Desafíos de medición de las desigualdades

Datos, problemas y soluciones

Ignacio Flores[1]

Seminario de Alto Nivel - Estadísticas Económicas
CEPAL Santiago,
18 Noviembre, 2019

[1]Coordinador LAC en World Inequality Lab -
Paris School of Economics -
INSEAD

Introducción General

Preguntas de investigación

- ¿Qué desafíos de medición para capturar a la parte más alta de la distribución? ¿Qué datos exísten y cómo podemos reconciliarlos? ¿Cómo mejorar las fuentes?

¿Qué datos?

Encuestas a Hogares

- Micro-datos, buena cobertura geográfica, covariables y representatividad
- Sólo años recientes. Incluye cesgos (e.g., sampling and non-sampling)

Impuestos - Administrativos

- Cubre mejor el top. Disponible más tiempo.
- Mayoritariamente tabulados sin covariables (a veces micro-datos). Inconsistencias en unidades y conceptos , calidad variable

Cuentas Nacionales

- Buena referencia para totales
- No incluye información distributiva (por ahora). Puede ser una caja negra

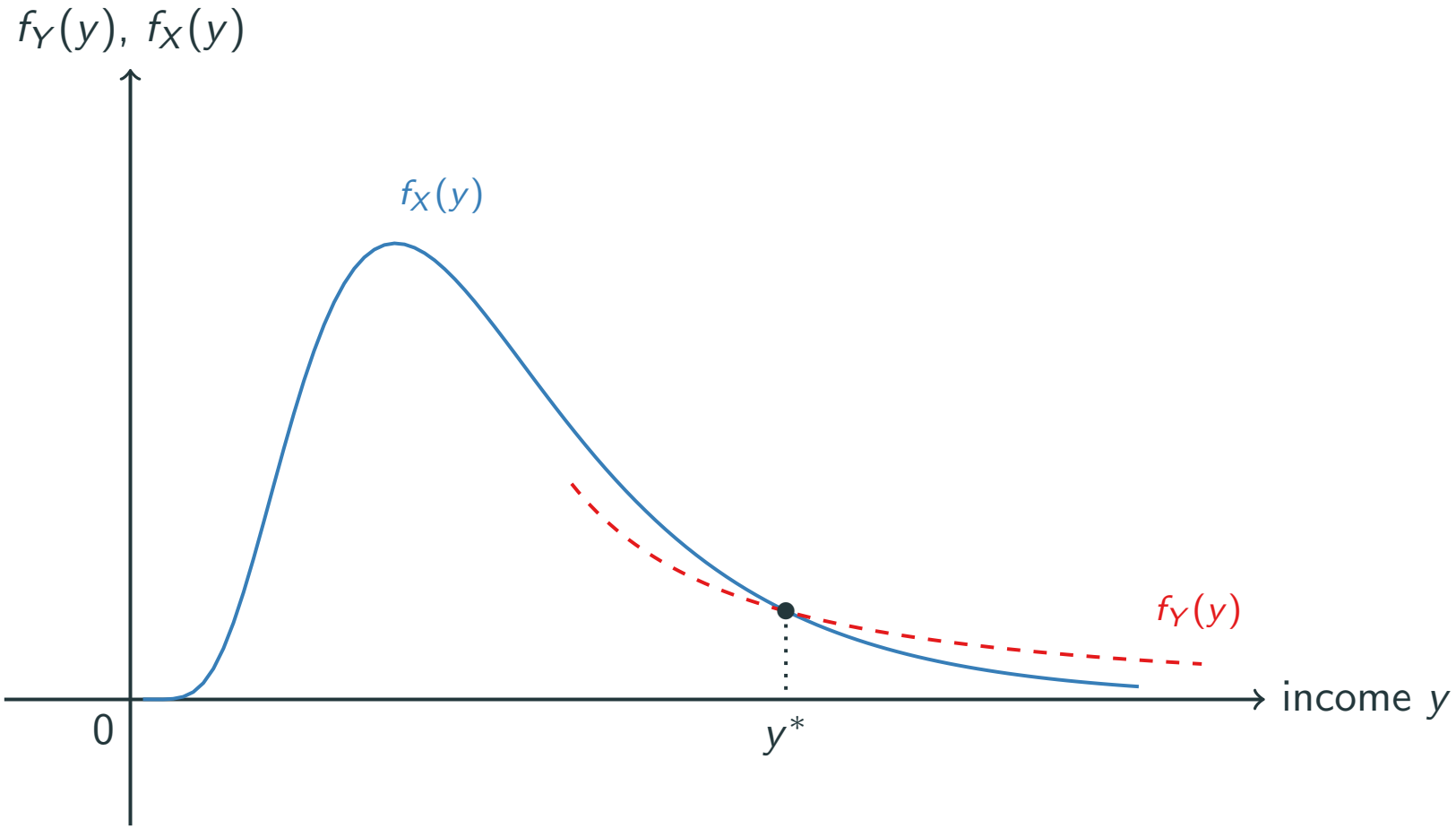
Otros

- Rankings de riqueza (e.g., Forbes)
- Fugas de datos o 'Leaks' (e.g., LuxLeaks, Panama Papers)

¿Se pueden combinar consistentemente?

- Si. Algunos ejemplos para países desarrollados: Garbinti, Goupille-Lebret and Piketty, 2016 (FR); Piketty, Saez and Zucman, 2018 (US) o OECD EG-DNA
- Pero los métodos (Alvaredo et al., 2017) no son fácilmente adaptables a países en desarrollo: Cobertura, confianza en datos fiscales (informalidad, peso de independientes)

Escena típica (LatAm): Encuestas vs. Tax



WID.WORLD

Interpretar la diferencia: Cesgos en encuestas

Sampling

- Small sample bias (Taleb and Douady, 2015)

Non-Sampling

- Misreporting (Bound and Krueger, 1991; Bollinger, 1998; Angel et al., 2017; Paulus, 2015)
- Heterogeneous response rates across the income distribution (Korinek, Mistiaen and Ravallion, 2006; Johansson and Klevmarken, 2007; Bollinger et al., 2015; Chenevert et al., 2016)

Interpretar la diferencia: Perspectiva formal

Non-Response

- $f_Z(y) = f_Y(y)\theta(y) \quad \Rightarrow \quad \theta(y) = f_Z(y)/f_Y(y)$

Misreporting

- $f_Z(y) = f_Y(y)(1 - p(y)) + f_M(y)\bar{p}$

Non-Response & Misreporting

- $f_Z(y)/f_Y(y) = \theta(y)(1 - p(y)) + f_M(y)/f_Y(y)\theta(y)\bar{p}$

Conclusiones

- Si ambos existen, se confunden *ex-post*.
- Misreporting sólo se puede resolver con *matching* individual

Metodología

$\Rightarrow$ Blanchet, Flores and Morgan, 2019 (BFM)

The Weight of the Rich: Improving Surveys with Tax Data

¿Un problema antiguo?

- Eliminar las inconsistencias entre las encuestas y otras fuentes de datos más confiables (e.g., censos) no es nada nuevo
- Las encuesta son corregidas rutinariamente para asegurar representatividad en terminos de edad, sexo, etc . . .

$\Rightarrow$ ¿Deberíamos entonces hacerlas también más representativas en términos de **ingresos**?

Calibración de encuestas

- Metodologías bien establecidas (e.g. Deville and Särndal, 1992)
- Variable en encuesta $x_1, \ldots, x_n$, con *design weights* $d_1, \ldots, d_n$
- Encontrar nuevos weights $w_1, \ldots, w_n$ que entreguen los totales observados en otra fuente:

$$\sum_{k=i}^{n} w_i x_x = T$$

- Condición: que minimize la distancia entre weights originales y corregidos:

$$\min_{w_1,\ldots,w_n} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}$$

$\Rightarrow$ Solución se interpreta cómo un modelo de *non-response*.

Software

Comandos de Stata `bfmcorr` (incl. `postbfm`)

- Flexibilidad en forma de datos, conceptos de ingreso, unidades estadisticas
- Selección automática de 'merging point' & funciones de extrapolación si necesario (datos limitados)
- Incluye varias herramientas de diagnóstico para realizar y analizar correcciones

Literatura sobre corrección de encuestas

- **Reweighting**
  - ⇒ Korinek, Mistiaen, and Ravallion (2006); Hlasny and Verme (2017; 2018); Alvaredo (2011, for Argentina)...
- **Replacing**
  - ⇒ Burkhauser et al. (2016); Piketty, Yang, and Zucman (2017); Chancel and Piketty (2017); Czajka (2017); DWP (2015); Alvaredo (2011, for U.S.); Burkhauser et al. (2018); Jenkins (2017)...
- **Hybrids**
  - ⇒ Bourguignon (2018); Medeiros et al. (2018); *nuestro método*

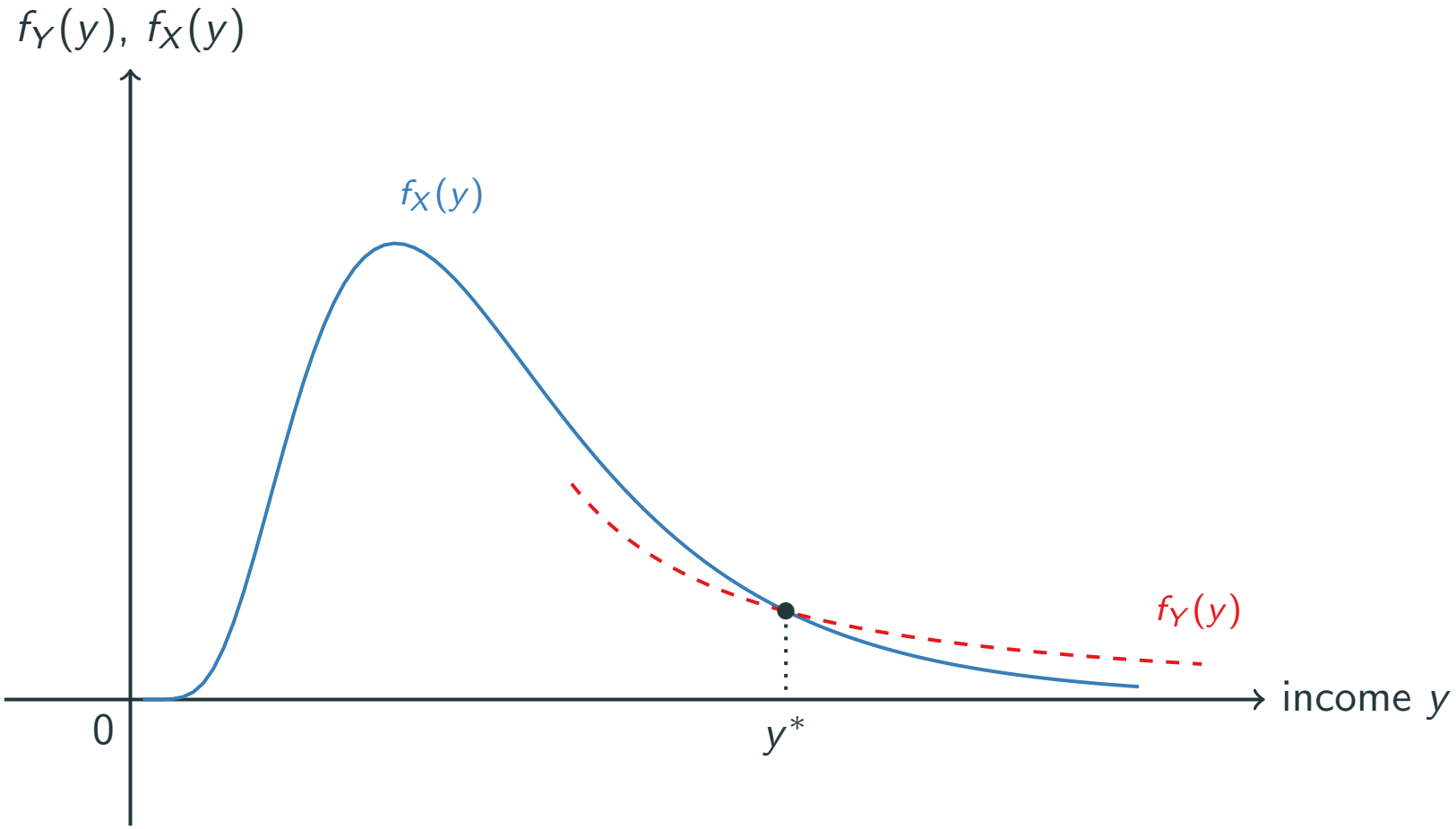
Nuestro método hibrido usa datos fiscales, encuentra un 'merging point' no-arbitrario y preserva la estructura de micro-datos, incluyendo totales de población (entre otros)

Teoría y metodología

Dos hipotesis fundamentales:

