Social Divisions and Conflict

18th Winter School on Inequality and Social Welfare

Laura Mayoral

IAE and Barcelona School of Economics

Alba de Canazei, January 10, 2025

This lecture, I

- Based on a chapter for the *Handbook of the Economics of Conflict*, edited by Dube, Morelli, Ray and Sjostrom.

- joint with Joan Esteban (IAE, CSIC)

This lecture, I

- Based on a chapter for the *Handbook of the Economics of Conflict*, edited by Dube, Morelli, Ray and Sjostrom.
- joint with Joan Esteban (IAE, CSIC)

+

- "Economic development in pixels: The limitations of nightlights and new spatially disaggregated measures of consumption and poverty", with John Huber (Columbia U.)

This lecture, II: What triggers internal conflict?

- Drivers of conflict, specifically:

  What's the connection between conflict and the society wide distribution of individual characteristics?

- Economic and non-economic characteristics
  - Economic markers: distribution of income, wealth ...
  - Non-economic markers: ethnic, religious, linguistic composition in a society
  - Interactions of both types of markers

Conflict and distribution: Research questions

1. "Class" conflict:
   - Is economic inequality a driver of conflict?
   - Is conflict born of economic difference (inequality) or economic similarity?

2. Ethnic conflict: distribution of ethnic, religious, linguistic traits.
   - Are "ethnic divisions"—broadly defined to include racial, linguistic, and religious differences— a potential driver of conflict?
     - If so, economic motives or ancestral hatreds?
     - If so, what are the characteristics of the groups that fight?

3. Are there interactions between the income distribution and the ethnic distribution that are relevant for understanding social conflict?

4. How can we construct new measures of spatial economic activity that allow us to explore the interactions between economic and non-economic variables (i.e., economic and ethnic divisions)?

https://www.spatial-economic-development.com/

Roadmap

1. Conflict: current trends
2. Economic Inequality and Conflict
   - 2.1. Income Inequality and conflict
   - 2.2. Income difference of economic similarity?
3. Ethnicity and Conflict
   - 3.1. Ethnic divisions: Fractionalization, Polarization
   - 3.2. Group size and Conflict
4. Ethnicity, Inequality and Conflict
   - 4.1. Within-group inequality
   - 4.2. Between group inequality
5. New measures of spatial economic activity

1. Conflict

In most of this lecture by social conflict, we refer to:

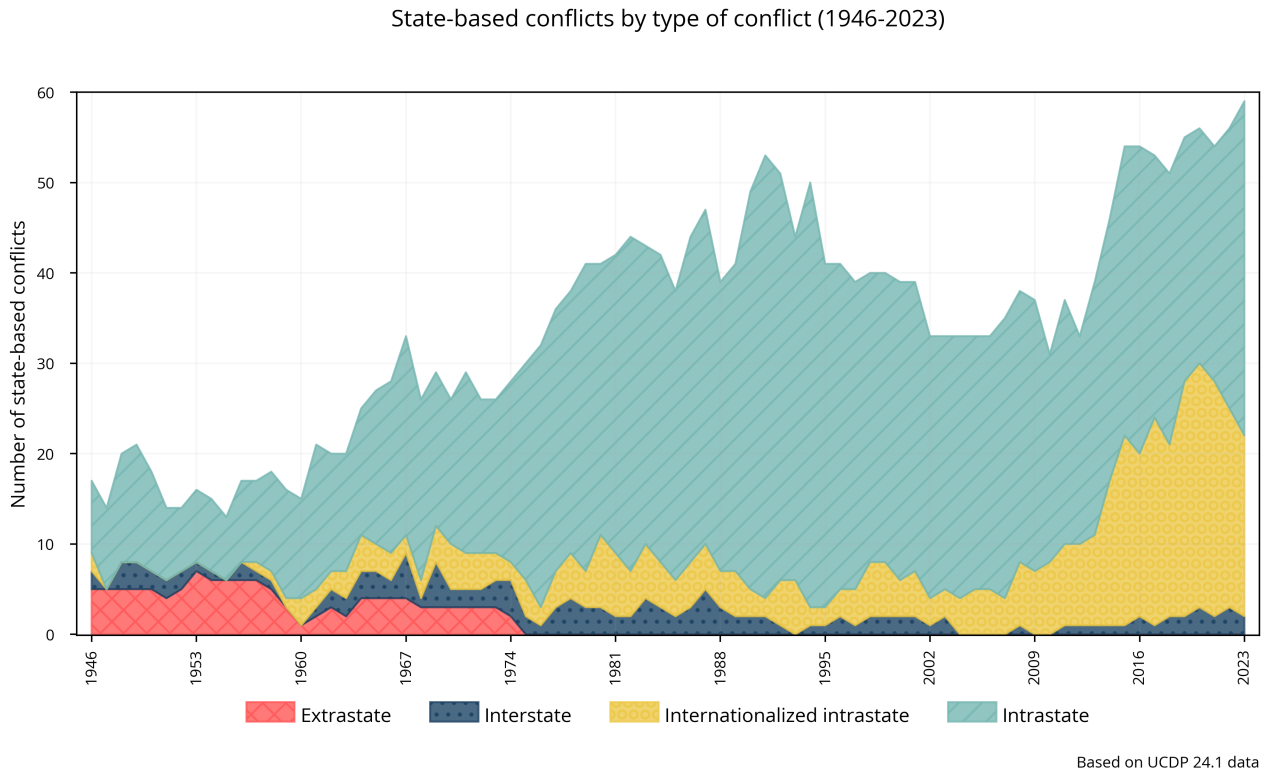
- Within-country unrest
- Many different manifestations: Peaceful demonstrations, strikes, violent riots, armed conflict, civil war
- Key feature: organized conflict

