

MAXIMUM CONSTRAINED PSEUDO- LIKELIHOOD ESTIMATION OF INCOME DISTRIBUTIONS, COMBINING SOURCES.

Alfredo Bustos (*)

Miriam Romo (*)

ESCoE Conference on Economic Measurement

May 11-13, 2021

(*) INEGI, Mexico

Summary

- The Maximum Constrained Pseudo Likelihood (MCPL) criterion to fit income parametric distributions using simultaneously several data sources is outlined.
- Discuss the evolution of income distributions in Mexico and its states, for the years 2010-2016.
- Comparisons will be based mainly on inequality measurements from estimated distributions and from National Household Income and Expenditure Surveys (ENIGH) or its Socioeconomic Conditions Modules (MCS).
- Comparisons with sample results indicate
 - National inequality has increased in the period which contrasts with ENIGH results.
 - Within state inequality does not seem to have changed much.
 - This seems to suggest that between states inequality has grown.

Motivation

- Income distributions important input in determination of poverty and inequality measures, and also useful in determining fiscal policy.
- Main available data source for estimation: Household income surveys
- Estimation of distributions from survey data exhibits limitations.
 - **Underreporting** of household income (i. e., gross or net of deductions)
 - **Truncation**: top income households not included in random sample or response missing
- Complementary data sources also show some limitations.
 - National Accounts (SNA) totals (not suited for distributional purposes)
 - Tax records (cover only partially the population)

Discrepancies between NAS and Surveys: The Mexican case.

- National Household Income and Expenditure Survey (ENIGH, Spanish acronym), sample size about 10,000 households (2012).
- Usually, total household current income from SNA (household institutional sector), nearly 2½ times larger than ENIGH estimates.
- Suggested answers, so far:
 - Adjust survey incomes, or their aggregates, to SNA figures assuming difference explained totally by either underreporting or truncation.

SOME INCOME ADJUSTMENT LITERATURE

Latin America

- Martinez (1970) discusses current income discrepancies between surveys and national accounts in Mexico and suggests ways to adjust survey data.
- Altimir (1987) suggests proportional adjustments, income source by income source assuming representativeness of the sample. ECLAC followed approach for many years when estimating poverty in the region.
- Leyva-Parra (2004) carries out a survey of adjustment proposals in the literature, while pointing out their similarities but also their differences. He suggests there ought to be an **optimality criterion** but stops short of suggesting any.
- Campos-Vázquez, et al. (2014) follow a proposal by Lakner et al. (2013) to correct only the upper part of the distribution while dealing with optimal taxation.

Limitations of surveys

- Korinek, et al. (2006) talk about some effects of survey non-response on income distributions

Discrepancies between surveys and national accounts

- Fesseau, et al. (2013) discuss income, consumption and saving discrepancies for many countries.
- Lakner, et al. (2013) propose a method for adjusting discrepancies in consumption, rather than income, at the top deciles of the consumption distribution via a Pareto distribution.

**CRITERION: MAXIMUM CONSTRAINED PSEUDO
LIKELIHOOD (MCPL)**

Purpose

- To produce likelier estimates for household income distributions, which agree with more than two data sources.

Proposed criterion

- Criterion makes comparisons possible, thus reducing arbitrariness.
- Initially, two parametric families fitted; also helps reduce arbitrariness.
- All available sources used

$$\text{Model: } f(y; \underline{\theta}) \Rightarrow \begin{cases} \ell(\underline{\theta}; \underline{Y}_{(i)}) = \ln(f(\underline{Y}_{(i)}; \underline{\theta})) \\ h(\underline{\theta}) \end{cases}$$

$$\text{Criterion: } \underset{\underline{\theta}, \underline{\lambda}}{\text{Max}} \left\{ \sum_{i=1}^n \frac{1}{\pi_{(i)}} \ell(\underline{\theta}; \underline{Y}_{(i)}) - \underline{\lambda}'(h(\underline{\theta}) - \underline{c}) \right\}$$

$$\text{Survey: } \begin{cases} Y_{(i)} \\ \pi_{(i)} \end{cases}, i = 1, \dots, n;$$

$$\begin{aligned} \text{SNA: } c_1 &= \text{Total}(\hat{Y}_{SNA}); \\ \text{Tax: } c_2 &= \text{Avge}(Y_{Max-k}, \dots, Y_{Max}); \end{aligned}$$