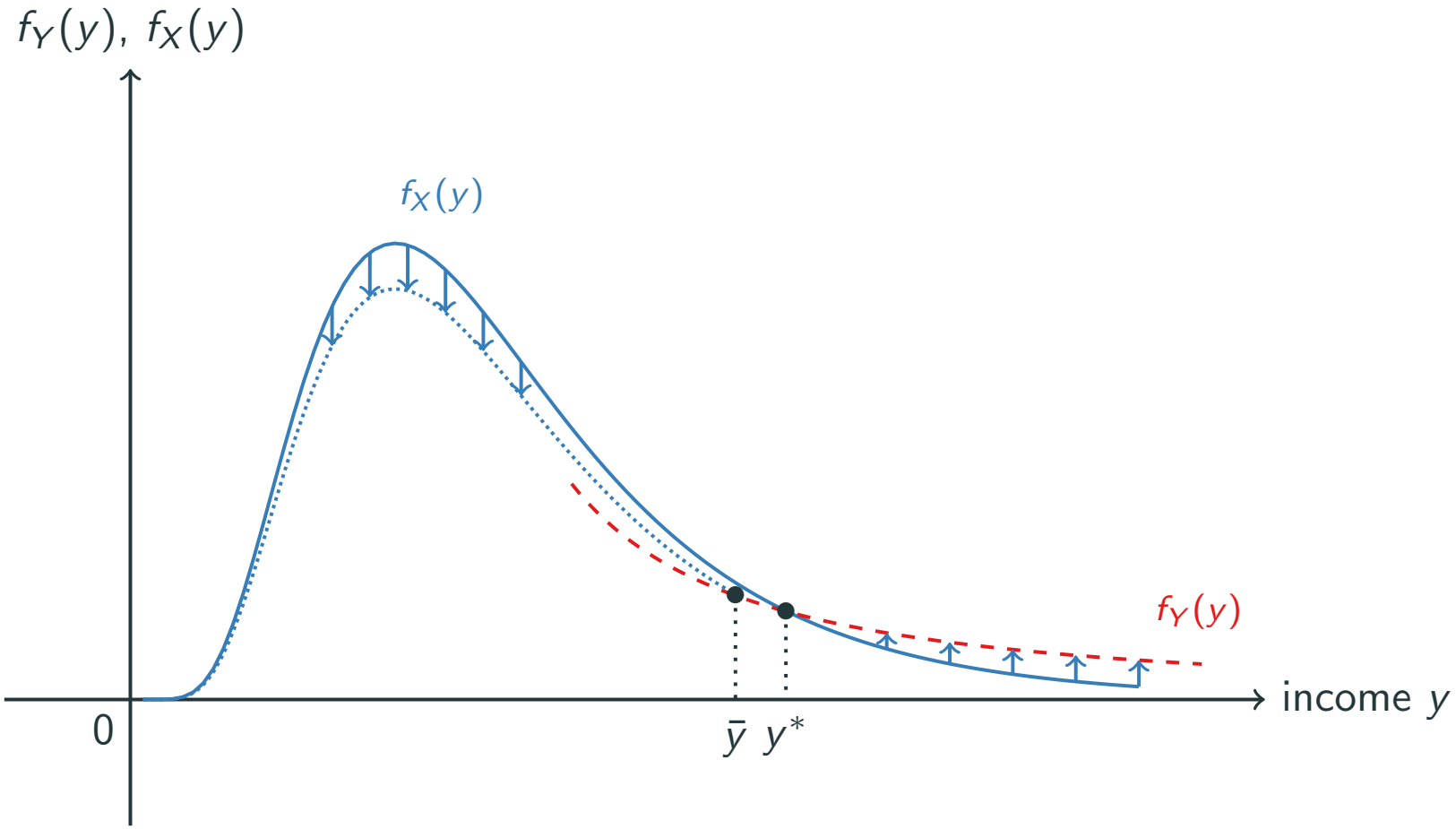
1. Declaraciones fiscales son un *lower bound* para las frecuencias en parte alta de la distribución
2. No cuestionamos las respuestas aportadas por individuos encuestados

Merging point



WID.WORLD

Merging point



WID.WORLD

Merging point

- Probabilidad relativa de respuesta

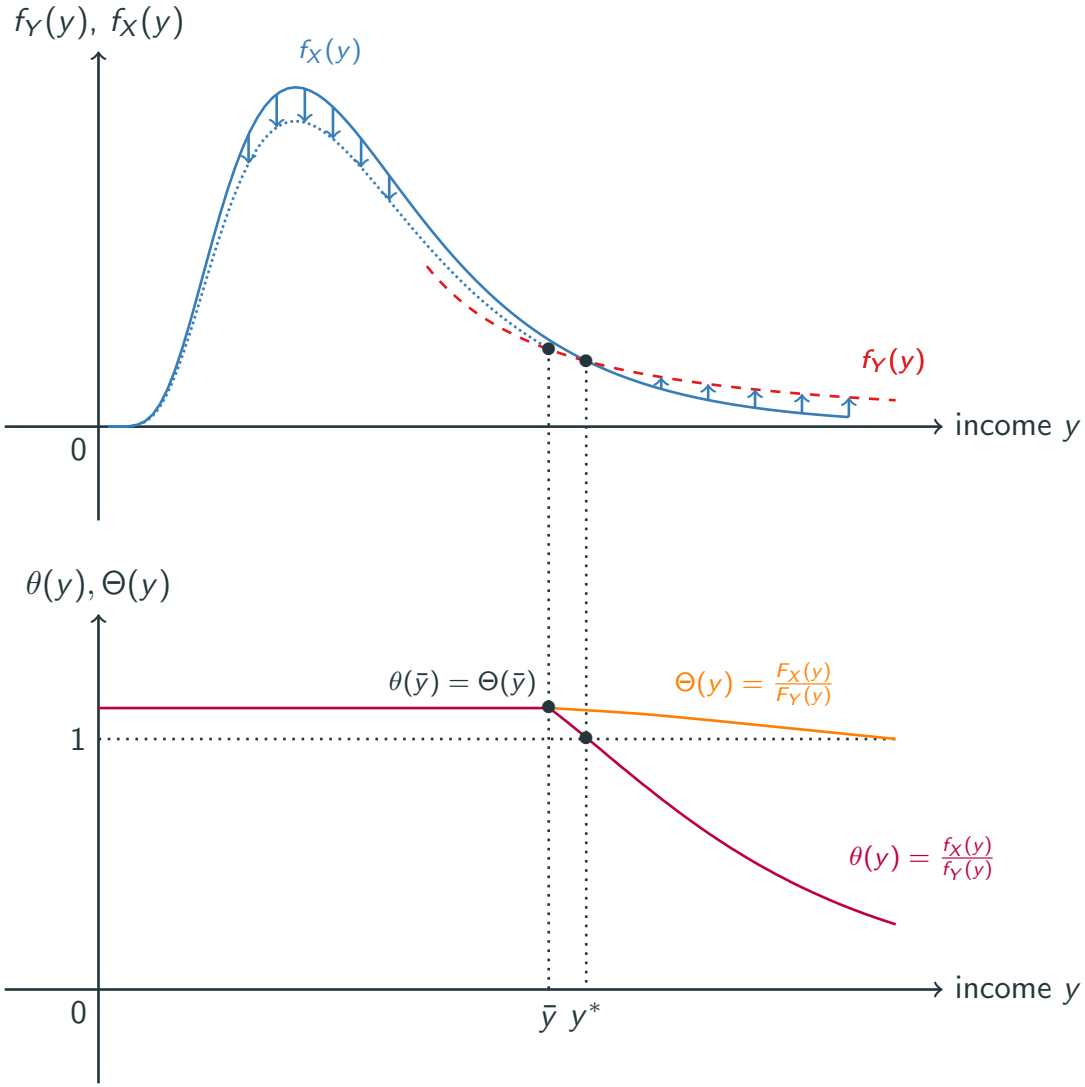
$$\theta(y) = \frac{f_X(y)}{f_Y(y)}$$

- Probabilidad relativa de respuesta cumulada

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)}$$

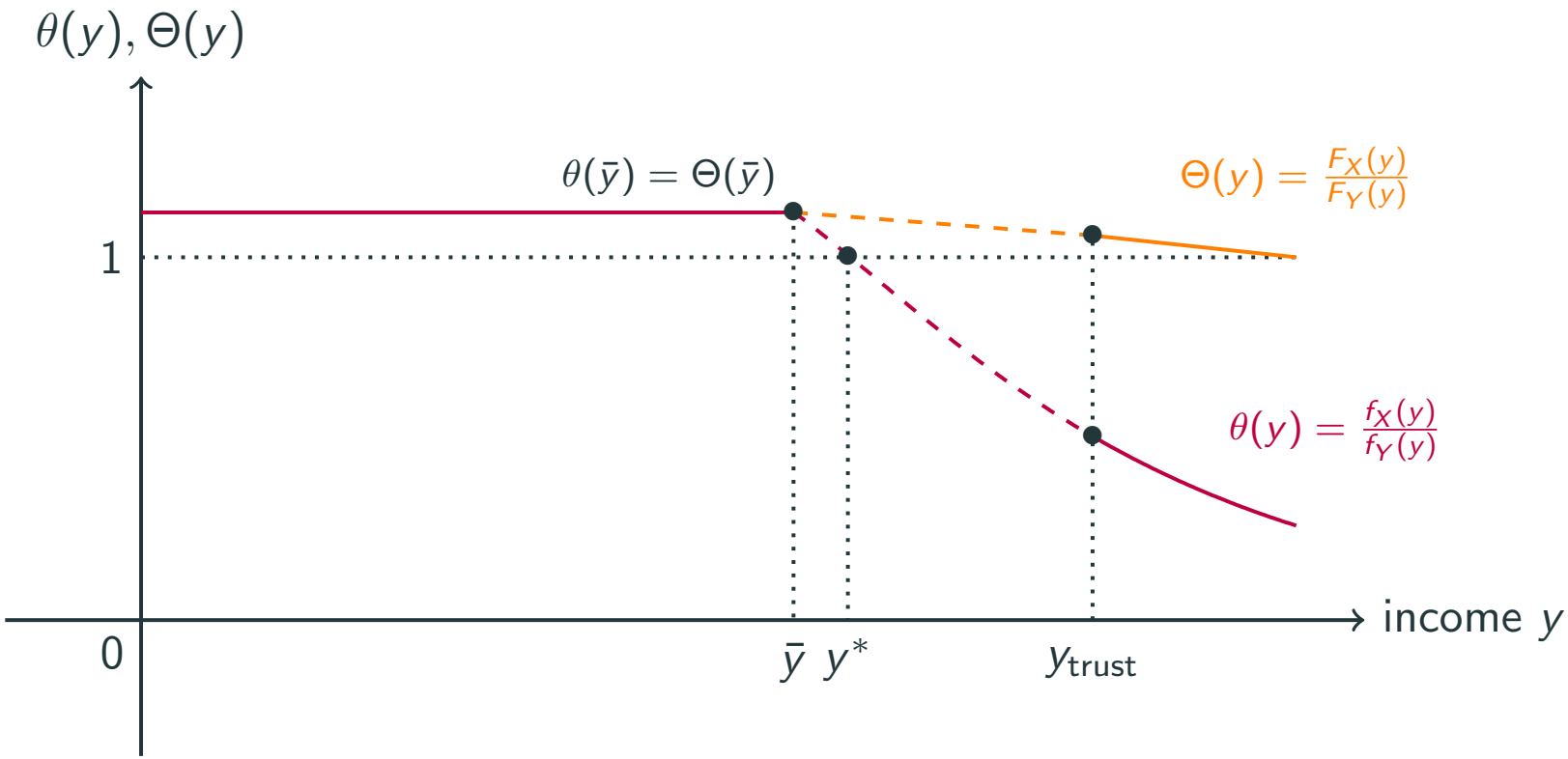
- Ambas se estiman empíricamente

Cesgo implícito y merging point



WID.WORLD

Datos fiscales limitados



WID.WORLD

Con datos fiscales limitados, merging point

Para extrapolar, asumimos la siguiente relación en el top:

$$\log \theta(y) = \beta_0 - \beta_1 \log y$$

Usando la *Ridge regression*:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{m} (\log \tilde{\theta}_k - \beta_0 - \beta_1 \log y_k)^2 + \lambda(\beta_1 - \beta_1^*)^2$$

Las estimaciones de β_1^* son calculadas a partir de datos reales

Calibración: Práctica

Aplicación a datos de ingreso

- Discretizar la distribución en grupos (percentiles...)
- ¿Cuanta gente en la encuesta para cada fractil?

La flexibilidad de los métodos de calibración permiten:

- ajustar weights respetando la consistencia macro de las variables (e.g. edad, sexo)
- introducir otros datos administrativos (caracteristicas, composición de ingresos, lidiar con distintas definiciones de ingreso)

Replacing and Matching

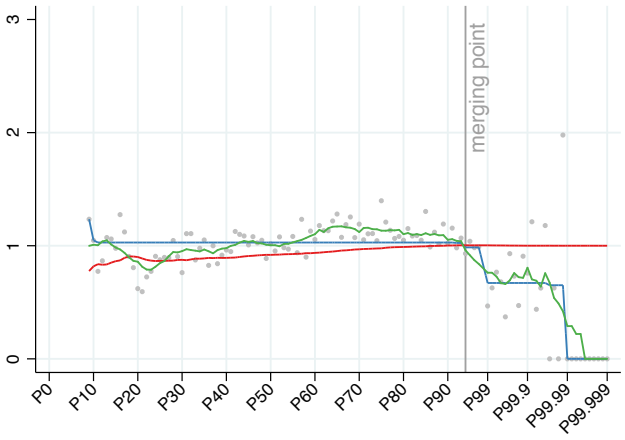
- We can gain more precision by replacing the survey distribution by the tax data distribution at the top.
- Addresses small-sample bias on top shares (Taleb and Douady, 2015).
- Duplicate observations and match them on their rank

  $\Rightarrow$ expand the support.
- Statistically:
  - Use the marginal distribution of income at the top provided by the survey data.
  - Keep the survey data for the marginal distribution of covariates.
  - Keep the survey data for the copulas (dependency) between income and the covariates.