Conflict Trends

- UCDP data (Uppsala Conflict Data Program), https://ucdp.uu.se:
  - State-based conflicts
  - Conflict: a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in one calendar year.
  - Civil war: A state-based conflict or dyad which reaches at least 1000 battle-related deaths in a specific calendar year.

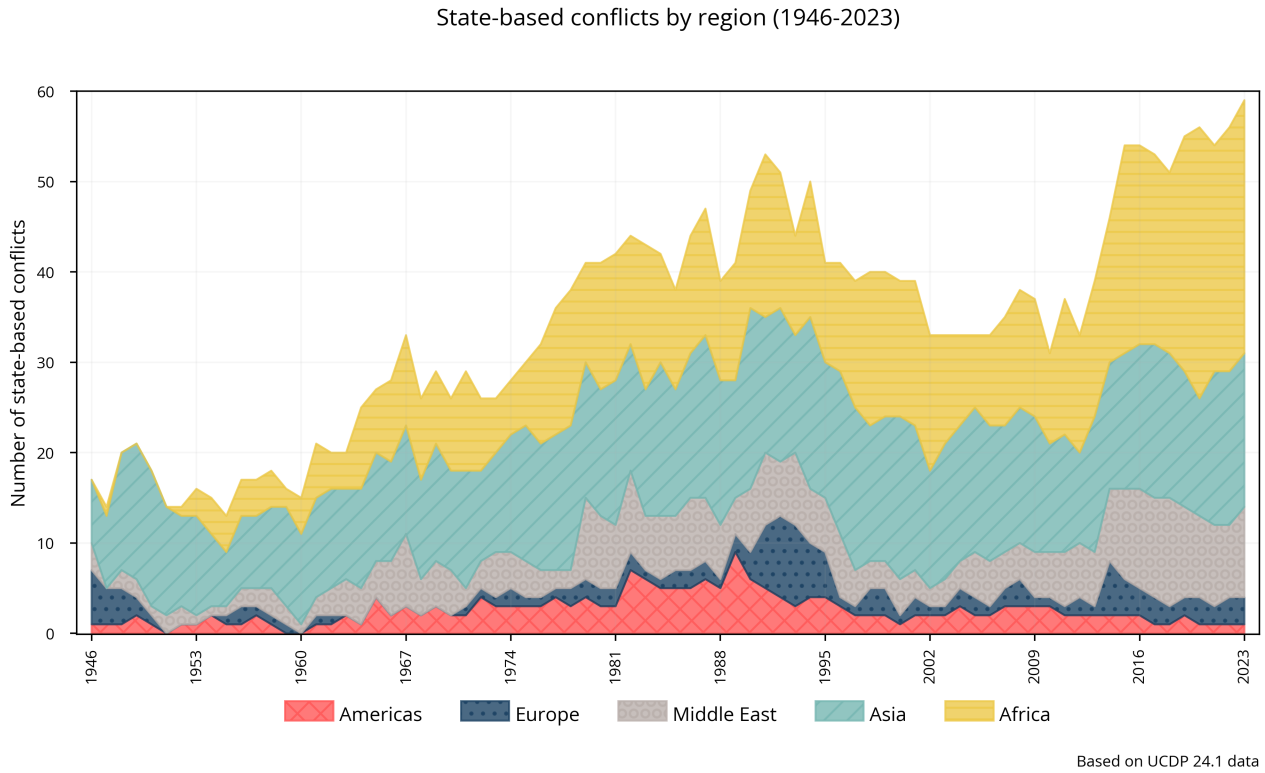
Conflict Trends

1. State-based conflicts are at a historic high



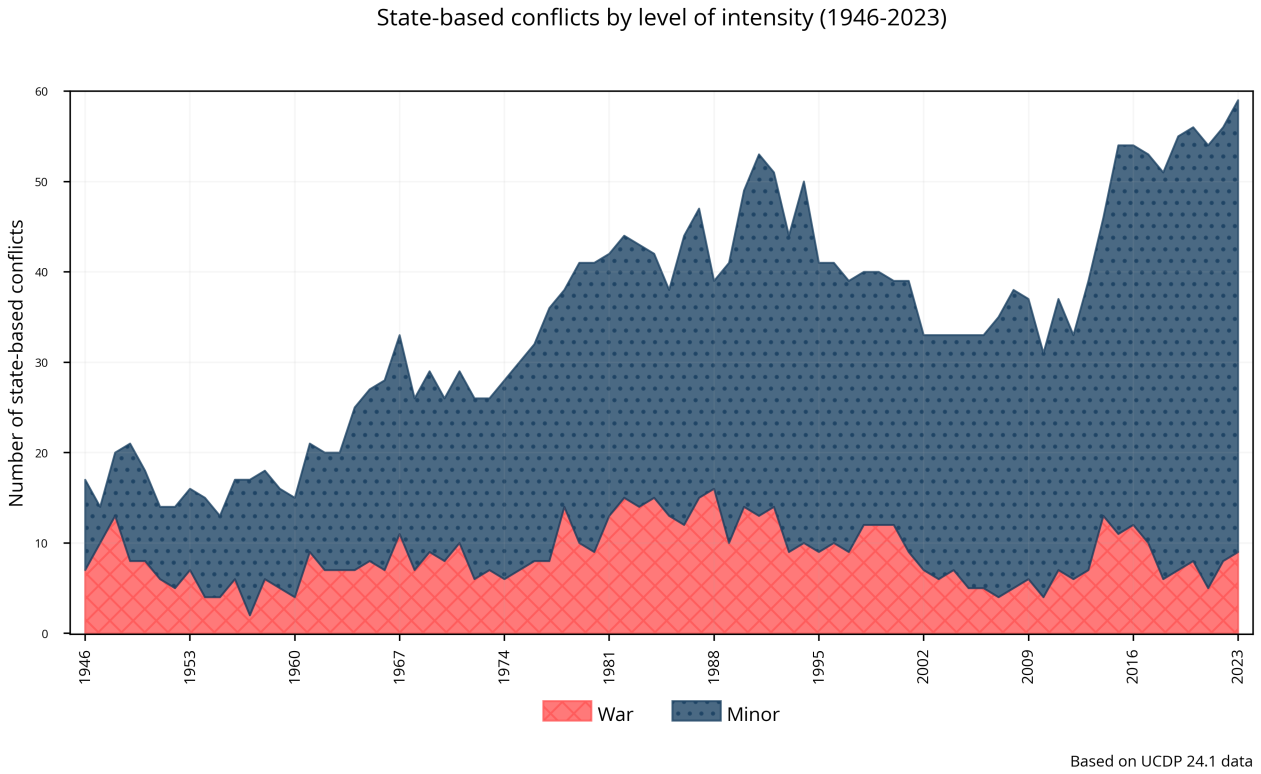
Conflict Trends

2. Most conflicts are concentrated in Africa and Asia, also Middle east



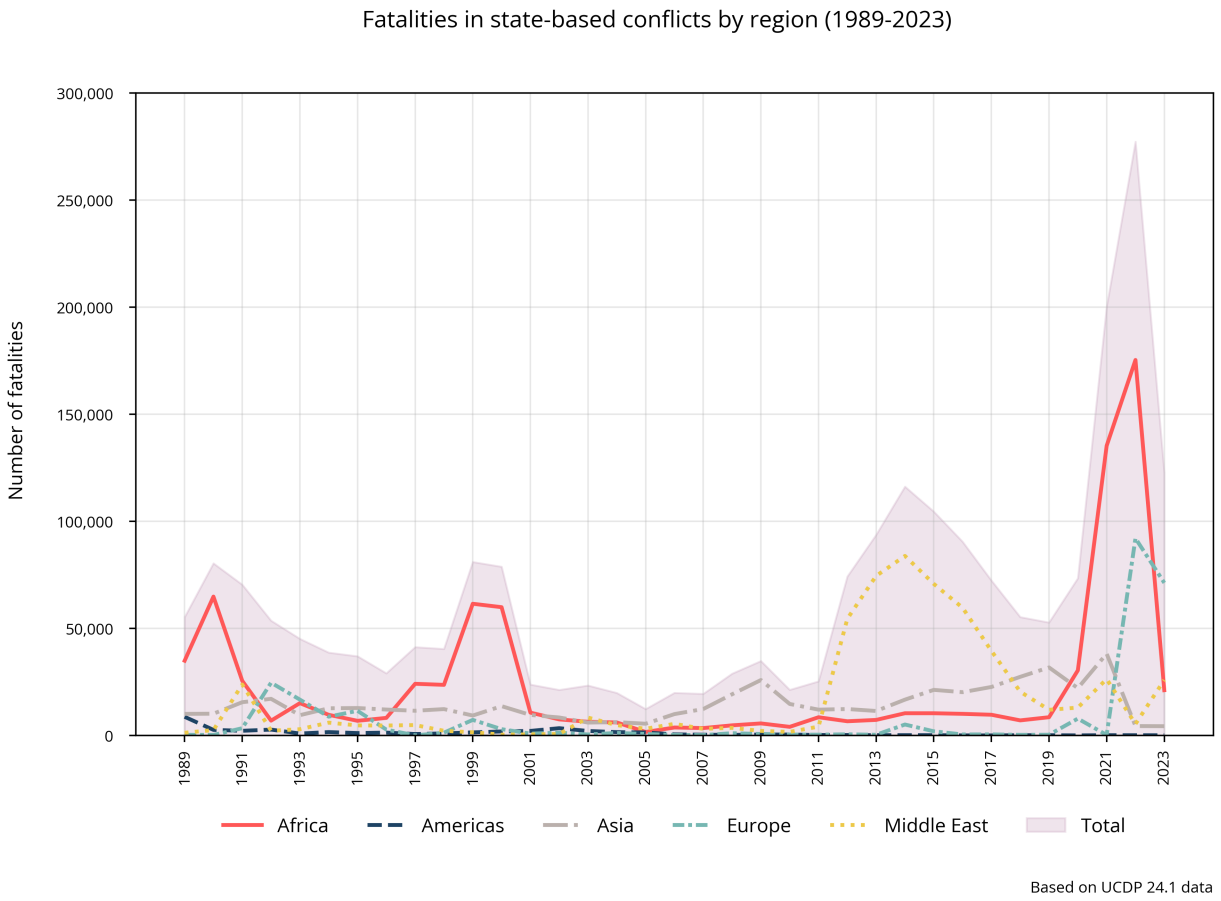
Conflict Trends

3. Most conflicts are not open wars



Conflict Trends

4. The number of fatalities due to conflict is very large



Main trends

- The predominant form of conflict is civil conflict
- There has also been an increase in the number of civil conflicts from the 1970s onwards.
- Battle-related deaths: significant increases. In 2020, there were over 49,000 battle-related deaths globally, while in 2021 there were more than 83,800: a 40% increase.