EXAMPLES OF CONSTRAINTS

Concept	Constraint:	Interpretation
Average household income (Source: SNA)	$h_1(\underline{\hat{\theta}}) = E[Y \underline{\hat{\theta}}] = c_1$	Average income from fitted model equals the one from SNA, c_1 .
Income Integral (Source: Tax data)	$h_2(\underline{\hat{\theta}}) = E(Y Y > \varphi_\alpha, \underline{\hat{\theta}}) =$ $= \frac{1}{\alpha} \int_{\varphi_\alpha}^{\infty} y f_Y(y \underline{\hat{\theta}}) dy = c_2$ $\text{where } \alpha = \int_{\varphi_\alpha}^{\infty} f_Y(y \underline{\hat{\theta}}) dy$	Average income c_2 of $100\alpha\%$ of households whose income above threshold φ_α , according to SAT, coincides with that of the group whose income is also above the same φ_α , according to the adjusted model. Typically, $\alpha = 10^{-5}$ or 10^{-6}

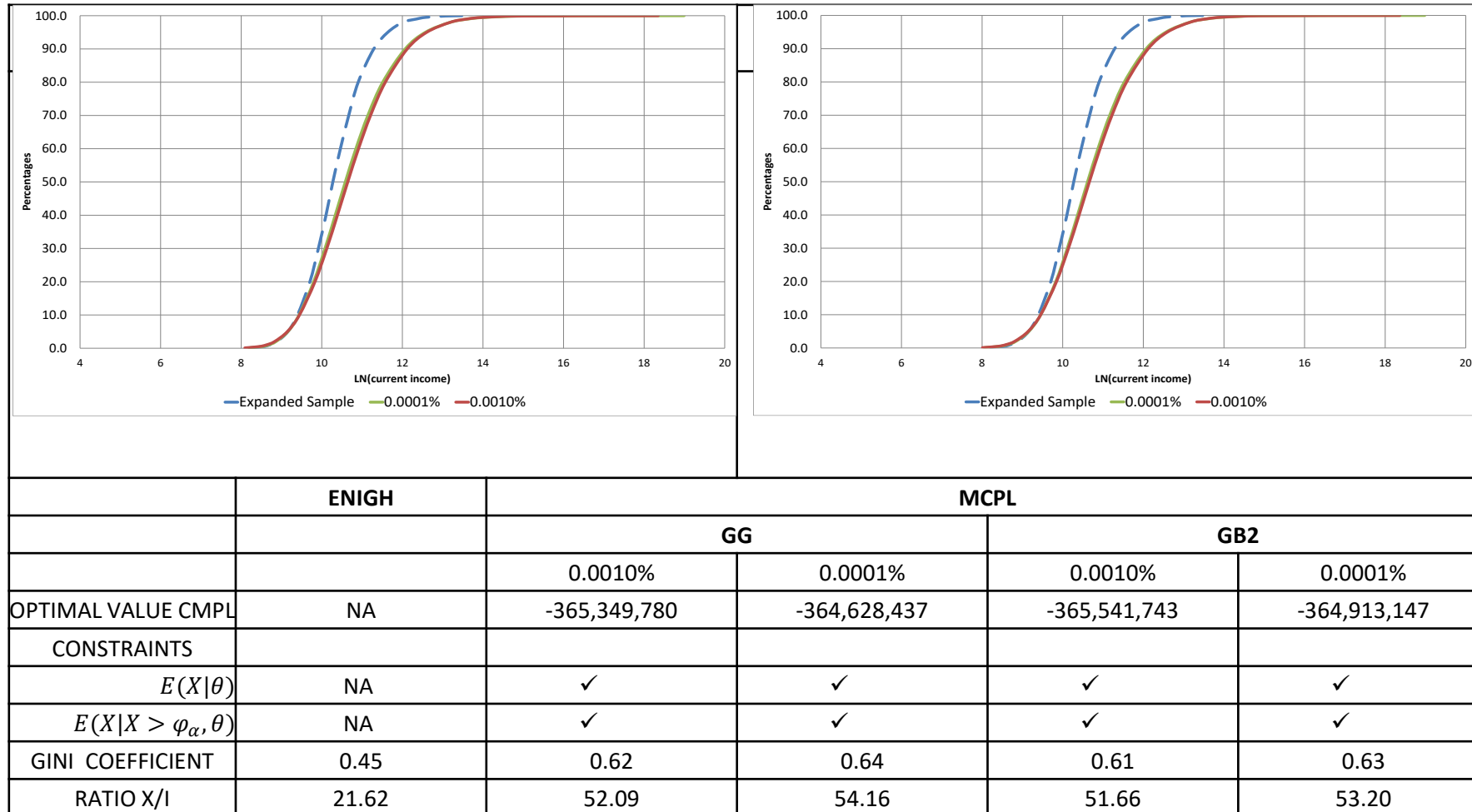
A R-script has been developed, and is currently under improvement, for performing estimation. It uses R-package Alabama, for constrained optimization, and gamlss.dist package, which makes available different distributional models.

