality Checks" UTIP Working Paper No. 68

Galbraith, James K., Jaehee Choi, Béatrice Halbach, Aleksandra Malinowska, and Wenjie Zhang. 2015. "A comparison of major world inequality data sets: LIS, OECD, SILC, WDI and EHII." UTIP Working Paper No. 69

Luxembourg Income Studies: http://www.lisdatacenter.org/.

Socio-Economic Database for Latin America and the Caribbean (SEDLAS and The World Bank), sedlac.econo.unlp.edu.ar http://sedlac.econo.unlp.edu.ar/eng/

University of Texas Inequality Project. 2011 EHII. http://utip.gov.utexas.edu/data.html

World Bank. 2007. World Development Indicators Online. http://www.worldbank.org/.

Appendix I: Gini Coefficients Compared Across EHII and SEDLAC

Figure 1: EHII Inequality Observations

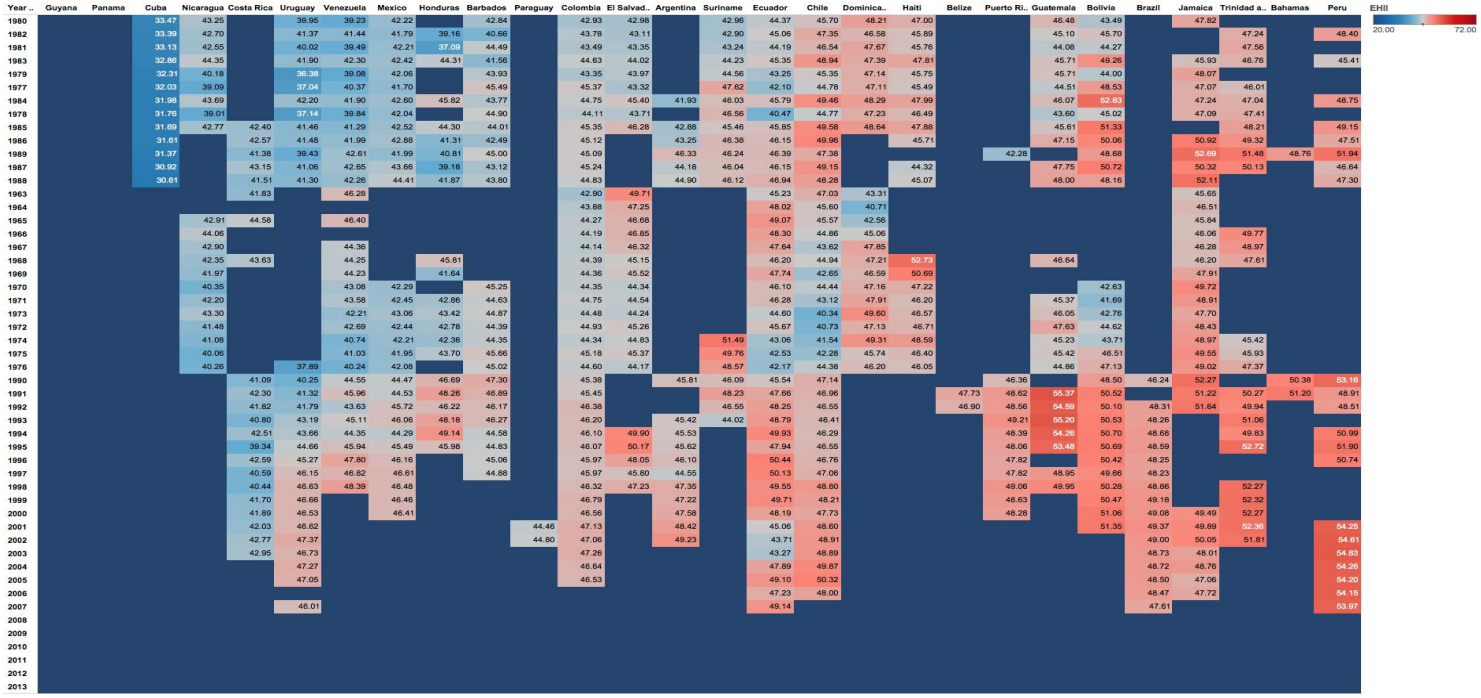


Figure 2: SEDLAC Inequality Observations

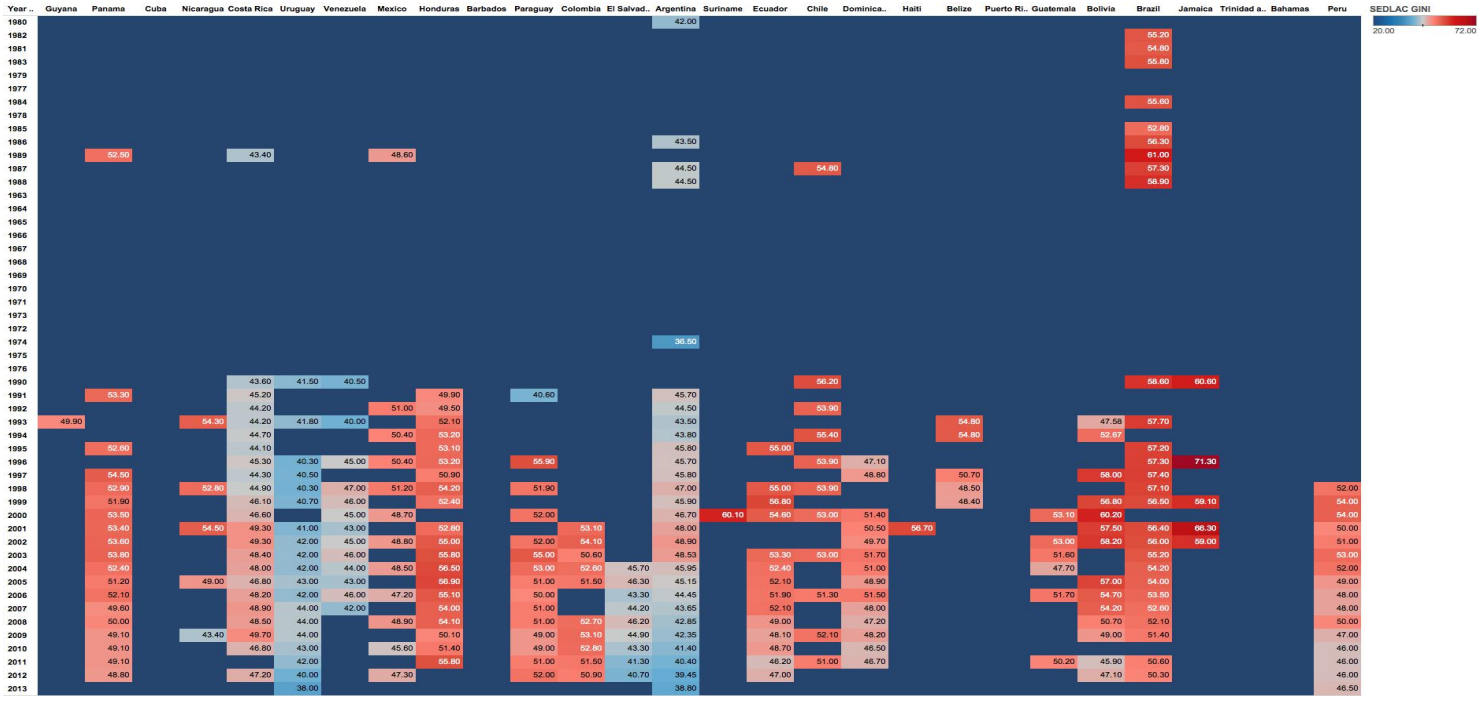
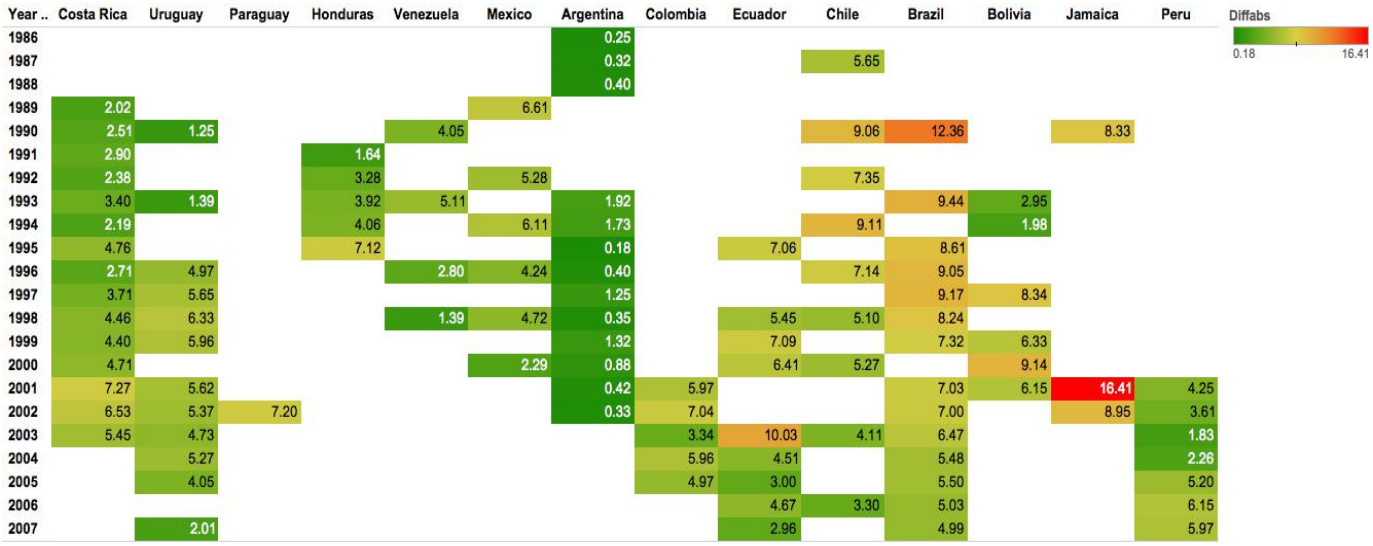