Why should economists care?

War is "development in reverse" (Collier et al., 2003)

Why should economists care?

War is "development in reverse" (Collier et al., 2003)

1. Loss of human life.

- Direct loss of human life.
  - Narrowly defined battle-related deaths from 1946 to 2019: 11 million fatalities (Lacina and Gleditsch, 2005; updated with current numbers from the UCDP, 2021).
  - Anderton and Brauer (2021) estimate 100 million mass atrocity-related deaths since 1900.
- Indirect effect of wars on human live:
  - mostly due to diseases after the end of conflicts. Indirect fatalities are at least as large as direct casualties (Ghobarah, Huth and Russett, 2003, APSR)

Why should economists care?, II

2. Large economic costs.

- Average drop in GDP of 18 percent, very slow economic recovery (Mueller and Tobias, 2016)
- Also large-scale destruction of infrastructure, human capital (Shemyakina, 2011, JDE) and of (inter-group) social capital (Rohner, Thoenig and Zilibotti, 2013, JOEG; Bauer et al., 2016, JEP)

Why should economists care?, III

3. Conflict tends to recur

- Various war traps
- 68 % of all civil conflict outbreaks after WWII took place in countries experiencing multiple wars.
- War traps hold countries persistently back, both economically and politically.
  - Political institutions
  - Large refugee flows
  - ...
  - Conflict as a major cause of poverty in the world (World Development Report, 2011)

2. Drivers of conflict, II

This lecture focuses on

- how the distribution of certain variables in society might affect the likelihood of conflict.
- In particular,
  - The distribution of income/economic well-being: equal or unequal societies
  - The distribution of non-economic variables such as ethnicity (broadly understood, encompassing religious or linguistic differences)
- Are some distributions more likely to lead to conflict than others?