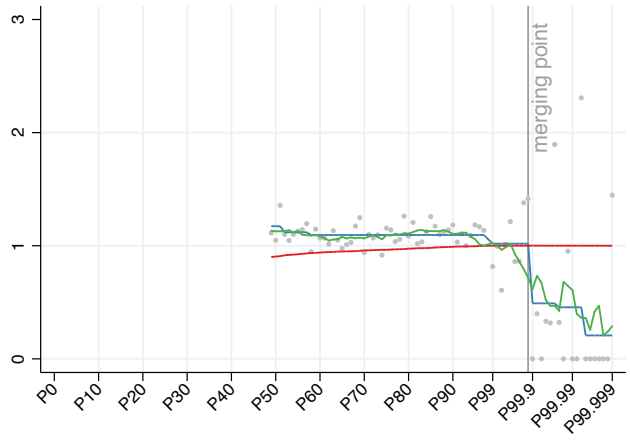
Aplicaciones

- Paso preliminar, armonizar definiciones (ingreso y unidad)
- Fuente: Encuestas locales (Chile, Brasil) y EU-SILCS (Francia, Noruega, Reino Unido)
- Series de tax: `http://wid.world`

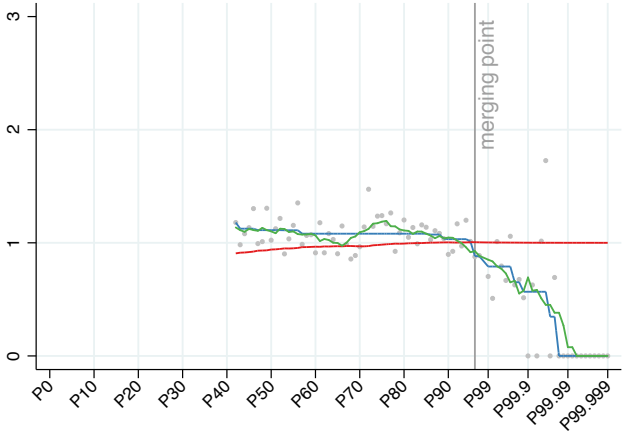
La forma del cesgo: Paises desarrollados



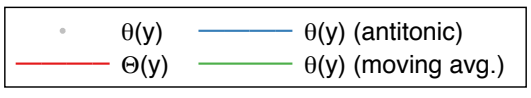
(a) Norway 2014



(b) France 2014

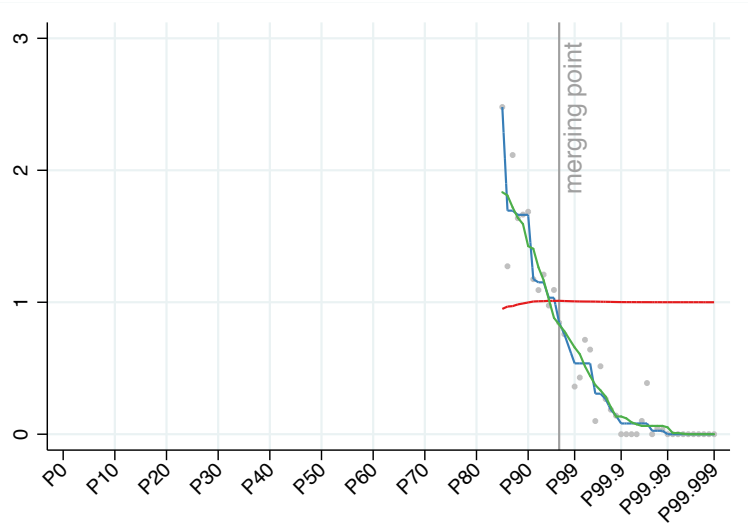


(c) UK 2014

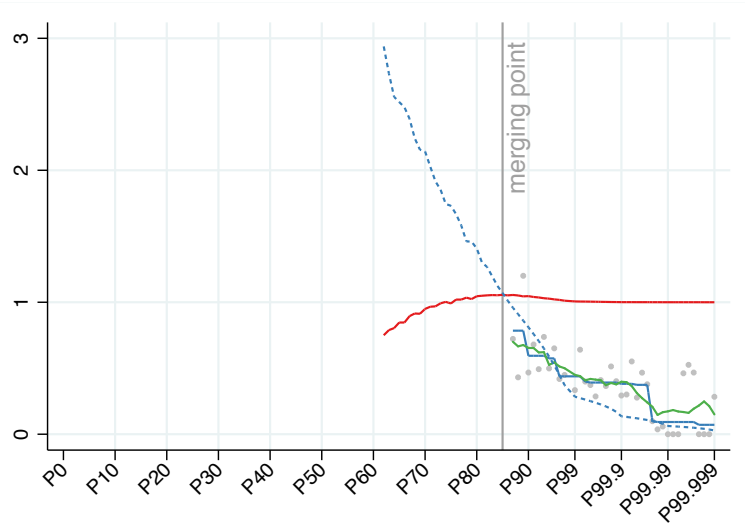


WID.WORLD

La forma del cesgo: paises en desarrollo



(e) Brazil 2015



(f) Chile 2015

θ(y)
θ(y) (antitonic)
Θ(y)
θ(y) (moving avg.)
θ(y) (extrapolation)

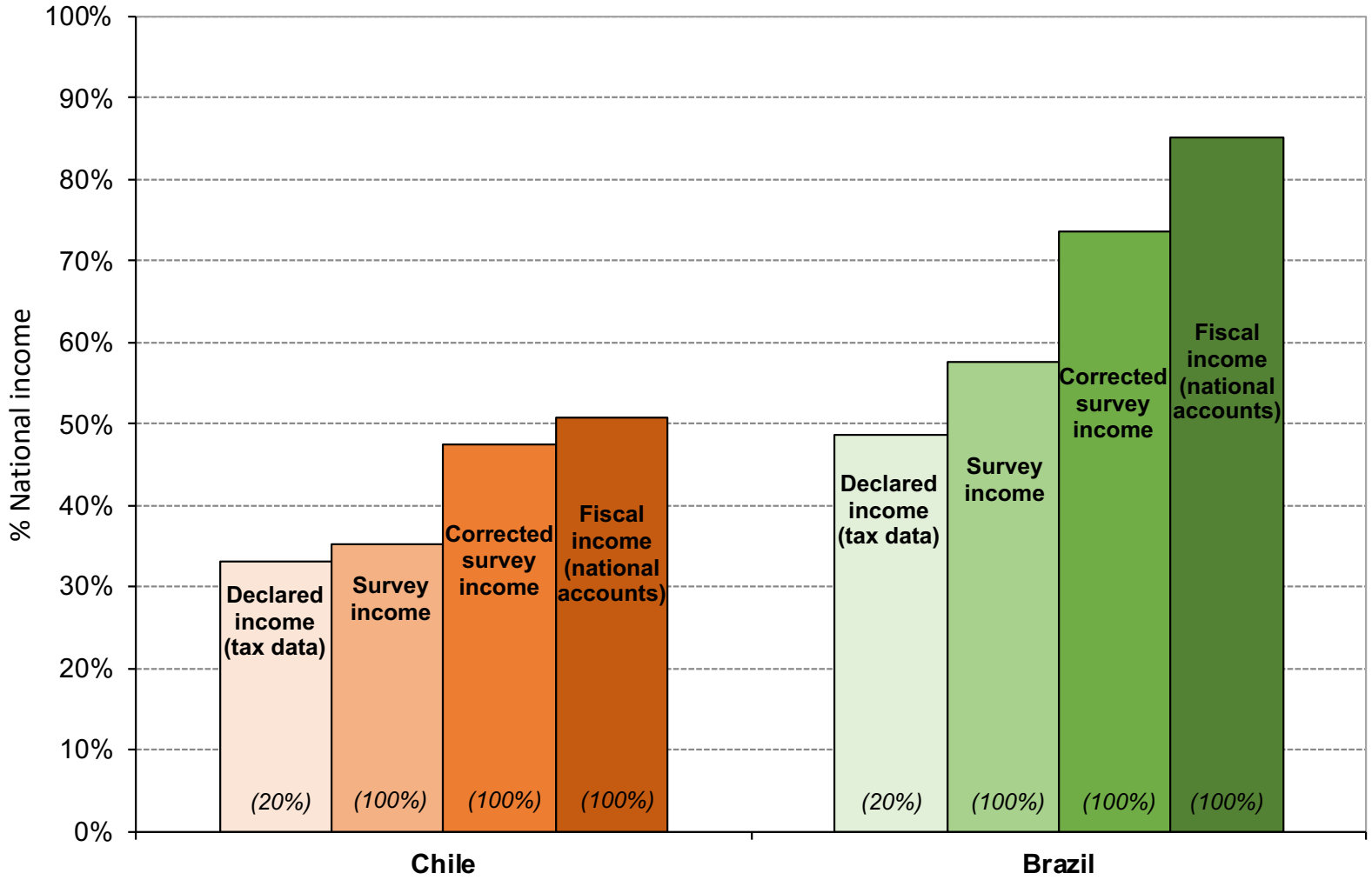
WID.WORLD

Población corregida

Table 1: Structure of Corrected Population: Latest Year

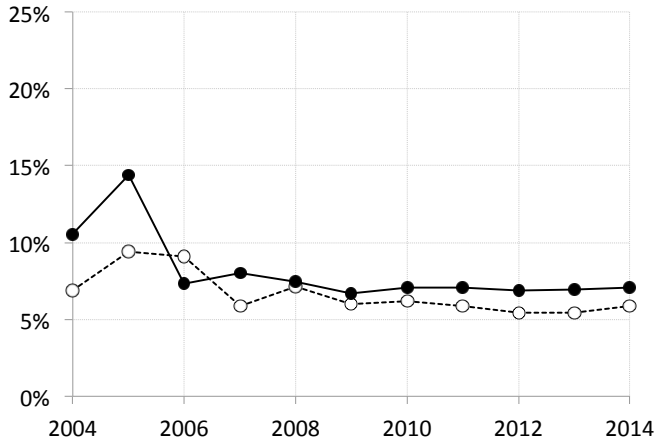
Country	Population over Merging Point (% total population)		Corrected population		
	Tax data [2]	Survey [3]	Total [4] = [2] − [3]	Share inside survey support [5]	Share outside survey support [6]
Chile	14.0%	9.2%	4.8%	99.99%	0.01%
Brazil	3.0%	1.9%	1.1%	98.2%	1.8%
UK	3.0%	2.5%	0.5%	93.6%	6.4%
Norway	5.0%	4.6%	0.4%	96.0%	4.0%
France	0.1%	0.05%	0.05%	99.0%	1.0%

Cobertura de Ingresos

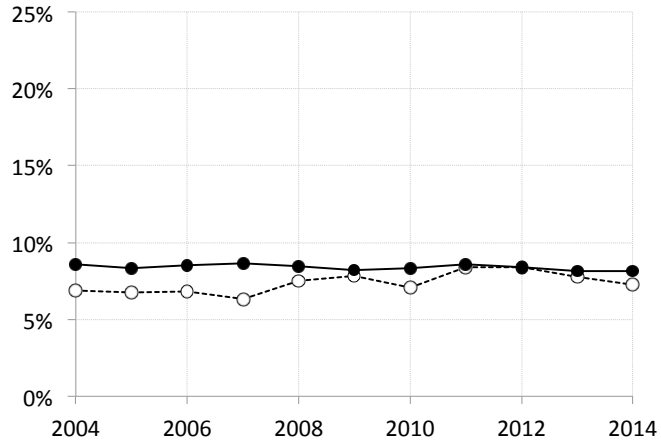


WID.WORLD

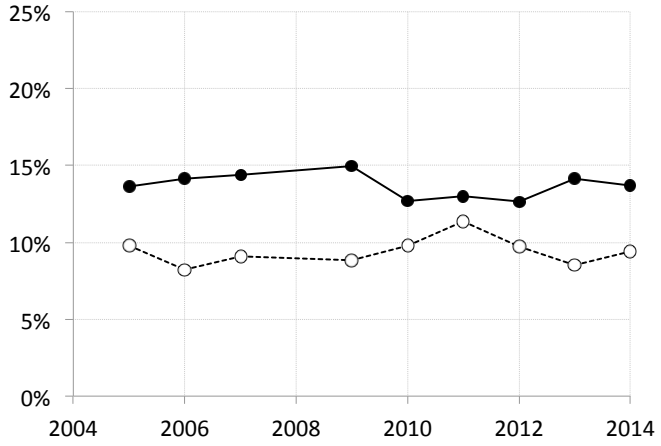
Resultados: Top 1% (paises desarrollados)



(h) Norway 2014



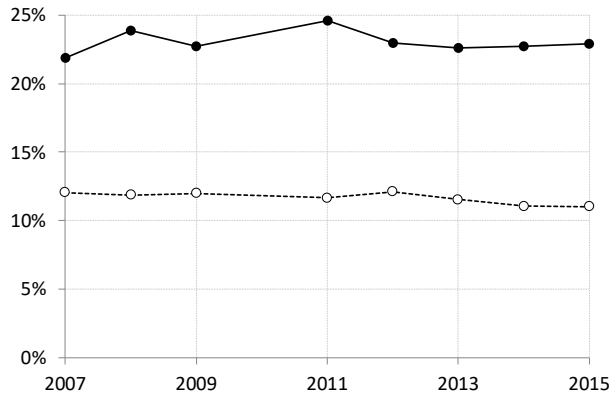
(i) France 2014



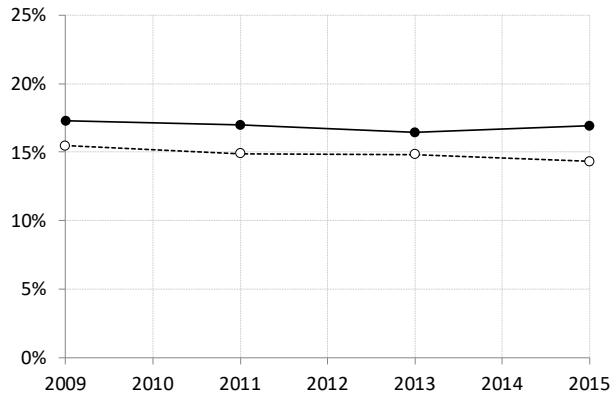
(j) UK 2014

WID.WORLD

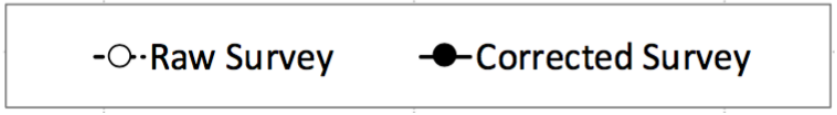
Resultados: Top 1% (paises en desarrollo)



(l) Brazil 2015



(m) Chile 2015



WID.WORLD

Resumen

Puntos fuertes

1. Data-driven approach
2. Observaciones razonables al nivel de la observación
3. Minimiza distorciones
4. Mantiene la representatividad en el uso de microdatos

Growing Unequal: The Distribution of Economic Prosperity in Latin America

Datos LatAm (1990-2015)

Encuestas a hogares

- CEPAL aplicación con micro-datos
- Datos armonizados para 18 paises = +95% de la población regional
- 14 con encuestas desde fin de 1980s/principios 1990s, 4 desde fines de 1990s/comienzos 2000s

Tax

- Disponible: Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Uruguay
- Pendiente: Ecuador, Peru, Bolivia
- Obstaculos: definiciones de ingreso

