Supplementary materials

S1: Small area estimation procedure

For the implementation of the small area estimation (SAE) procedure, our primary data source is the 2015 EIC survey implemented by the National Institute of Statistics and Geography (INEGI) with the objective of updating the socio-demographic information between the 2010 census and the one to be carried out in 2020. This survey covers 6.1 million households (more than 22 million individuals) and is representative at the national, state and municipal levels. It provides basic information on households' assets, housing, education, ethnicity, health, etc. However, this survey fails to collect accurate data on household income. This is the reason why we also use the 2016 ENIGH survey, which covers more than 70,000 households and provides precise information on household income and its different components. Despite many recent refinements in SAE methods, we adopt the standard approach developed by Elbers, Lanjouw & Lanjouw (ELL) (2003) because of its multiple implementations.

The first step in ELL methodology estimates a welfare model (called the Beta model), based on household survey data (ENIGH data in our case), following equation (1):

$$lnY_{hm} = X_{hm}\beta + \eta_m + \varepsilon_{hm} \quad (1)$$

where Y_{hm} is the per capita income of household h in municipality m and X_{hm} are income predictors. The error terms η_m and ε_{hm} represent the unexplained variation at municipality and household levels, respectively, and are treated as random effects. This specific structure of the error component explains why model (1) is estimated using Generalized Least Squares (GLS). Two additional elements are important components when estimating the welfare model. First, in addition to household-level variables, ELL recommend to include municipal-level variables as covariates to account for heterogeneity between municipalities. Second, in

the ELL specification, the household-specific error component $\widehat{\varepsilon_{hm}}$ is assumed to be heteroscedastic (i.e. to vary between households). The ELL strategy for modelling heteroscedasticity consists of estimating a model to explain the squared predicted household-level residuals by household-level and municipality-level characteristics through a parametric logistic transformation (called the Alpha model).

In the second step of the methodology, the parameter estimates from equation (1) are applied to census data (EIC data in our case) in order to predict income for all households and then to estimate welfare indicators (inequality indices in our case). More precisely, a series of k Monte Carlo simulations (usually around 100) are implemented. In each simulation, a set of values $\hat{\beta}$, $\widehat{\eta_m}$ and $\widehat{\varepsilon_{hm}}$ are drawn from their estimated distributions and an estimate of income and the Gini index are produced. We have also calculated the generalized entropy indices to test the robustness of our results in relation to alternative inequality measures. After k simulations, we can calculate the average income and the average of inequality indices which can be treated as representative at the municipal level.

The numerous applications of SAE methods provide practical guidelines for constructing the first-stage model. One important issue is that variables are available and comparable between the survey and the census (both in their definition and in their distribution). Among comparable variables, it is necessary to include a large set of predictors with characteristics for the head of the household (age, sex, employment, education) and the household (assets, housing, demographic composition, employment, education, migration, etc.). In addition, ELL recommend the inclusion of municipal-level variables (aggregated means from census data, for instance) in order to reduce the magnitude of the unexplained municipal-level component of the error term η_m. Moreover, as recommended by Tarozzi & Deaton (2009), we include non-linear functions of quantitative variables by including their squared terms. We also take into account some interaction terms as recommended by Fuji (2010).

The final set of variables included in the income model has been determined by a stepwise procedure and ex-post diagnostics. More precisely, once controlling for the comparability of variables between the EIC and ENIGH surveys, we set the model specification in such a way as to maximize the number of significant variables, to maximize the adjusted R-squared and to minimize the variance in the municipal component of the error term η_m. Our SAE estimates also include a heteroscedasticity model (Alpha model) in which residuals predicted from the income model are regressed on all the explanatory variables.

In Table S1, GLS estimates for the logarithm of monthly per capita household income are reported. Following the above-described procedure, more than 40 explanatory variables have been included. The estimates perform to a highly competitive extent with an adjusted R-squared close to 0.60 and the variance of η_m being residual (less than 0.015). It is also worth noting that heteroscedasticity is found to be negligible ($R^2 < 0.02$ in the Alpha model). The parameter estimates from this model are then applied to EIC data through 100 Monte Carlo simulations. From these simulations, the mean per capita household income and the main measures of income inequality are calculated.