2.1. Economic Inequality and Conflict

- Is income inequality a correlate of conflict?
- Marxist/traditional view:
    - class is the only/main relevant social cleavage and class conflict the fundamental source of social unrest.

2.1. Economic Inequality and Conflict

- Is income inequality a correlate of conflict?
- Marxist/traditional view:
    - class is the only/main relevant social cleavage and class conflict the fundamental source of social unrest.
- However: Lack of empirical backing

Empirics of the inequality-conflict nexus

- Key dependent variable: armed conflict (Collier and Hoeffler 2004, Fearon and Laitin 2003...)
- Key independent variable: Gini coefficient
- Conclusion: no support for a positive relationship between these variables
  - Lichbach (1989) reviews 43 papers on the subject: overall evidence obtained by all these works is thoroughly mixed.
  - "...fairly typical finding of a weak, barely significant relationship between inequality and political violence [...], rarely is there a robust relationship discovered between the two variables." Midlarsky (1988, p. 491):

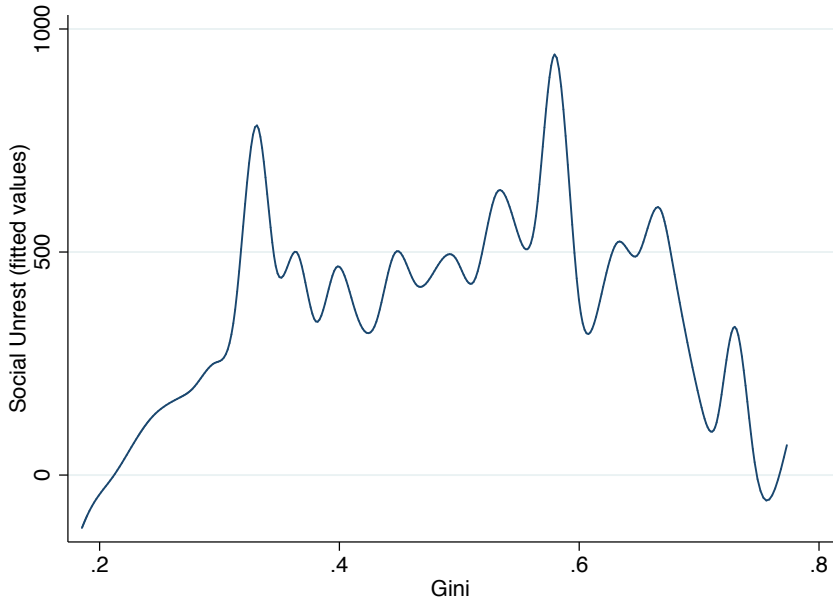
Why we do not find a clear link between Inequality and Conflict?

1. Fundamental problem with the logic of class conflict:
   - the rich have the means but not the motive;
   - the poor have the motive but lack the means.

2. Lack of theoretical model that informs and shapes the empirical test:
   - Gini? income polarization? shape of the relationship? interactions? ...
   - Type of conflict: Recent empirical exercises consider civil war. How about "lower voltage" social unrest? —strikes, demonstrations, ...

Non-linear relationship, lower-voltage conflict

- Financing problem $\rightarrow$ intermediate levels of inequality can be the most dangerous
- Dependent variable: Social unrest (strikes, demostrations, . . .)
- Independent variable: Gini.
- Non-parametric regression



Source: Esteban, Mayoral and Ray (2021)

Economic difference or economic similarity?

- Situations of economic similarity can also be highly conflictual.
- When employment, land, or business resources are scarce, "similar" groups are likely to fight
- Examples: immigration in developed countries; land grabbing in developing countries (Rwandan genocide)
- (Non-class) conflict is the outcome, ethnicity is a convenient marker to categorize individuals on either side of some quasi-artificial divide.
- Instrumentalist view: ethnic markers can become salient in conflicts whose ultimate goal is economic.

Takeaway

Does inequality breed conflict?

- No empirical support
- Conceptual problems: more inequality might make conflict more difficult, not easier
- Economic similarity, and not economic differences can precipitate conflict
- More research, both theoretical and empirical, is needed to clarify all these issues

3. Ethnicity and Conflict

- Non-economic markers: ethnicity (broadly defined to include religious or ethnolinguistic differences).
- Main question:
  - Do Ethnic Divisions Matter?