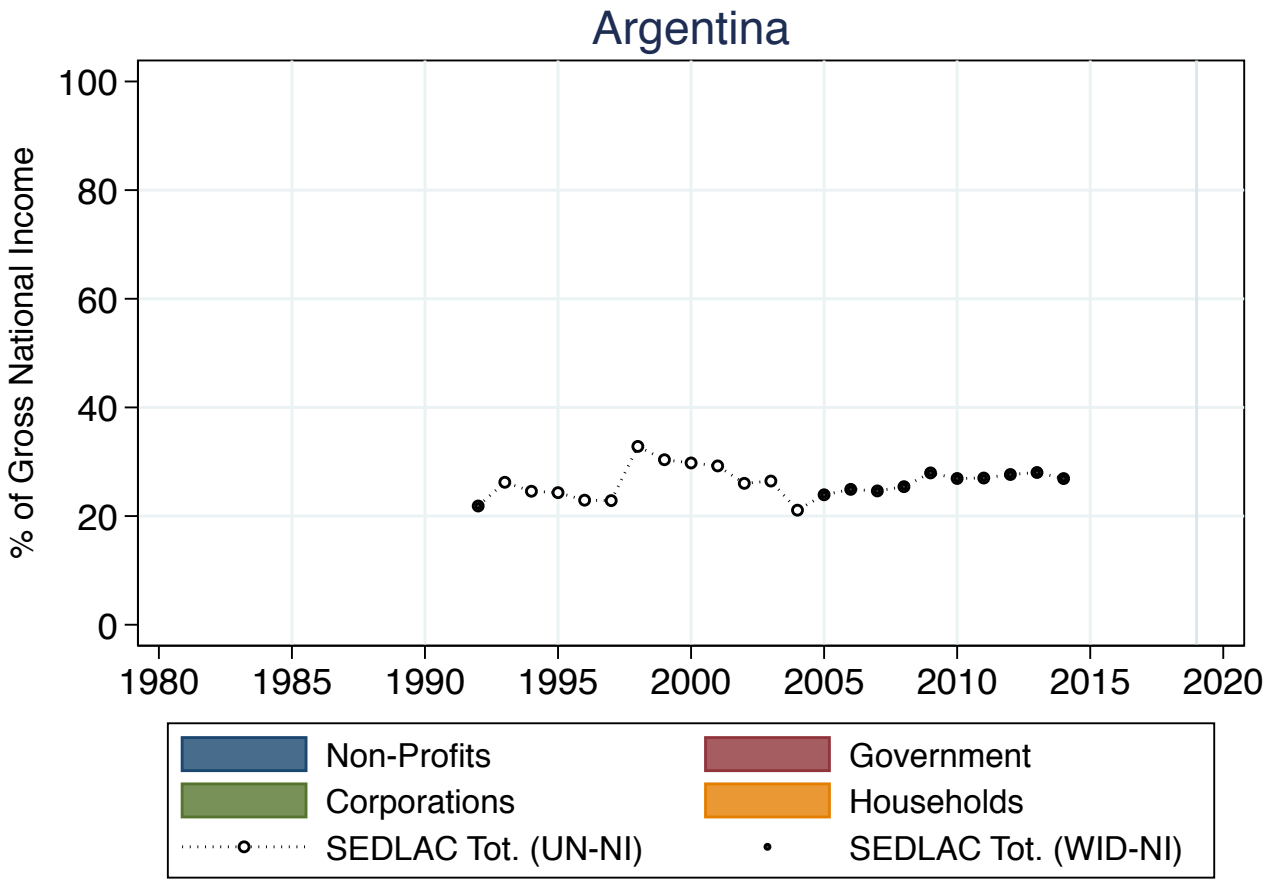
Cuentas nacionales

- Descomposición sectoral desde mediados de 1990s
- Nivel de detalle y cobertura temporal variable

Data coverage

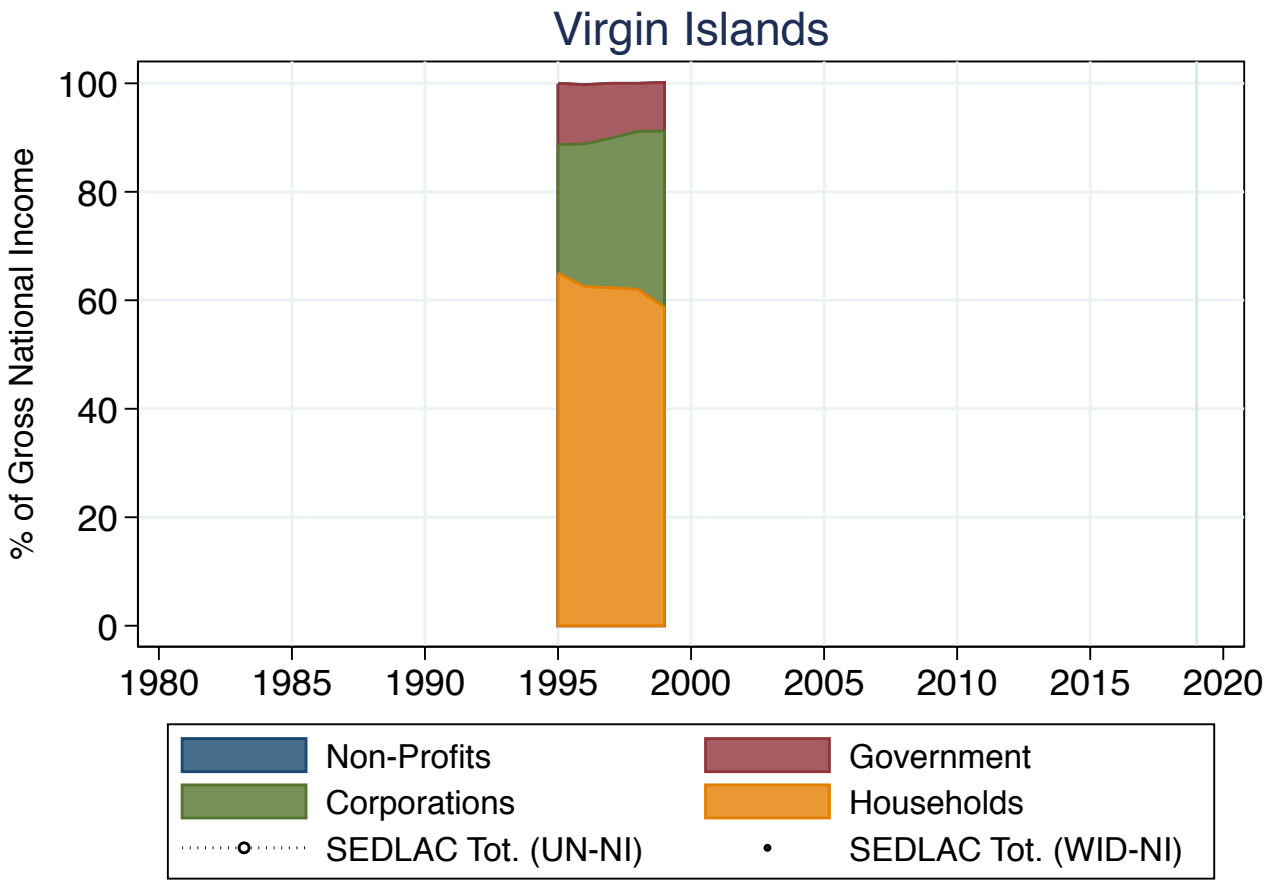
[LINK]