Table S1: Income model for small area estimation (GLS estimates).

Variables	Coefficient	z	p-value
Constant	8.0757***	184.34	0.000
Household head characteristics			
Male	-0.0093	-1.54	0.124
Age	-0.0043***	-6.05	0.000
Age squared	0.00005***	8.92	0.000
Indigenous (self-description)	-0.0142***	-3.24	0.001
Literate	0.0610***	8.39	0.000
Secondary education or higher	0.0188***	2.68	0.007
In a couple	0.0164**	2.56	0.010
Household characteristics			
Urban	0.1265***	10.00	0.000
Migration (=1 for households whose head lived in another municipality in 2010)	0.0509***	4.98	0.000
Household size	-0.3053***	-63.38	0.000
Household size squared	0.0189***	45.34	0.000
Proportion of male	-0.4073***	-13.58	0.000
Proportion of male squared	0.5153***	17.15	0.000
Proportion of children (11 y.o. or less)	-0.0606*	-1.91	0.056
Proportion of children squared	-0.2857***	-5.28	0.000
Proportion of hh members (15 y.o or more) with at least secondary education	0.1791***	8.41	0.000
Proportion of hh members with at least secondary education squared	0.0259	1.30	0.193
Employment rate (for 12-65 y.o. members)	0.3805***	14.43	0.000
Employment rate squared	0.0692***	2.84	0.005
Number of rooms per capita	0.0244***	4.94	0.000
Number of rooms per capita squared	0.0055***	9.74	0.000
HH with access to piped water into dwelling	0.0529***	10.22	0.000
HH with access to piped sewer system	0.0336***	5.78	0.000
HH equipped with a car	0.1498***	21.07	0.000
HH equipped with a mobile phone	0.1228***	15.50	0.000
HH equipped with a computer	0.1823***	32.28	0.000
HH with access to the internet	0.1579***	12.34	0.000
HH equipped with a washing machine	0.0668***	14.60	0.000
HH equipped with a refrigerator	0.0519***	8.31	0.000
HH equipped with a flat screen tv	0.0695***	17.07	0.000
HH with access to pay tv	0.1355***	33.60	0.000
Interaction terms			
Urban * household size	-0.0150***	-6.97	0.000
Urban * internet	-0.0329**	-2.44	0.015
Urban * mobile phone	-0.0348***	-3.31	0.001
Urban * car	0.0777***	9.15	0.000
Municipal controls			
Municipal employment rate	0.9586***	10.80	0.000
Municipal secondary education rate	0.2569***	3.94	0.000
Municipal migration rate	0.3522***	4.98	0.000
Municipal car equipment rate	0.2543***	7.19	0.000
Municipal computer equipment rate	0.2026***	2.61	0.009

N	69,078
Adjusted R-squared (Beta model)	0.583
Adjusted R-squared (Alpha model)	0.015
Sigma eta squared	0.013
Variance of epsilon	0.270

Notes: Robust t-statistics are reported into brackets. Level of statistical significance: 1 %***, 5 %**, and 10 %*.
Source: Authors' calculations based on ENIGH data.