Ethnic Salience in conflict

- Conflicts are largely ethnic: half of the PRIO conflicts since 1945 are classified as ethnic (Political Instability Task Force, 2012)
- 100 of the 700 known ethnic groups participated in rebellions over the period 1945–1998 (Fearon)
- "...In much of Asia and Africa, it is only modest hyperbole to assert that the Marxian prophecy has had an ethnic fulfillment." (Horowitz)
- "Ethnically divided" societies are more likely to engage in conflict. Do we have evidence to support this view?

Do Ethnic Divisions Matter?

1. Do "ethnic divisions" predict conflict within countries?
2. How do we conceptualize those divisions?
3. If it is true that ethnic divisions matter, how do we interpret such a result?
   - "primordial" –ancestral ethnic hatreds– versus
   - "instrumental" –ethnic markers to achieve political power or economic gain–
4. And if they matter, what are the characteristics of the groups that fight?

Early empirical evidence

(Collier-Hoeffler 2002, Fearon-Laitin 2003, Miguel-Satyanath-Sergenti 2004)

- Typical variables for conflict: demonstrations, processions, strikes, riots, casualties and on to civil war.
- Explanatory variables:
  - Economic per-capita income, inequality, resource holdings . . .
  - Geographic mountains, separation from capital city . . .
  - Political "democracy", prior war . . .
  - And, of course, ethnic. But how is it measured?

How to measure ethnic divisions: the Fractionalization index

First approach to measure "divisions": ethnic diversity

- Fractionalization: index of ethnic diversity

$$F = \sum_{j=1}^{m} n_j(1 - n_j)$$

n_j: is the population share of group j.

- Interpretation: choosing two people at random in a population, F=prob. of they belonging to two different groups.
- It reaches a maximum when everyone belongs to a different group.

- Fractionalization used in many different contexts:
  - growth, governance, public goods provision.

- But it shows no correlation with conflict.
(Collier-Hoeffler (2002), Fearon-Laitin (2003), Miguel-Satyanath-Sergenti (2004))

- Fearon and Laitin (APSR 2003): "... The estimates for the effect of ethnic and religious fractionalization are substantively and statistically insignificant . . . The empirical pattern is thus inconsistent with . . . the common expectation that ethnic diversity is a major and direct cause of civil violence."

An alternative measure: Polarization, Esteban and Ray (1994, 2011)

It aims to capture deep cleavages in society

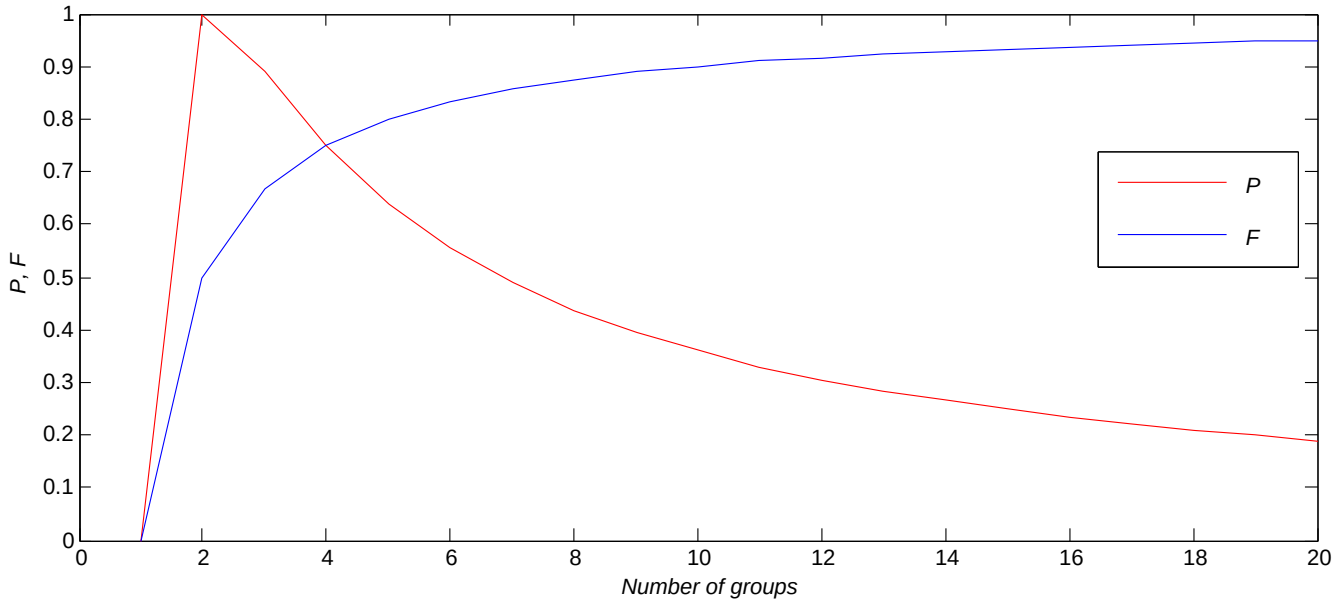
$$P = \sum_{i=1}^{m} \sum_{j=1}^{m} n_i^2 n_j d_{ij}$$