Filling the national accounts gap

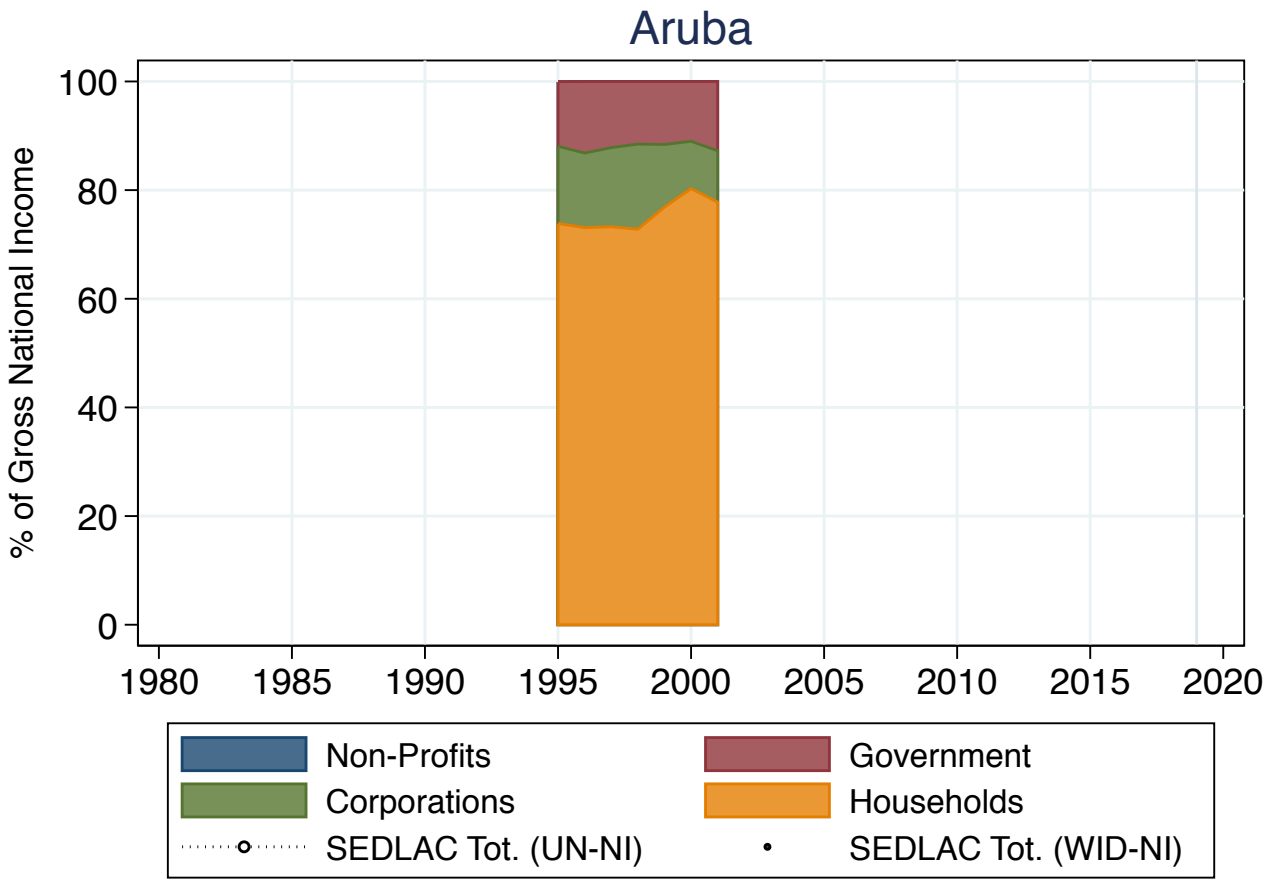


WID.WORLD

Filling the national accounts gap

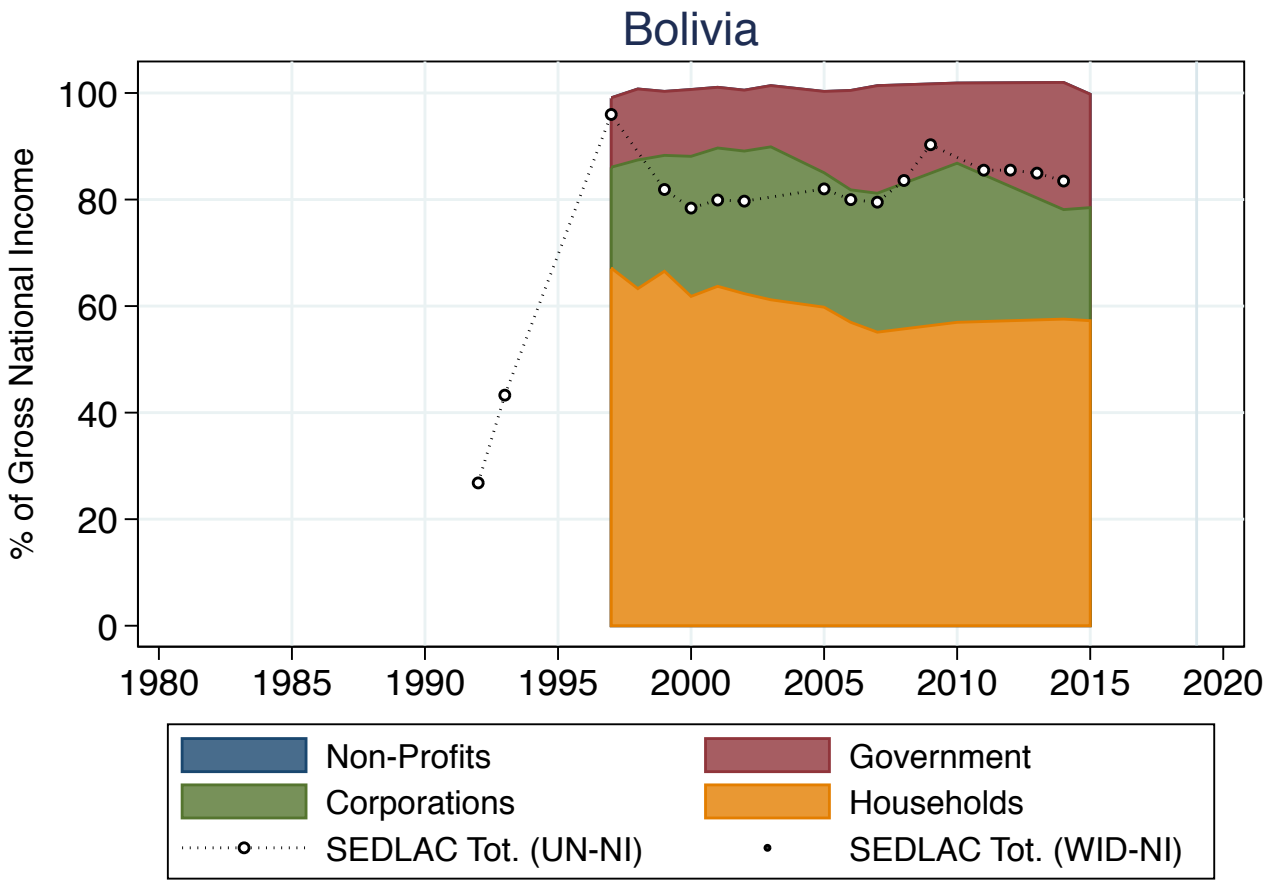


Filling the national accounts gap



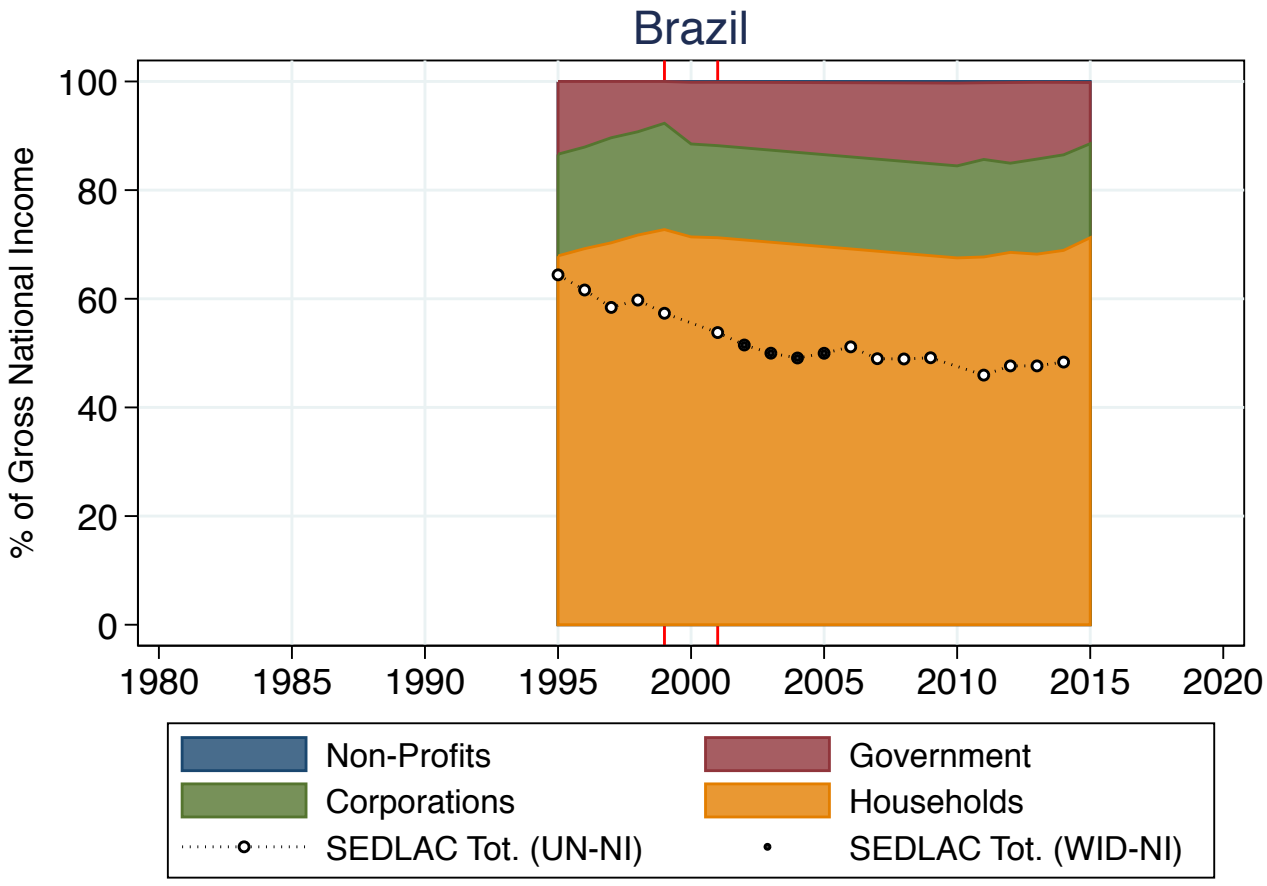
WID.WORLD

Filling the national accounts gap



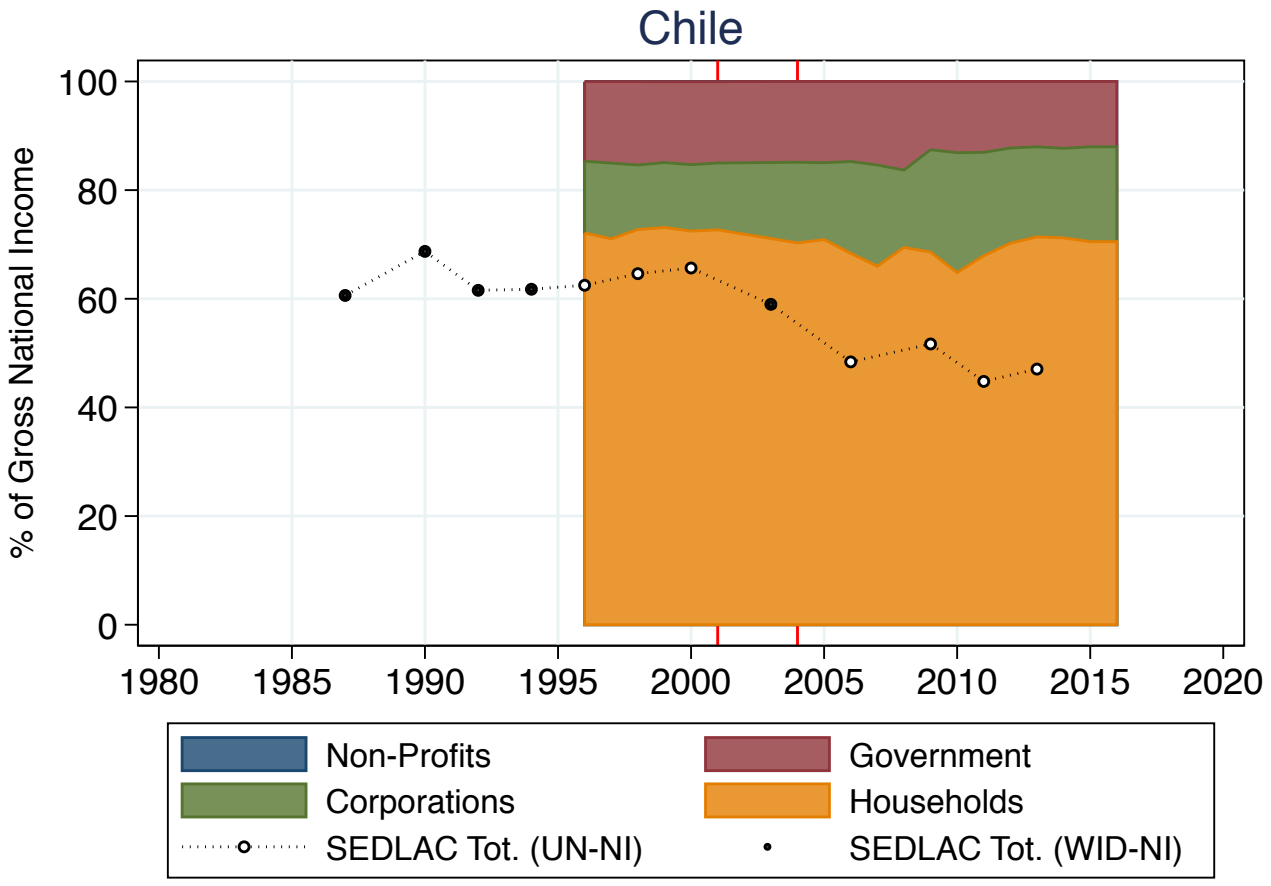
WID.WORLD

Filling the national accounts gap



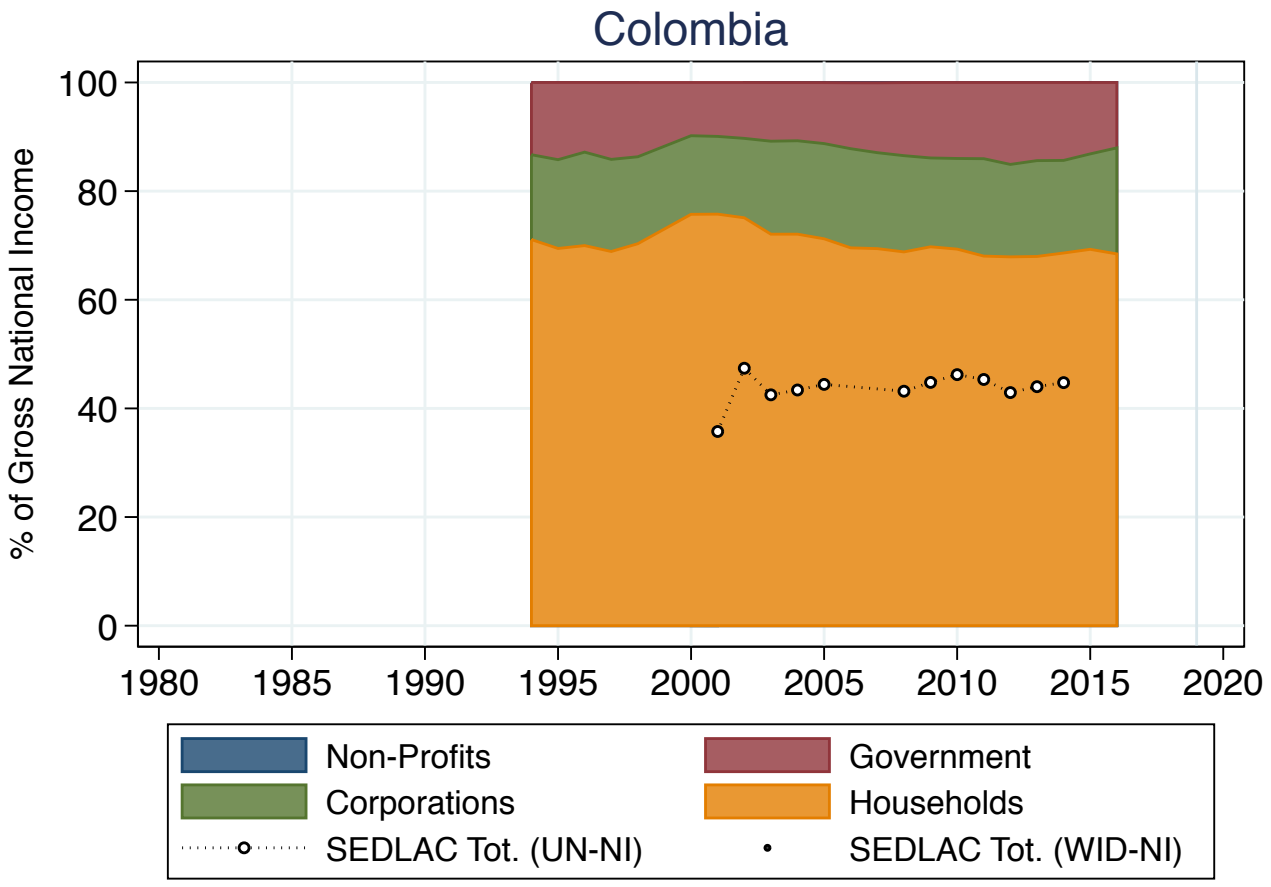