CONSTRAINTS ARE IMPORTANT

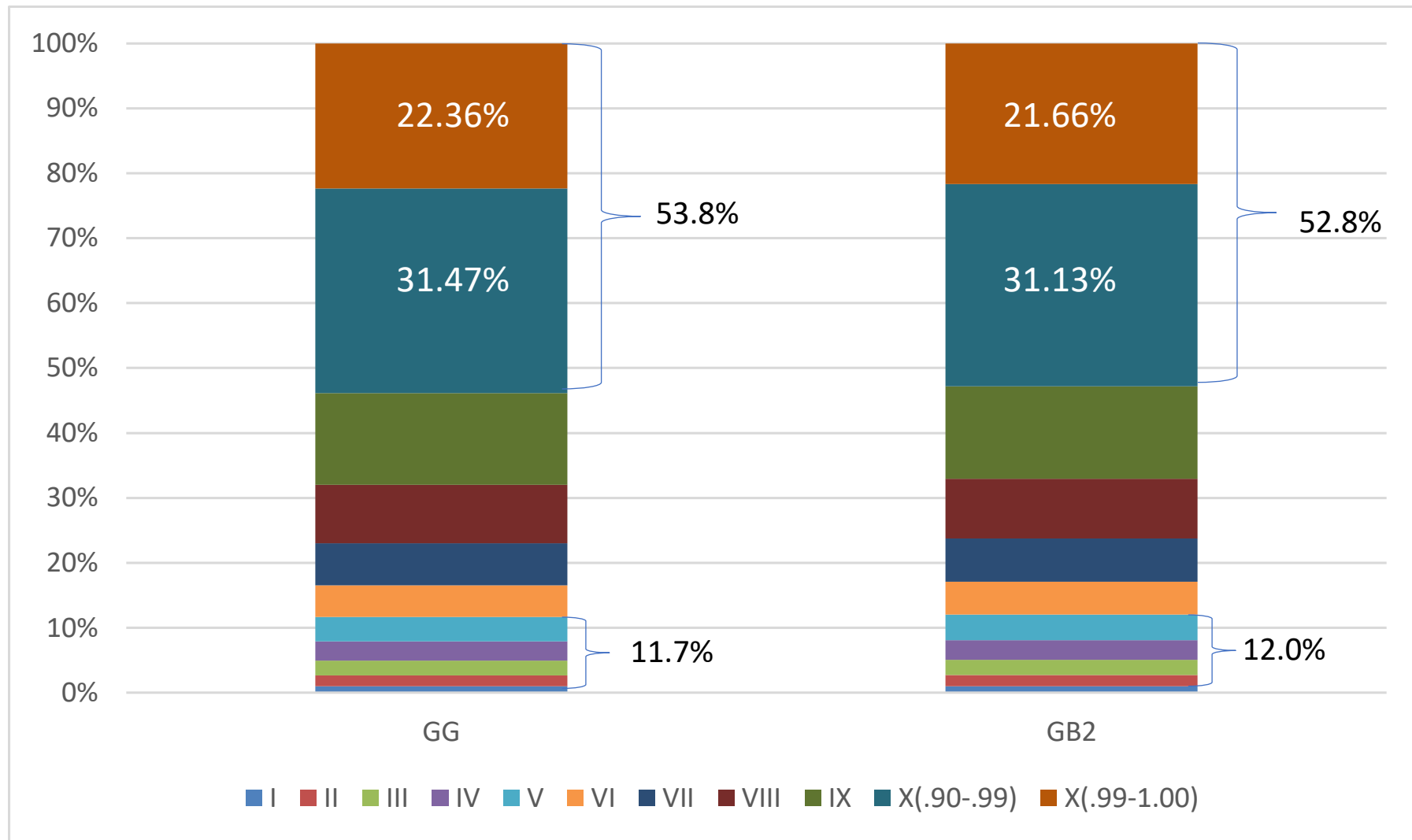
CONSTRAINTS ARE IMPORTANT

- Two fitted models,
 - Generalized Gamma (GG) and Type II Generalized Beta (GB2);
 - 3 or more parameters (p), since two constraints (c) applied;
 - $p > c$ required for survey data to play any role.
 - For each model, two thresholds (0.0010% & 0.0001%) for top incomes.
 - Same value of SNA household average income.