Figure 3: Absolute Differences between EHII and SEDLAC for Latin America



Appendix II: Comparing different data sets in Latin America

Figure 4: Country-year observation in each data set per country

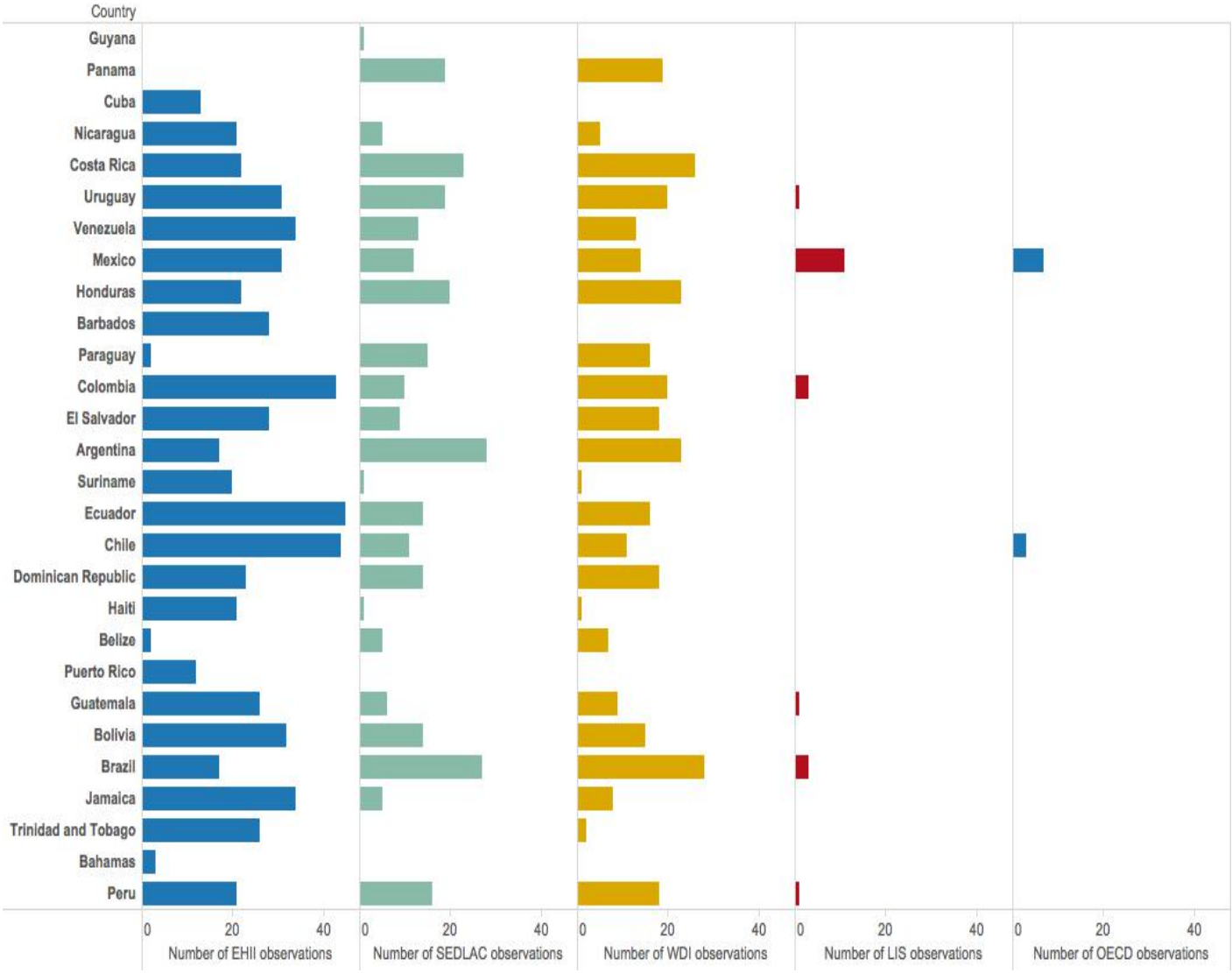


Figure 5: Country Gini coefficient observations per year per data set

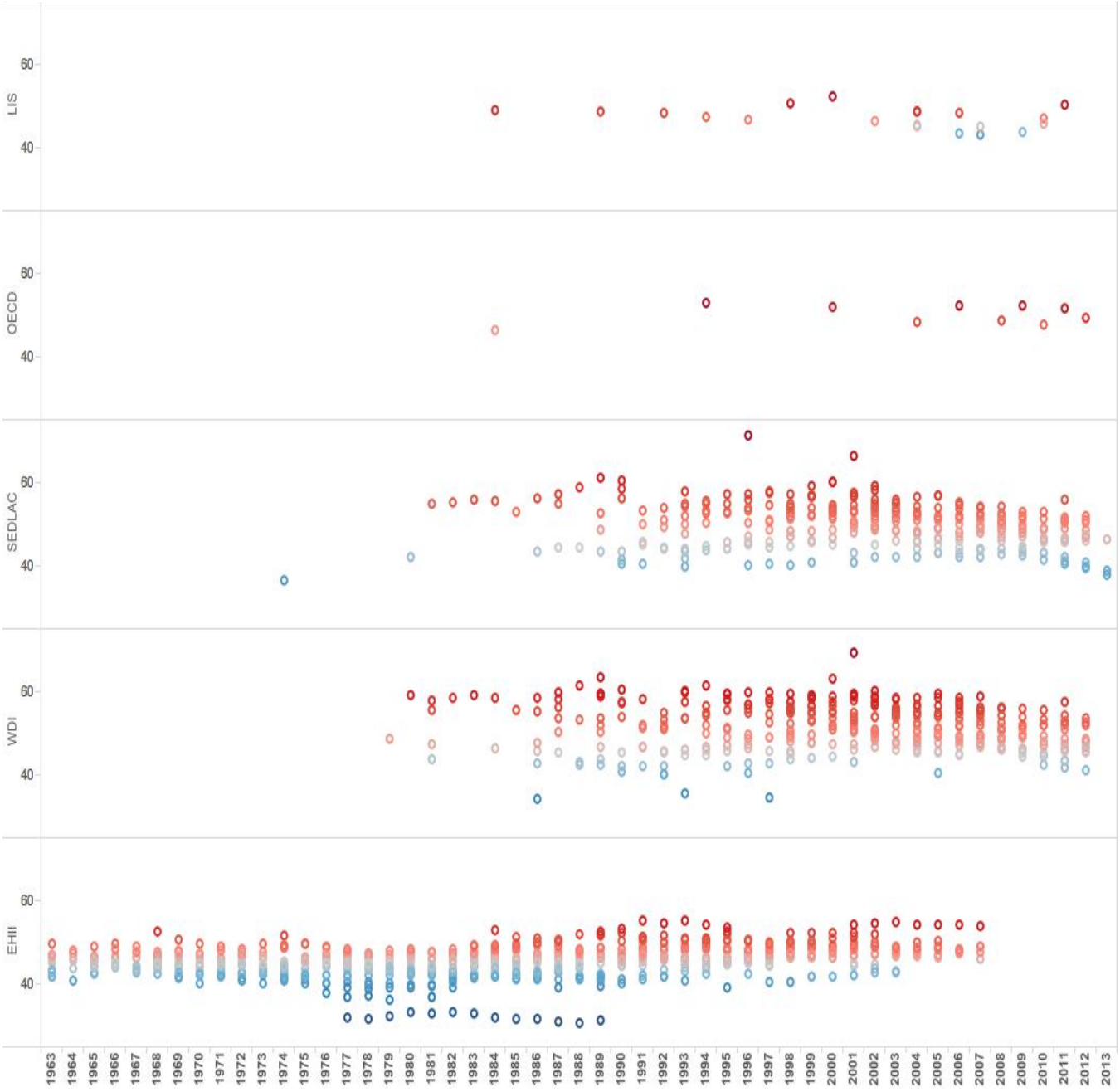