WID.WORLD

Filling the national accounts gap



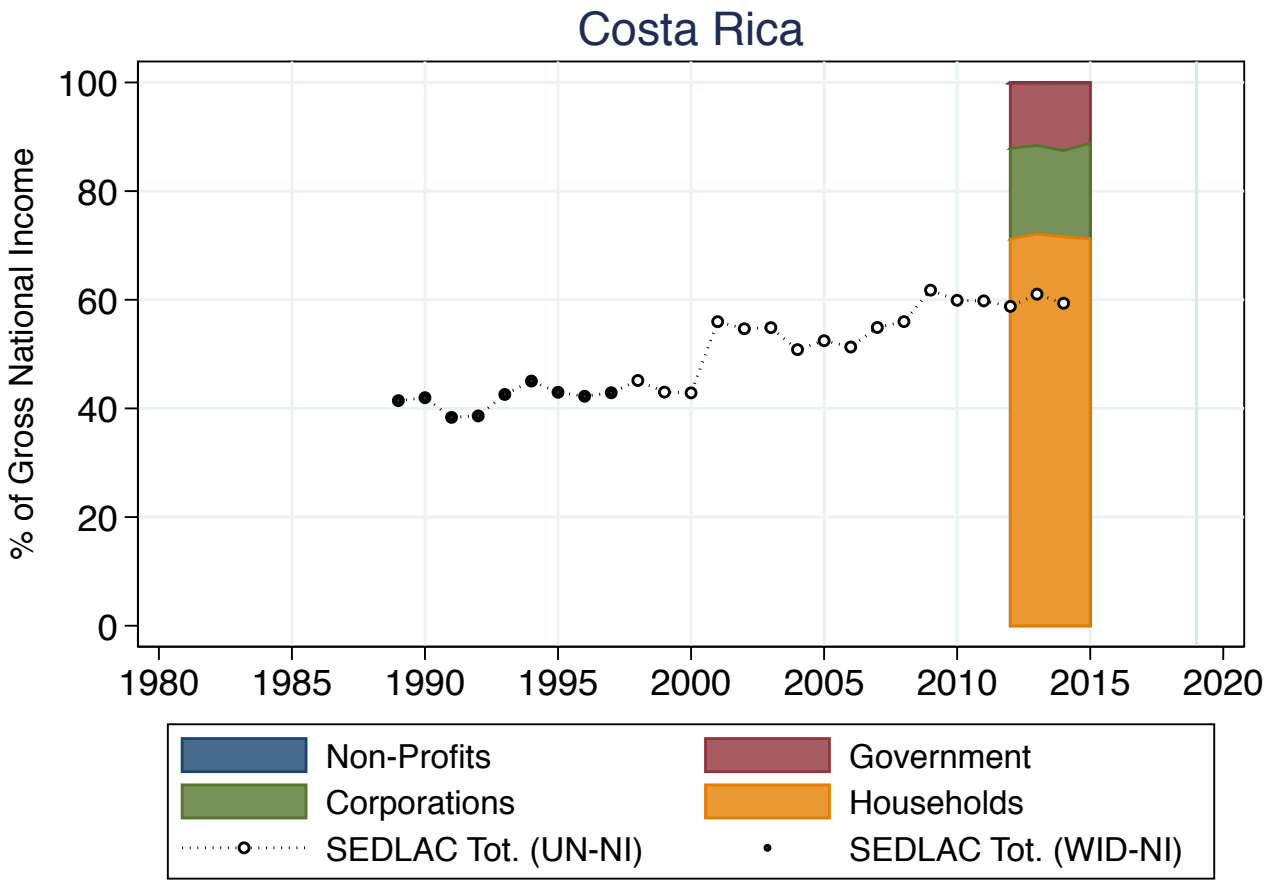
WID.WORLD

Filling the national accounts gap



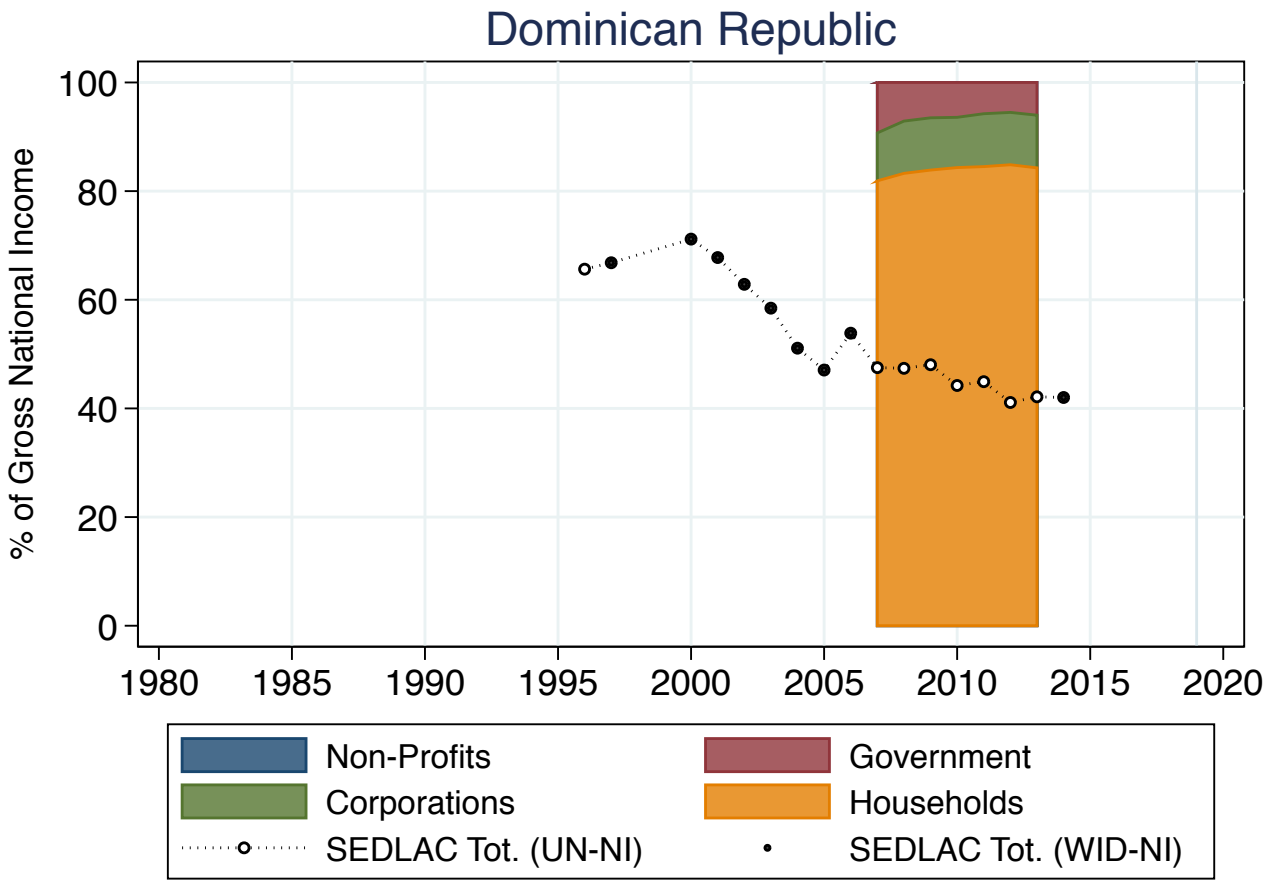
WID.WORLD

Filling the national accounts gap



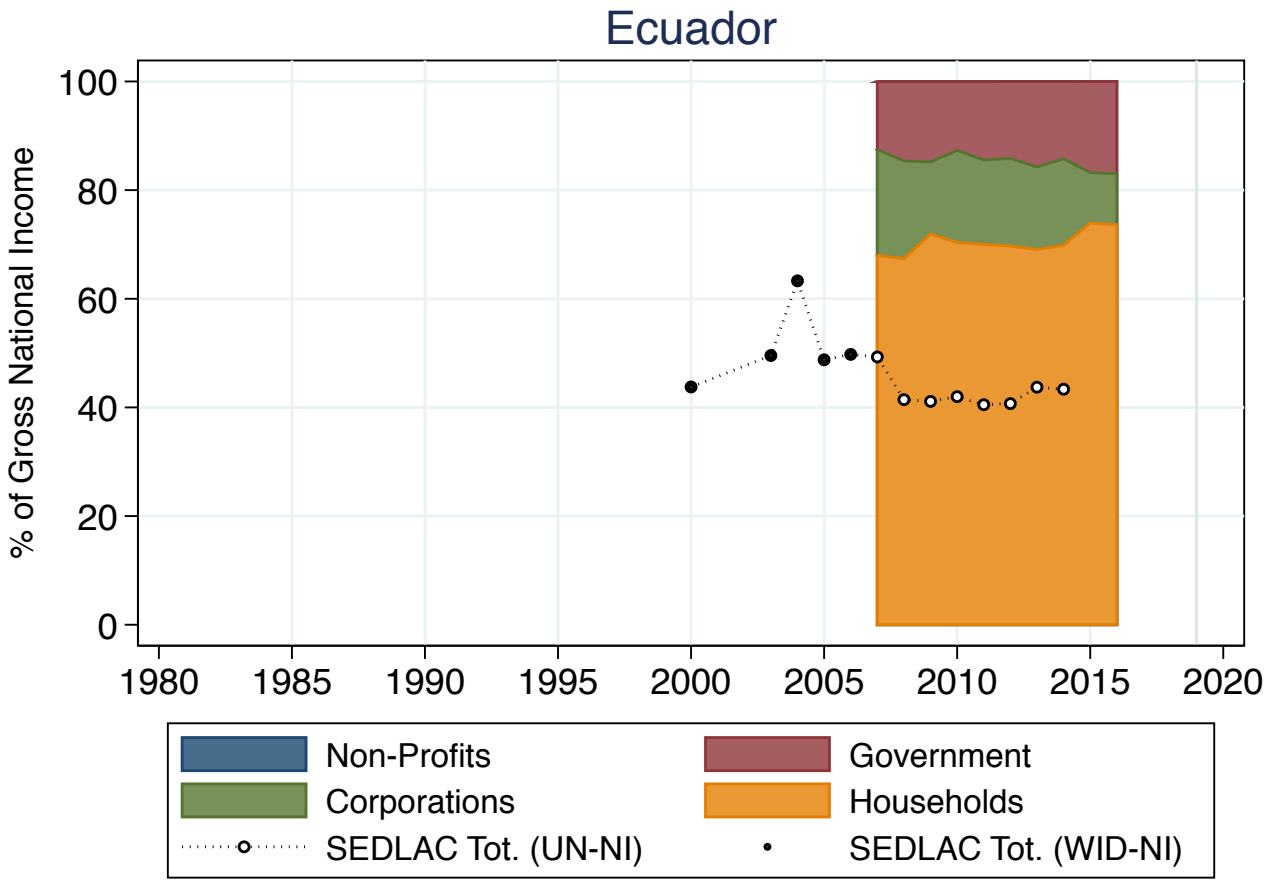
WID.WORLD

Filling the national accounts gap



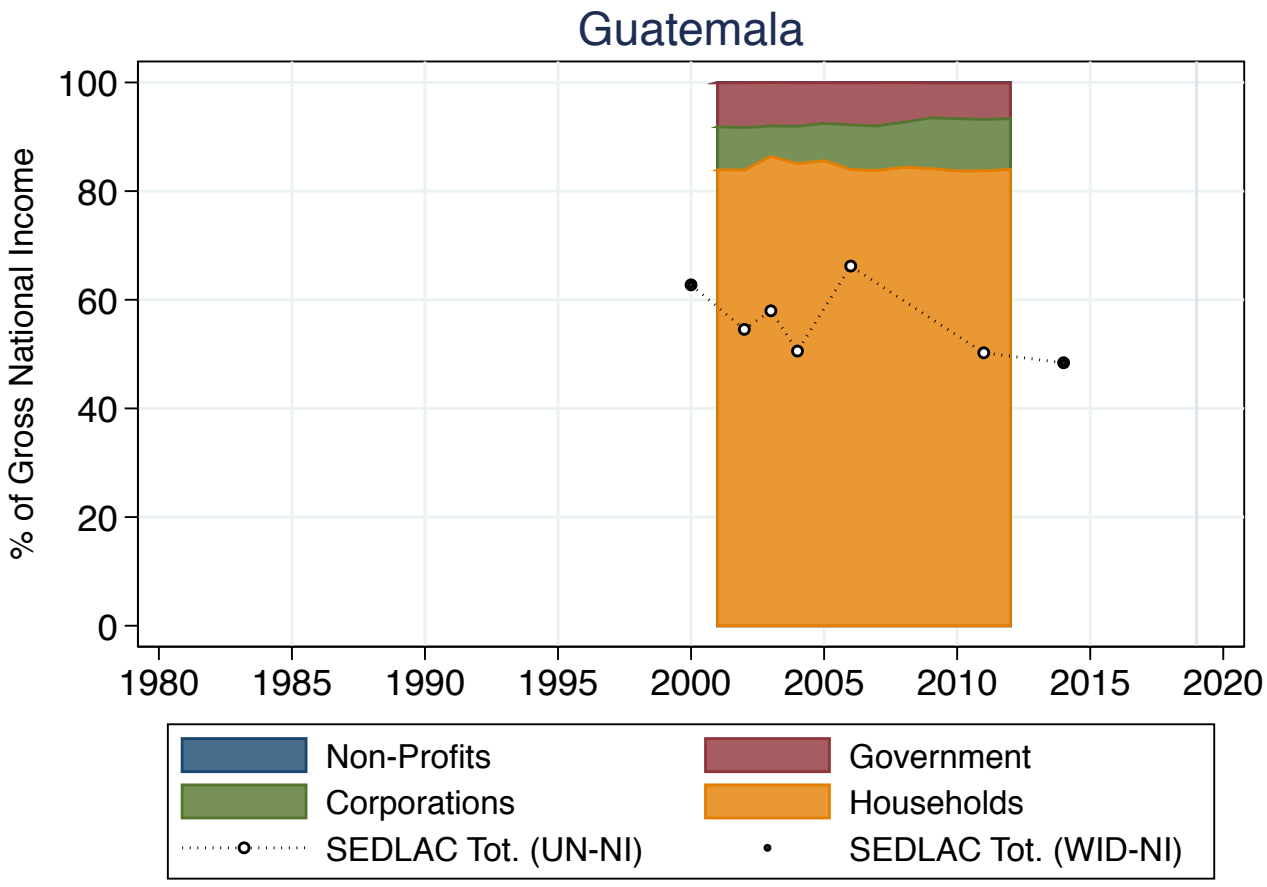
WID.WORLD

Filling the national accounts gap