Optimal fit of two models using two different income thresholds, Mexico, 2012.



MCPL Income distribution by decile, Mexico, 2012



Observations

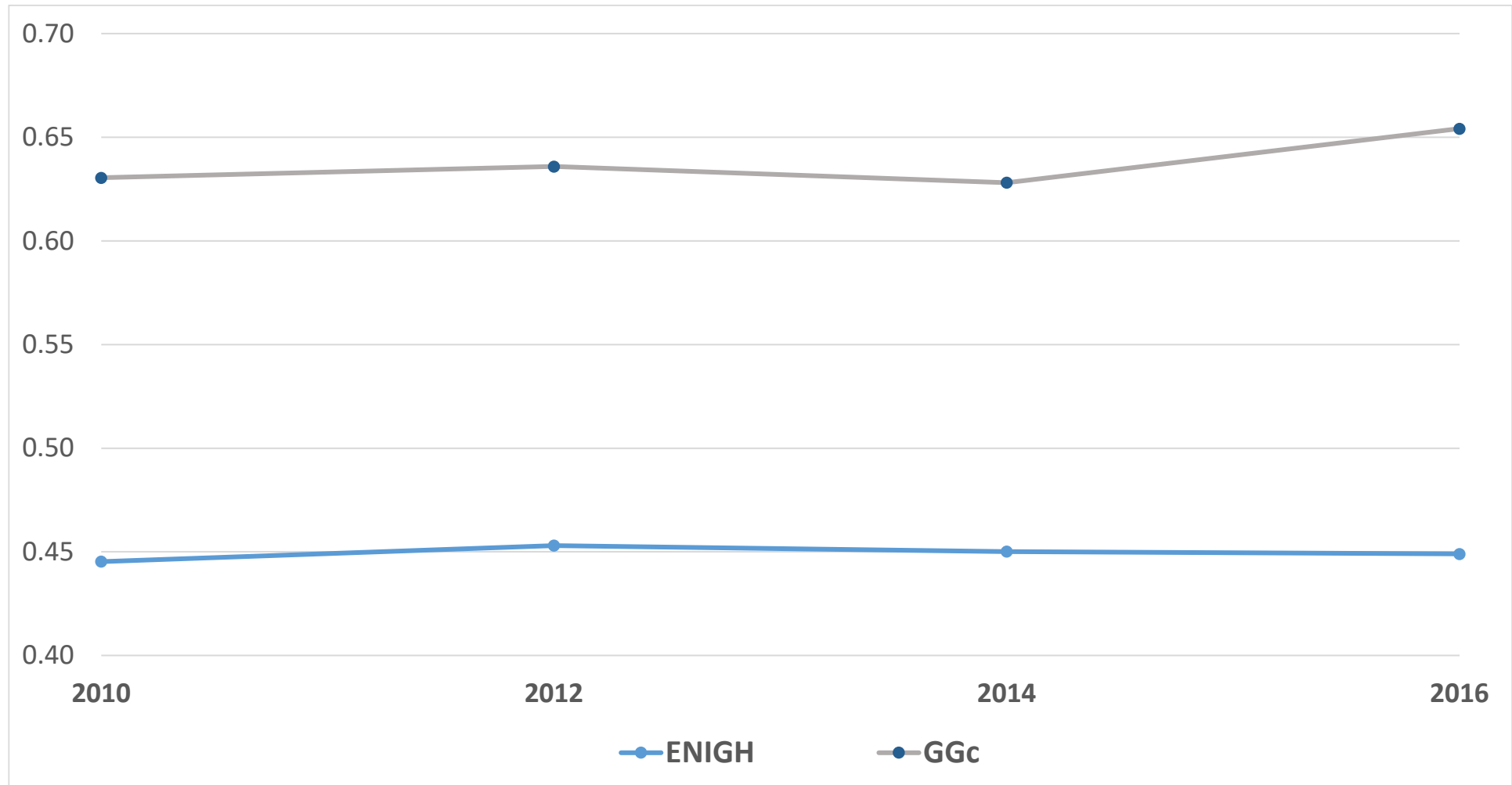
- Constraints weigh heavily on estimates
- Too many constraints render data useless
- Changes in data have little or no effect on fitted model
- Evidence of both underreporting and truncation in the sample
- Optimal fit shows underreporting grows more than proportionally with income.