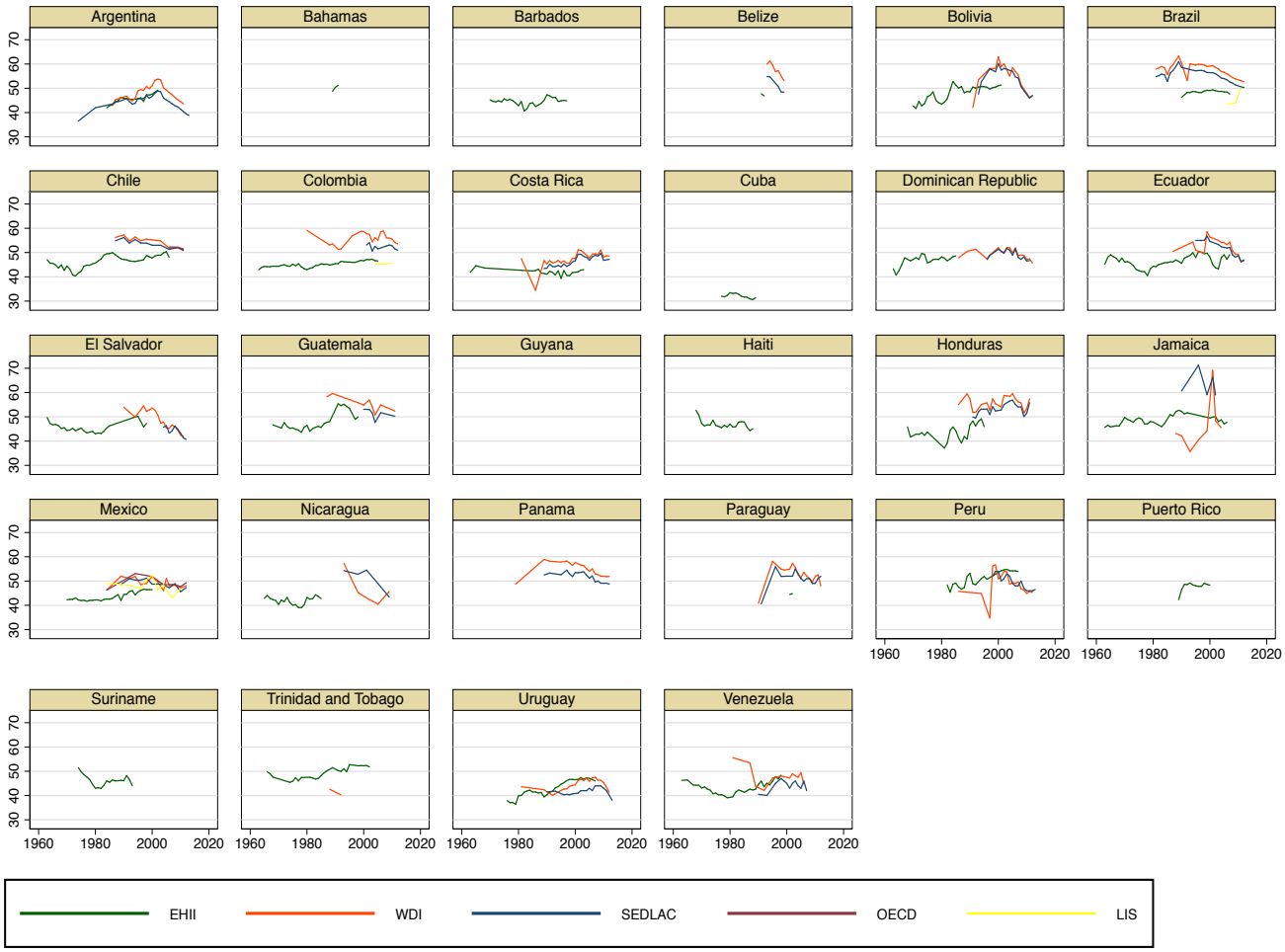


Figure 6: Gini coefficients by country as reported by the different data sets



Note: Guyana has only one observation for SEDLAC in 1993 of 49.9 Gini points.