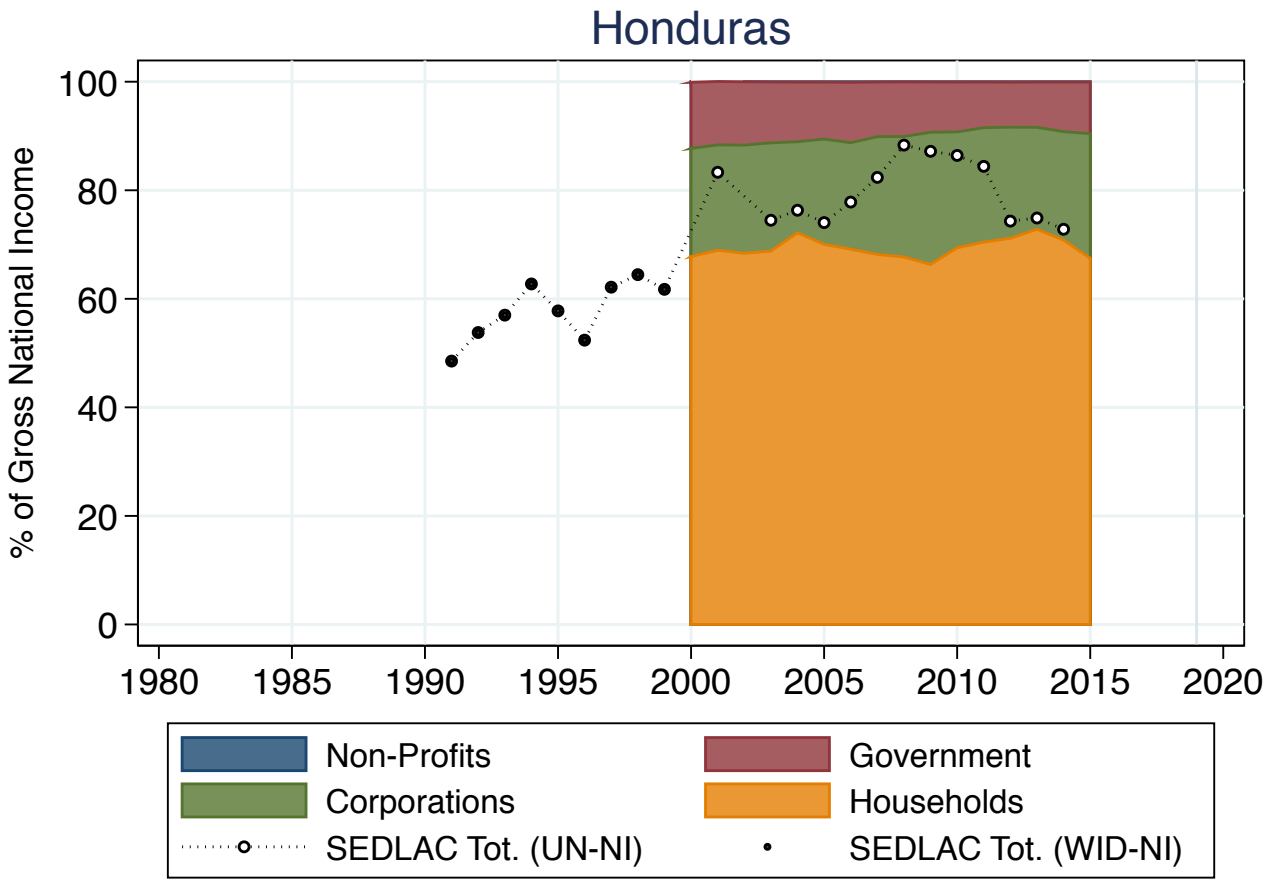
WID.WORLD

Filling the national accounts gap



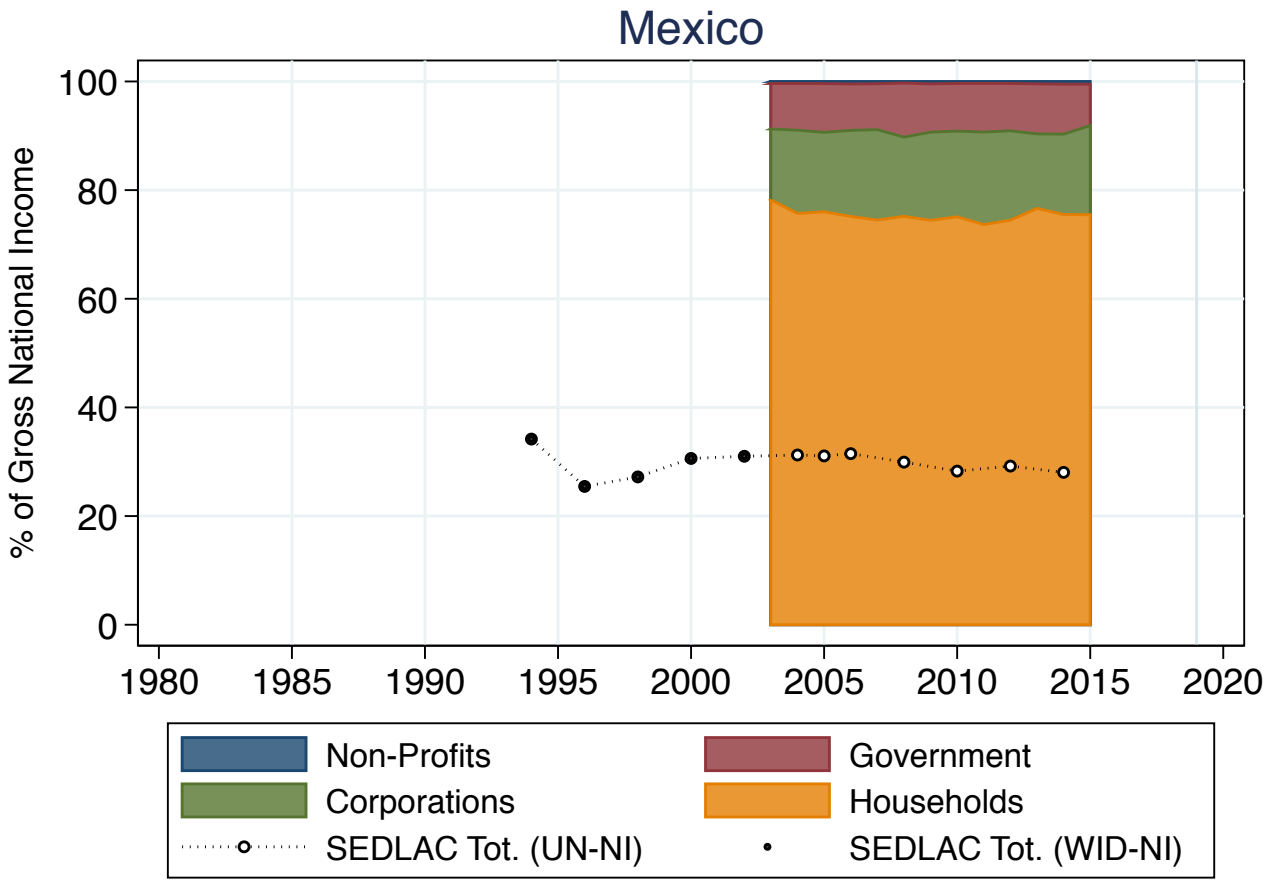
WID.WORLD

Filling the national accounts gap



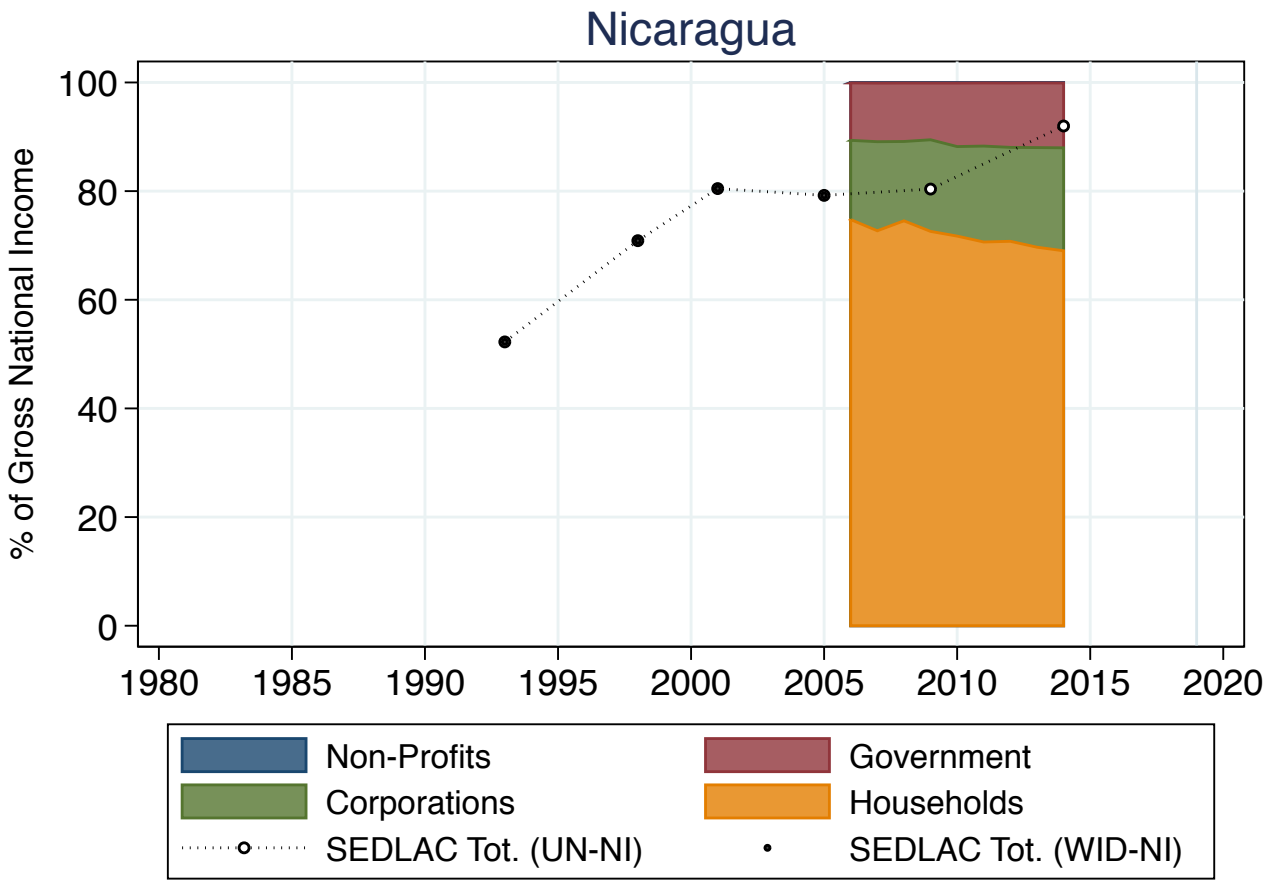
WID.WORLD

Filling the national accounts gap



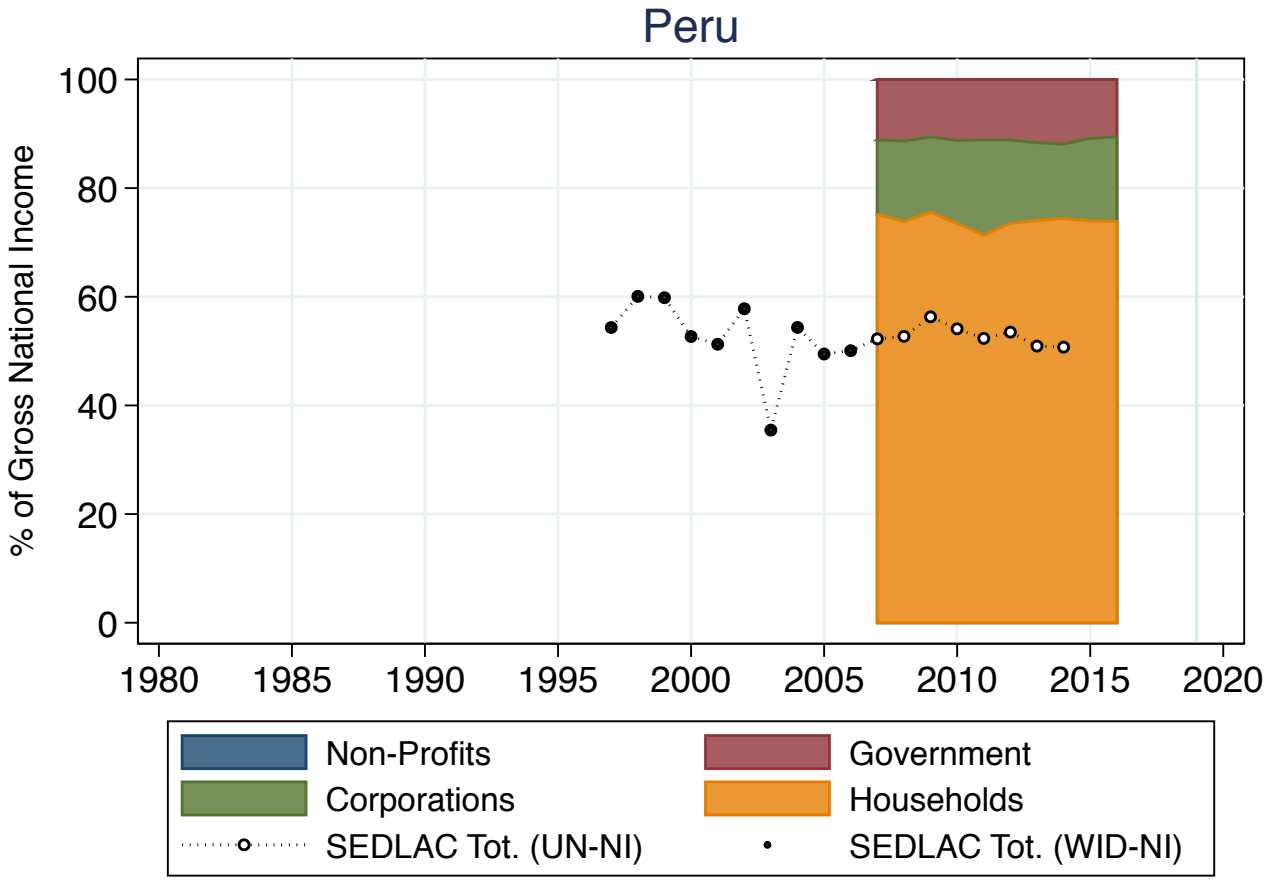
WID.WORLD

Filling the national accounts gap



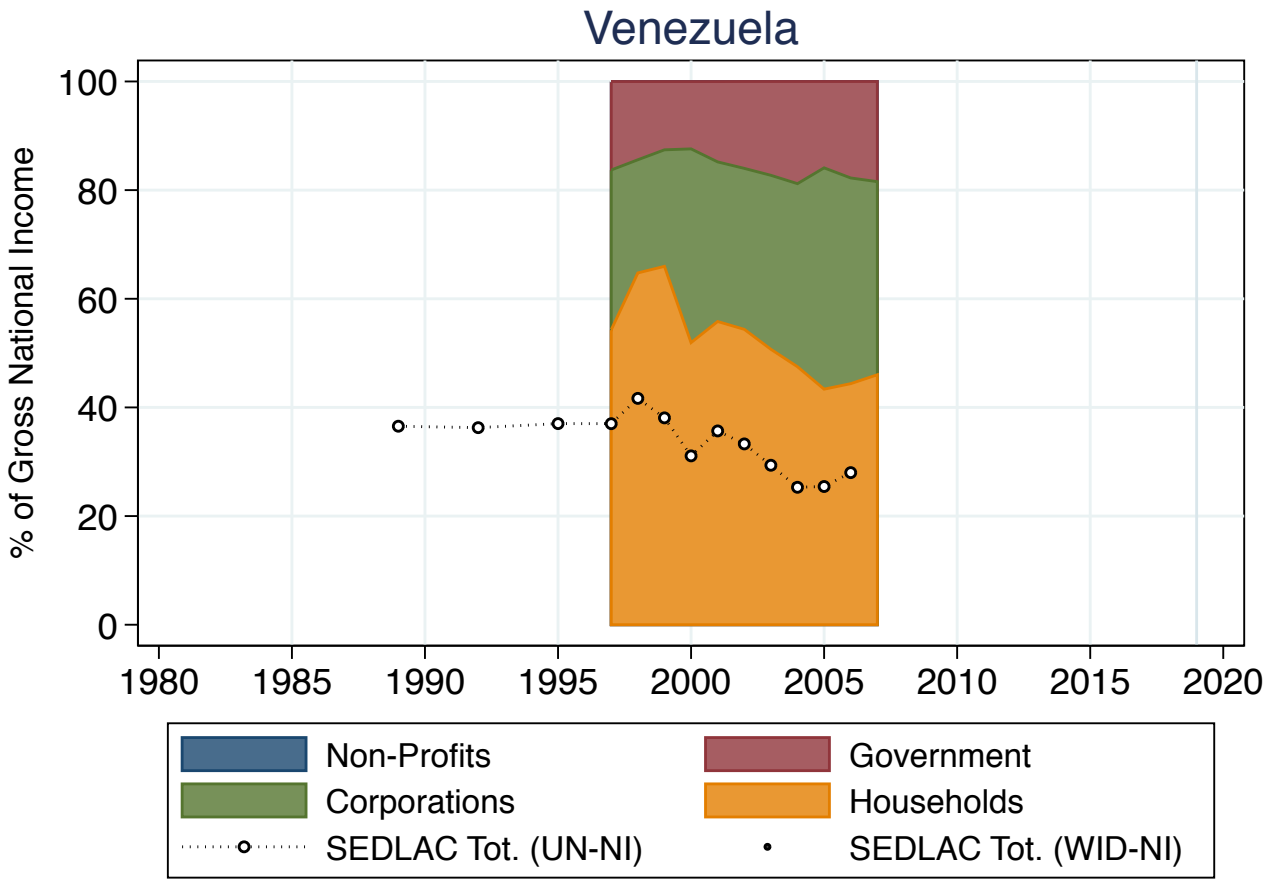
WID.WORLD

Filling the national accounts gap