COMPARISON OVER TIME

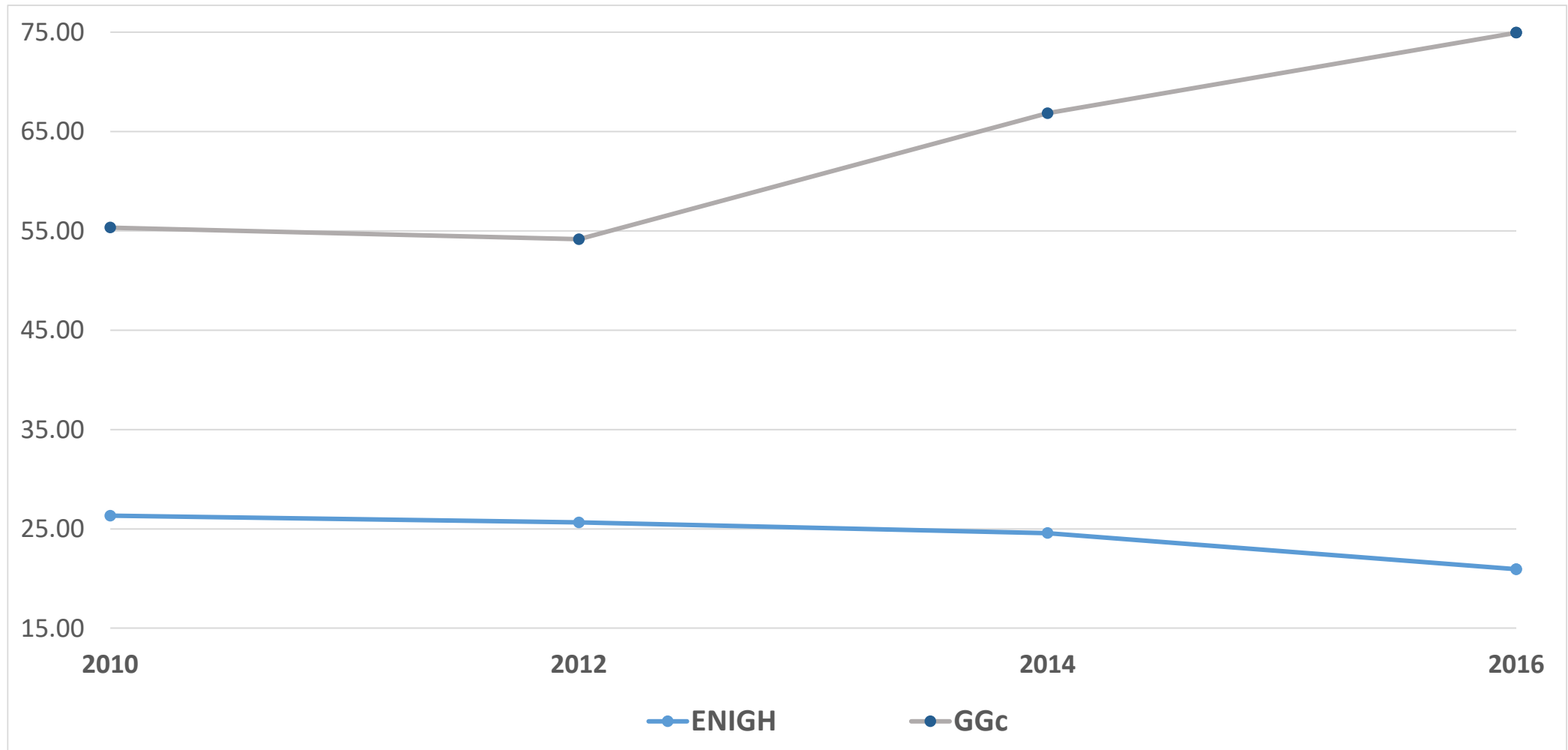
Income distribution in Mexico, 2010-2016