- Society is divided into "groups" and polarization is designed to measure social "antagonism," fueled by two factors
  - Identity: "homogeneity" within each group (proxied by group size).
  - Alienation: "distance" to other groups (proxied by d_{ij}, perceived distances).
  - Particular case: binary distance (belong or not belong)

$$Q = \sum_{i=1}^{m} n_i^2 (1 - n_i)$$

Fractionalization versus Polarization

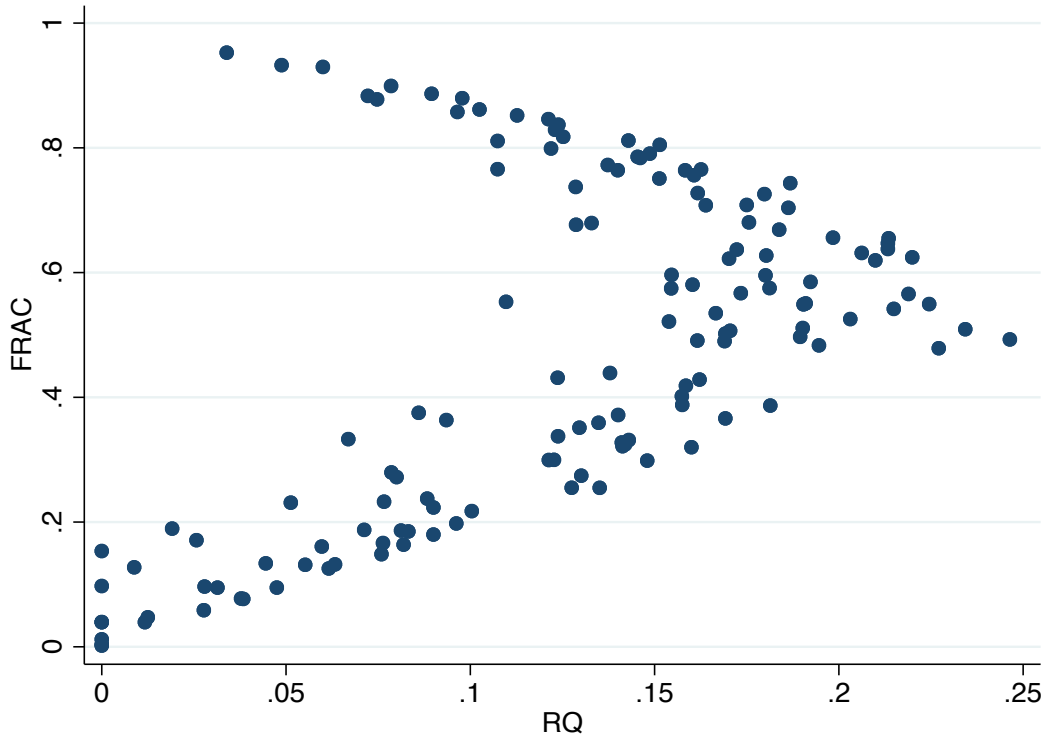
Conceptually very different



Note: equally sized groups and binary distances

Fractionalization versus Polarization

. . . and empirically as well



Note: Polarization with binary distances

Source: Esteban, Mayoral and Ray, (2012).

Data on ethnic groups: Fearon (2003)

Polarization and Conflict

- Montalvo and Reynal-Querol (AER, 2005): positive correlation between ethnic polarization and conflict
- Q index (polarization with "binary" distances: distance is 1 if different groups, 0 if same group)

Ethnicity and Conflict: Theory

- Many potential indices, which one should we use? and why? how can we interpret the results?
- Esteban and Ray (AER 2011): Write down a "natural" theory which links conflict with relevant indices.

Ethnicity and Conflict: Theory

- Many potential indices, which one should we use? and why? how can we interpret the results?
- Esteban and Ray (AER 2011): Write down a "natural" theory which links conflict with relevant indices.
  - m groups engaged in conflict.
  - Two types of prizes at stake: "public" and "private."
    - Public prizes: examples include cultural supremacy, political power, etc.
    - Private prizes: examples include access to natural resources, loot, etc.
  - Conflict equilibrium as the induced Nash equilibrium with extended payoff structure.