References

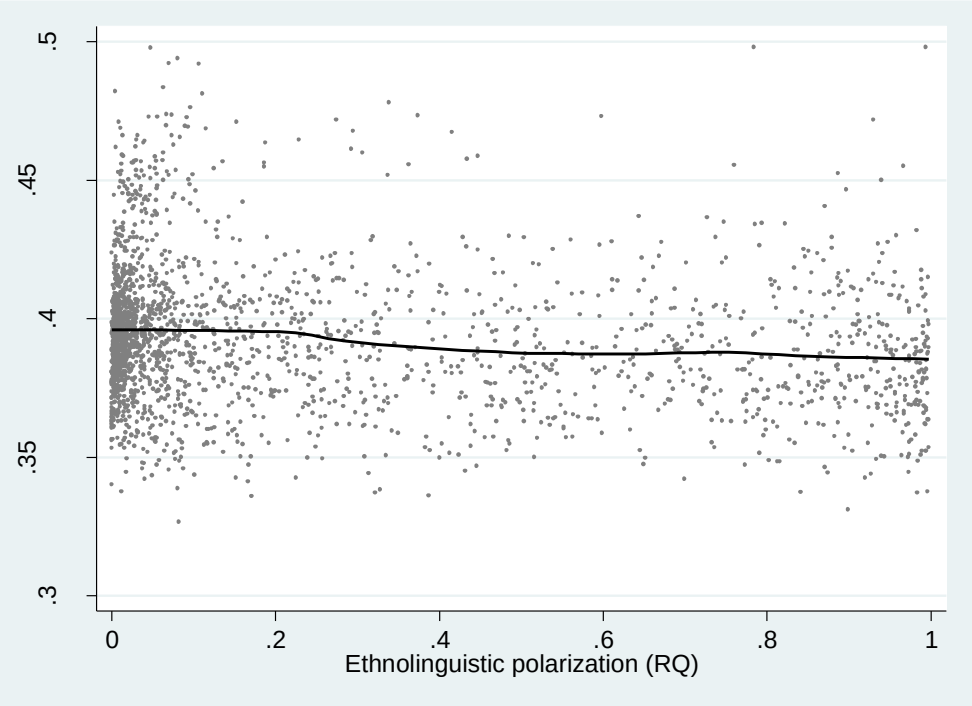
Elbers, C., Lanjouw, J.O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica, 71(1),* 355-364.

Fujii, T. (2010). Micro-level estimation of child undernutrition in Cambodia. *World Bank Economic Review, 24(3),* 520-553.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics, 91(4),* 773-792.

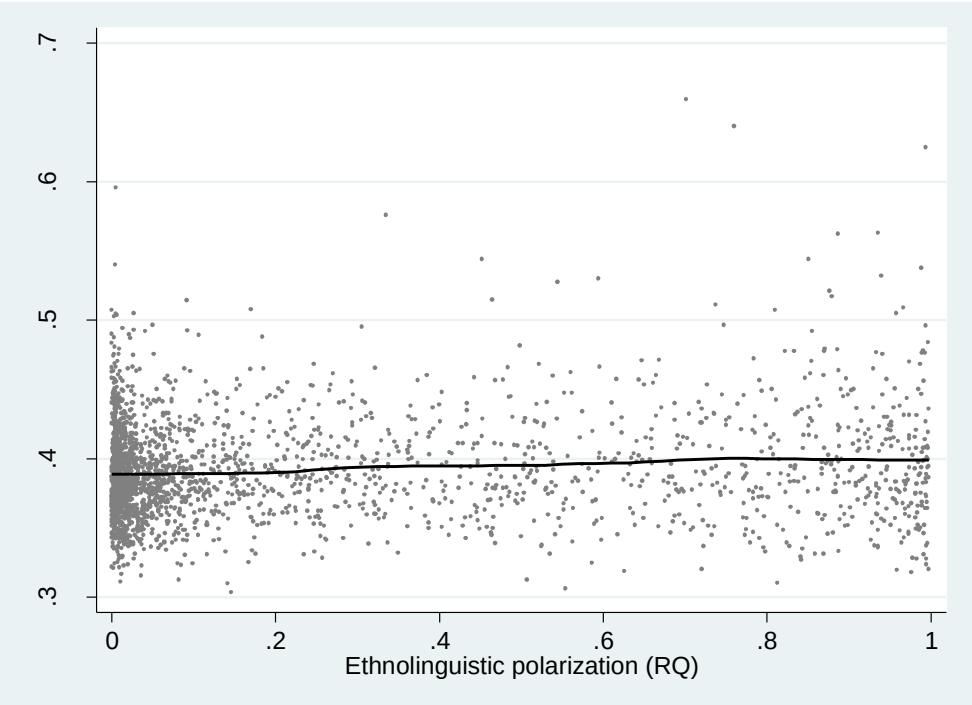
S2: Additional descriptive statistics

Figure S2.1: Non-parametric fit (local polynomial smoothing) between income Gini and ethnolinguistic polarization.