- Mexico's Household Income and Expenditure Survey (ENIGH) takes place every other year.
- SNA summaries are published yearly.
- Anonymized tax records for 2010, 2012, 2014 and 2016 became available.
- According to ENIGH, inequality in Mexico did not increase:
 - Gini coefficients have fluctuated around 0.45, and
 - Ratio of X-th decile income to that of the 1-st one (X/I) shows a downward trend.
- However, when all 3 sources are taken into consideration through MCPL, a different picture appears.
 - In 2016, Gini coeff. was above 0.65 for the first time
 - X/I rose from 55 times, in 2010, to 75, in 2016.

Survey (ENIGH) and optimal model Gini coefficients, 2010-2016.



Household income inequality: Ratio of X-th to I-st Decile income.



EVOLUTION OF INEQUALITY FOR MEXICAN STATES.

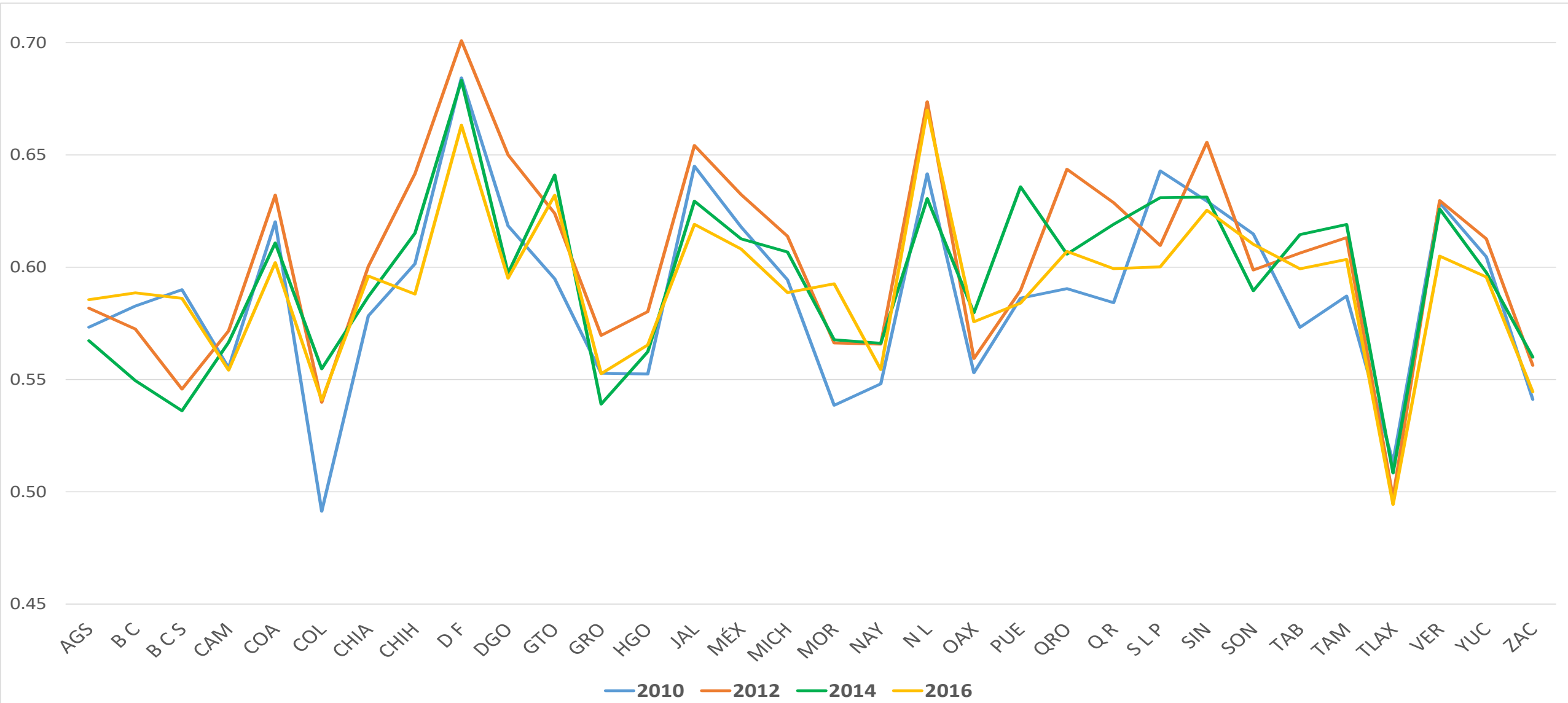
Procedure

- Social Conditions Module (MCS) is an ENIGH module which does not collect expenditure information.