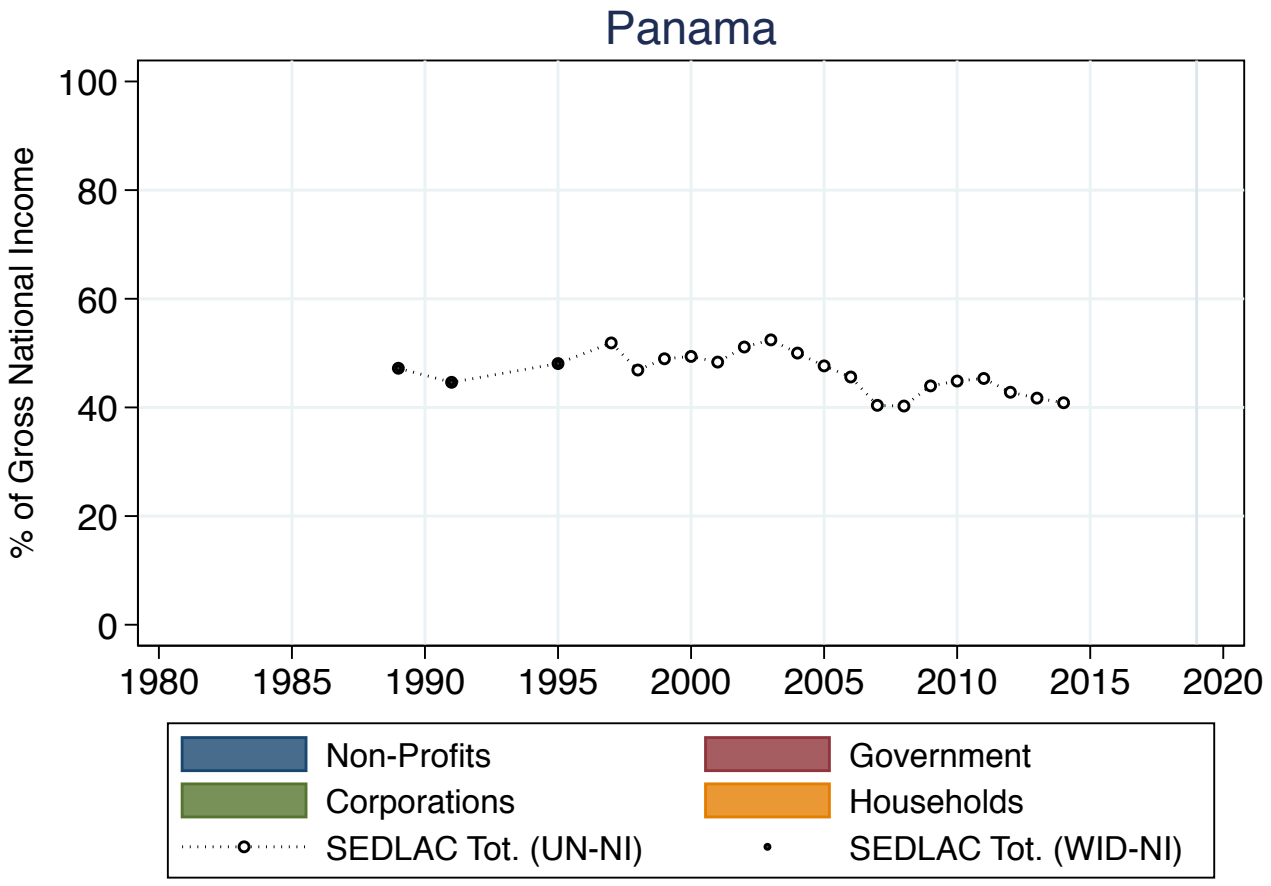
WID.WORLD

Filling the national accounts gap



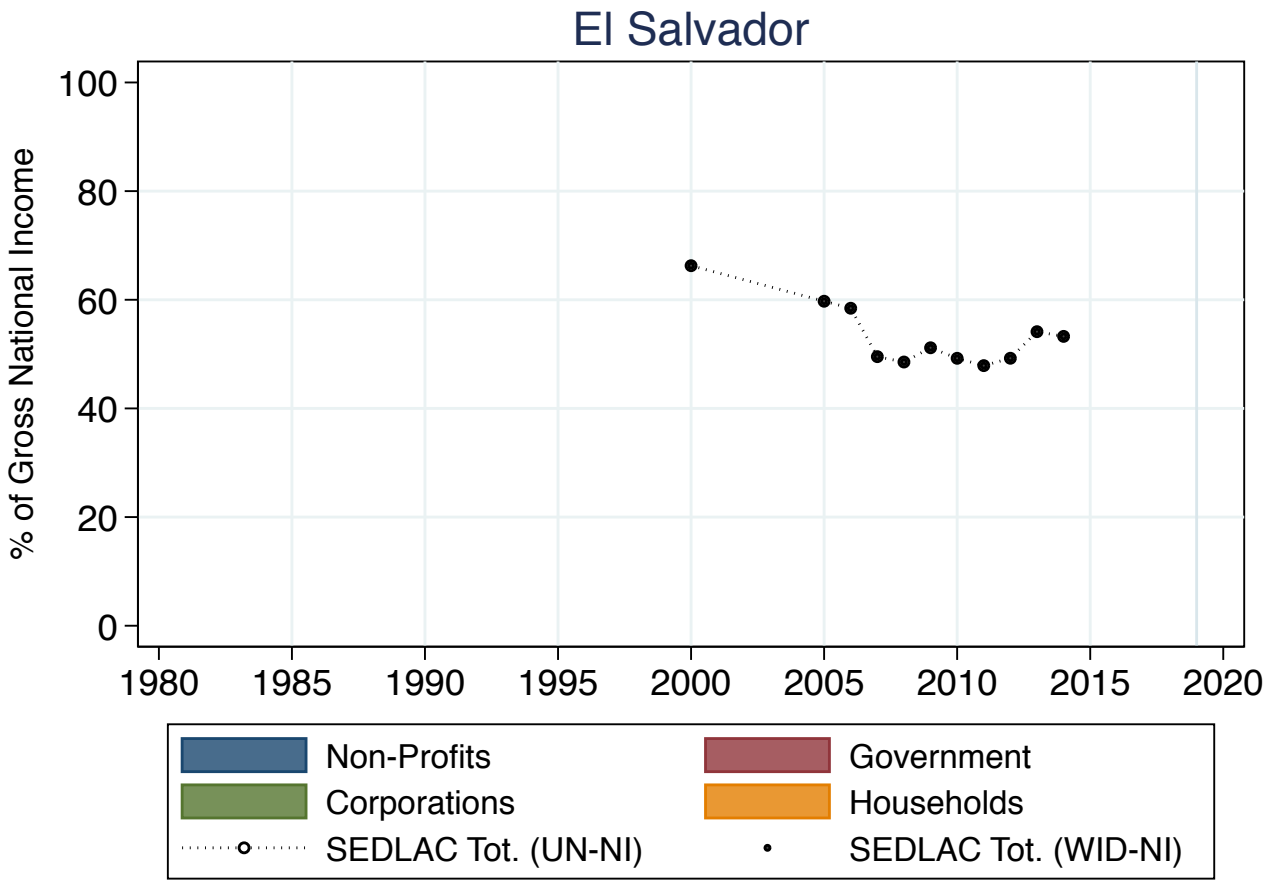
WID.WORLD

Filling the national accounts gap



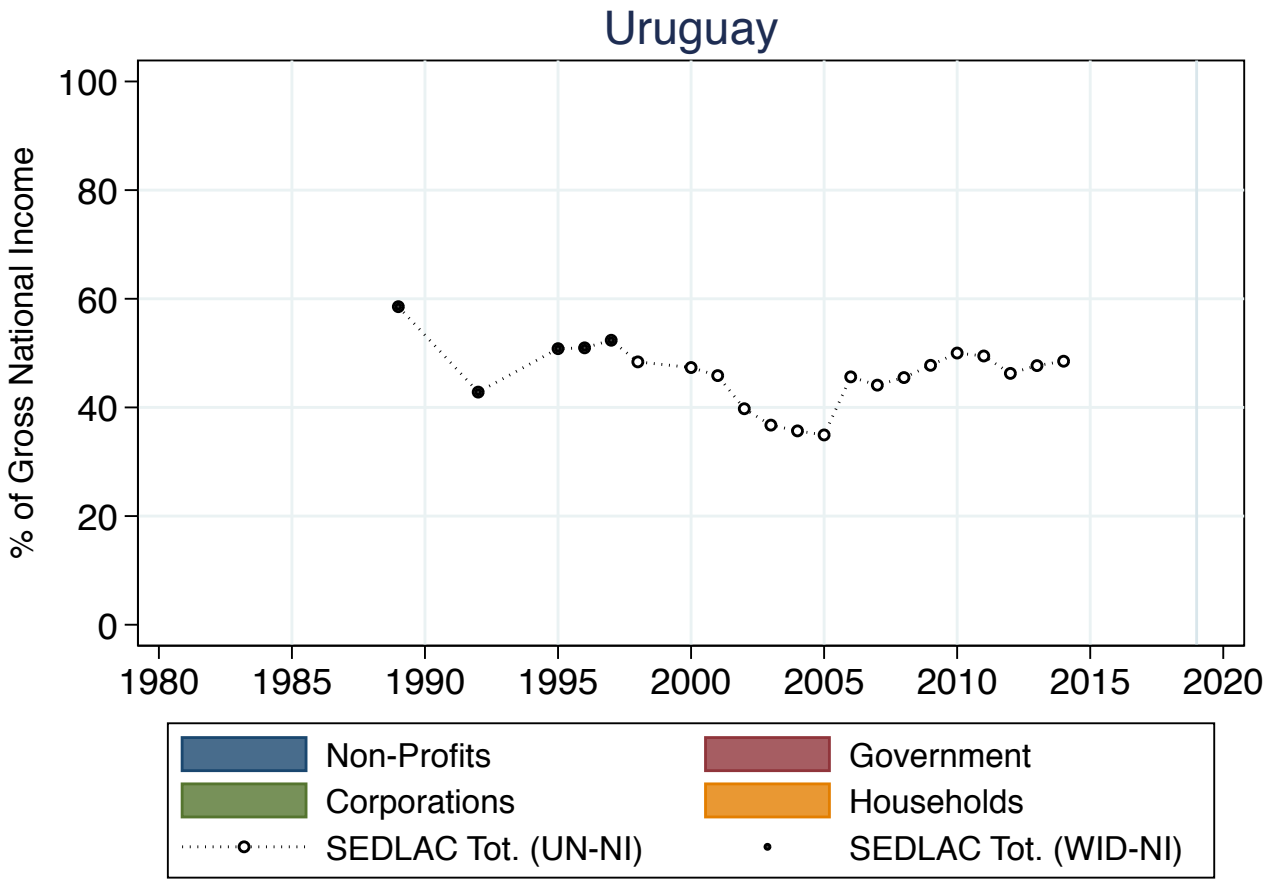
WID.WORLD

Filling the national accounts gap



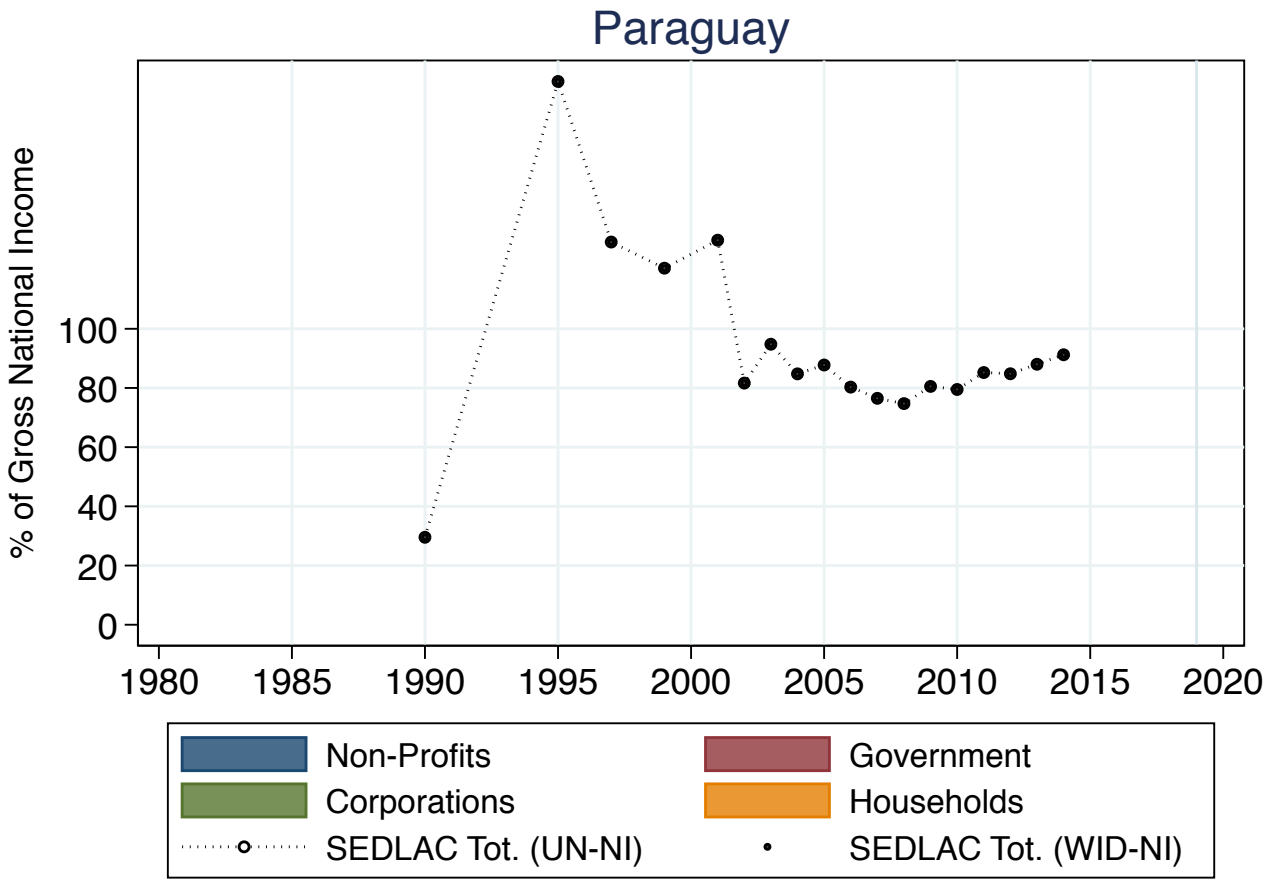
WID.WORLD

Filling the national accounts gap



WID.WORLD

Filling the national accounts gap



WID.WORLD

Gracias!