Note: Epanechnikow kernel; Bandwidth=0.1.
Source: Authors' calculations.

Figure S2.2: Non-parametric fit (local polynomial smoothing) between CONEVAL's income Gini and ethnolinguistic polarization.



Note: Epanechnikow kernel; Bandwidth=0.1.
Source: Authors' calculations.

S3: Additional robustness checks

Table S3.1: Robustness checks with alternative inequality indices (Spatial IV).

	Local transfers				Local taxes		
	GE(0)	GE(1)	GE(2)		GE(0)	GE(1)	GE(2)
RQ	-2.4797	-2.2430	-0.8847	RQ	-1.5930*	-1.0821	13.32865
	(-1.39)	(-0.70)	(-0.03)		(-1.79)	(-0.90)	(0.79)
Transfers	-0.0070***	-0.0079*	0.0163	Taxes	-0.0142***	-0.0131***	0.0262
	(-2.78)	(-1.84)	(0.55)		(-3.36)	(-3.04)	(1.23)
Transfers * RQ	0.0149**	0.0143	0.0505	Taxes * RQ	0.0336***	0.0310***	-0.0231
	(2.26)	(1.14)	(0.43)		(3.99)	(3.67)	(-0.26)

Notes: The entropy indices have been multiplied by 100 to rescale the values of coefficients. Coefficients on control variables are not reported. Instruments are the same as in Tables 3 and 4. Robust t-statistics are reported into brackets. Level of statistical significance: 1 %***, 5 %**, and 10 %*.
Source: Authors' calculations based on multiple datasets.

Table S3.2: Additional robustness checks (Spatial IV).

	Gini outliers excluded		Alternative instruments		Alternative ethnic diversity variable (NGV)		Additional control (PROSPERA)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RQ	-1.3246*	-5.6164***	-1.6758***	-0.0745			-3.4287***	-1.3503***
	(-1.66)	(-2.69)	(-2.61)	(-0.28)			(-3.351)	(-2.82)
NGV					-1.0467	-1.7078*		
					(-0.75)	(-1.86)		
Transfers	-0.0043***		-0.0032***		-0.0025*		-0.0074***	
	(-3.46)		(-4.06)		(-1.92)		(-4.71)	
Transfers * RQ	0.0071***		0.0034**				0.0124***	
	(2.89)		(2.18)				(3.82)	
Transfers * NGV					0.0104**			
					(1.99)			
Taxes		0.0005		-0.0010***		-0.0093***		-0.0062***
		(0.15)		(-2.57)		(-3.33)		(-2.95)
Taxes * RQ		0.0219		0.0045***				0.0190***
		(1.36)		(3.72)				(3.73)
Taxes * NGV						0.0359***		
						(3.98)		
N	1924	1852	1756	1685	1931	1859	1952	1888

Notes: The Gini index has been multiplied by 100 to rescale the values of coefficients. Coefficients on control variables are not reported. With the exception of regressions (3) and (4), instruments are the same as in Tables 2 and 3. Robust t-statistics are reported into brackets. Level of statistical significance: 1 %***, 5 %**, and 10 %*.
Regressions (1) and (2): Municipalities with the lowest (bottom 1%) and highest (top 1%) degrees of income inequality are excluded.
Regressions (3) and (4): In the spatial IV procedure, we use the second-order spatial lags of the selected instruments instead of the first-order spatial lags.
Regressions (5) and (6): Ethnic diversity is measured with the ethno-linguistic fractionalization index (NGV) instead of the ethno-linguistic polarization index (RQ).
Regressions (7) and (8): The share of PROSPERA beneficiaries at the municipal level in 2015 (data from https://datos.gob.mx/busca/organization/prospera) is included as an additional control variable.
Source: Authors' calculations based on multiple datasets.