 - Larger sample size (54,184 vs. 10,062 in 2012) allows for state level disaggregation of income.
 - MCS total household income smaller than ENIGH's;
 - constraints used to correct.
- No institutional sector accounts are produced at state level.
 - SNA national household income proportionally distributed according to MCS shares.
- Tax records include taxpayer's state of residence.
 - Individual state income thresholds were developed, by year, from them.
- Separate models were fitted by state and by year (32 by 4)

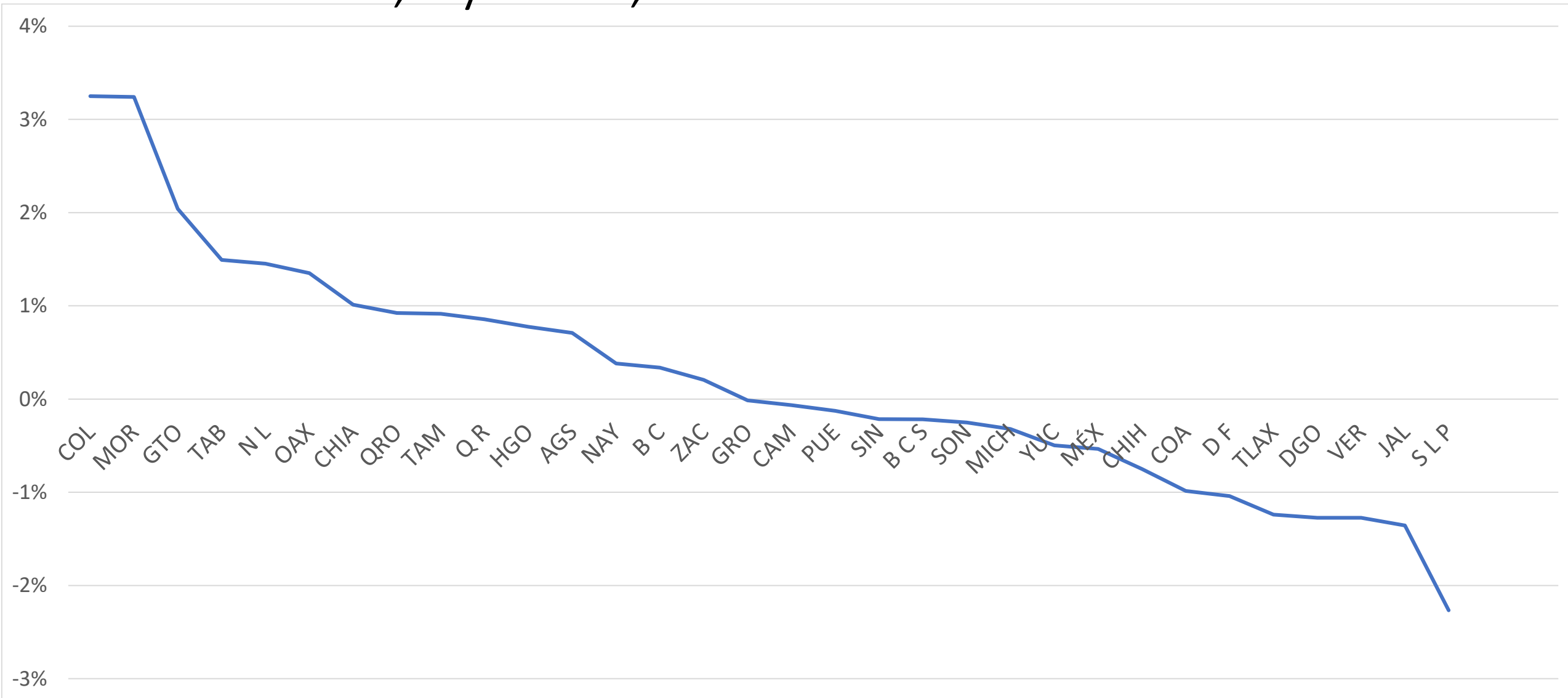
Comments

- In contrast with country-wide behaviour, 2016 does not appear to be the most unequal year within each state.
- Therefore, within state inequality may have in fact decreased.
- Is this evidence that between states inequality grew?
- We have yet to develop a procedure to prove this is the case.

GINIS



Geometric mean of biennial growth of Gini coefficients, by state, 2010-2016.



Concluding remarks

- Comparisons between countries from survey results, ignore differences in sampling design (e.g., oversampling of high income households) and/or analysis (e.g., calibration of sampling data to known totals).
- Standard analysis procedures, useful.
- For more general purposes, our approach allows for fairer comparisons:
 - SNA constraints stem from sound and comparable methodology
 - Constraints reduce sampling design or data analysis influence and
 - Makes model choice less arbitrary.

Summary

- The Maximum Constrained Pseudo Likelihood (MCPL) criterion to fit income parametric distributions using simultaneously several data sources is outlined.
- Discuss the evolution of income distributions in Mexico and its states, for the years 2010-2016.
- Comparisons will be based mainly on inequality measurements from estimated distributions and from National Household Income and Expenditure Surveys (ENIGH) or its Socioeconomic Conditions Modules (MCS).
- Comparisons with sample results indicate
 - National inequality has increased in the period which contrasts with ENIGH results.
 - Within state inequality does not seem to have grown, in general.
 - This seems to suggest that between states inequality has grown.

References

- [1] O. Altimir, Income distribution statistics in Latin America and their reliability, *Review of Income and Wealth* 33(2) (1987).
- [2] Bergsman, J., Income Distribution and Poverty in Mexico, 1963–1977, World Bank Staff Working Paper, No. 395, Washington, The World Bank, July 1980.
- [3] Bustos, A., *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources*, Workshop on Measuring Inequalities of Income and Wealth, High-Level Expert Group on the Measurement of Economic Performance and Social Progress, Berlin, Sep. 15-16, 2015.
- [4] Bustos, A., *Estimation of the distribution of income from survey data, adjusting for compatibility with other sources*. *Statistical Journal of the IAOS*, Statistical Journal of the IAOS 31 (2015) 565–577
- [5] Bustos, A., Leyva, G., *Towards a More Realistic Estimate of the Income Distribution in Mexico*. *Latin American Policy*, Volume 8, Issue 1, June 2017, Pages 114-126.
- [6] Campos Vázquez, R., Chávez Jiménez, E., Esquivel Hernández, G., *“Los Ingresos Altos, la Tributación Óptima y la Recaudación Posible”*, primer lugar, Premio Nacional de Finanzas Públicas, 2014.
- [7] Economic Commission for Latin America and the Caribbean (ECLAC), *“Income poverty measurement: updated methodology and results”*, ECLAC Methodologies, No. 2 (LC/PUB.2018/22-P), Santiago, 2019.
- [8] Faulkner, Ch., Using G2 to measure income inequality in two Latin American upper middle-income countries, *Statistical Journal of the IAOS*, 30 (2014), 321–329.
- [9] Félix, D., Income distribution Trends in Mexico and the Kuznets Curves, in: *The Political Economy of Brasil and Mexico*, Weinert and Hewlett (eds.), Philadelphia, ISMI Press, 1979.
- [10] Fesseau, M., and Mattonetti, M.L., Distributional Measures across Household Groups in a National Accounts Framework: Results from an Experimental Cross-country Exercise on Household Income, Consumption and Saving, *OECD Statistics Working Papers*, No. 2013/04, OECD Publishing, 2013.
- [11] Korinek, A., Mistiaen, J., and Ravallion, M., Survey nonresponse and the distribution of income, *Journal of Economic Inequality* 4(1) (2006), 33–55.
- [12] Lakner, Ch., Milanovic, B., "Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession". World Bank policy research paper 6719, 2013.
- [13] Leyva-Parra, G., *El ajuste del ingreso de la ENIGH con la Contabilidad Nacional y la medición de la pobreza en México*, Serie: Documentos de Investigación, No. 19, SEDESOL, México, 2004
- [14] Martinez, I., *La Distribución del Ingreso en México: Tendencias y Proyecciones a 1980*, Vol. 1, México D.F., Siglo Veintiuno Editores, 1970.