Private and Public prizes

Two key differences:

- if prize is private (e.g., money, natural resources. . .)
  - group size dilutes individual benefits and
  - the identity of the winner is irrelevant to the loser (distance between groups is irrelevant
- if prize is public
  - group size doesn't dilute the prizes
  - Identity of the group matters (more or less dislike to cultural norms, etc.)

Equilibrium Conflict Intensity (C)

- Conflict intensity (C) measured by the money value of average per capita resources expended in conflict.
- Approximate formula for large populations:
- Λ: relative "publicness" of the prize

Interpretation

$$C \approx \Lambda P + (1 - \Lambda) F$$

- Both F and P matter for conflict
- F matters on conflicts over private prizes (intergroup distances become irrelevant)
- P connected to publicness of prize (intergroup distances matter)
- No other indexes enter the equilibrium equation (for large populations)

Ethnicity and Conflict: Empirics

- Esteban, Mayoral and Ray (AER 2012, Science 2012)
- Main variables
  - Conflict: Different measures of conflict intensity, from armed conflict and less serious manifestations of conflict (PRIO, Banks)
  - Groups: Fearon database, Ethnologue (linguistic groups)

Distances

Linguistic distances based on language trees

- E.g., all Indo-European languages in a common subtree
- Spanish and Basque diverge at the first branch; Spanish and Catalan share first k nodes. Max: T steps of branching
- Similarity s = common branches/maximal branches down that subtree
- Distance $dij = 1 - s_{ij}^{\delta}$, for some $\delta \in (0, 1]$. Baseline $\delta = 0.05$ as in Desmet et al. (2009)

Additional Variables and Controls

Among the controls:

- Population
- GDP per capita
- Dependence on oil
- Mountainous terrain
- Democracy
- Governance
- Civil rights

Also:

- Indices of publicness and privateness of the prize
- Estimates of group concern from World Values Survey

Empirical evidence

Table 3 Ethnicity and conflict

Variable	[1] PRIO-C	[2] ISC	[3] PRIO-C	[4] ISC
P	***5.16 (0.001)	***19.50 (0.002)	−1.48 (0.606)	−16.33 (0.227)
F	*0.93 (0.070)	*3.56 (0.061)	0.76 (0.196)	0.31 (0.878)
$P\Lambda$			***11.174 (0.003)	***61.89 (0.001)
$F(1-\Lambda)$			*1.19 (0.097)	***10.40 (0.000)
GDPPC	**−0.34 (0.047)	***−2.26 (0.004)	*−0.36 (0.080)	***−3.02 (0.001)
POP	***0.24 (0.000)	***1.14 (0.000)	***0.21 (0.001)	***1.30 (0.000)
NR	−0.27 (0.178)	−0.53 (0.497)	−0.00 (0.570)	0.00 (0.432)
MOUNT	0.00 (0.537)	0.02 (0.186)	0.00 (0.362)	*0.03 (0.061)
NCONT	***1.06 (0.001)	***4.55 (0.001)	**0.77 (0.026)	***4.28 (0.001)
Politics	0.18 (0.498)	0.29 (0.789)	−0.00 (0.328)	**−0.00 (0.026)
LAG	***1.99 (0.000)	***0.46 (0.000)	***1.94 (0.000)	***0.44 (0.000)
CONST	–	0.90 (0.915)	–	9.19 (0.398)
(Pseudo)-R^2	0.35	0.43	0.36	0.44
Observations	1,125	1,111	1,104	1,090
Countries	138	138	138	138

138 countries over 1960–2008, with the time period divided into 5-year intervals. Dependent variables PRIO-C and ISC are indices of conflict described in the text. Variables P and F are measures of polarization and fractionalization described in text. Variable Λ is an index of relative publicness described in text. All specifications employ region and time fixed effects, not shown explicitly. p-values are in parentheses, with *, **, and *** representing associated p-values lower than 0.05, 0.01, and 0.001, respectively. Robust standard errors adjusted for clustering have been employed to computez-statistics. Columns

Interpreting the results, I

Primordial or Instrumental conflict?

Recall:

- Primordial view: ancestral ethnic hatreds lead to conflict
- Instrumental view: instrumental use of ethnicity to achieve political power or economic gain

Our results: indirect evidence that ethnic conflicts are likely to be instrumental, rather than driven by primordial hatreds.

Ethnic divisions: takeaway

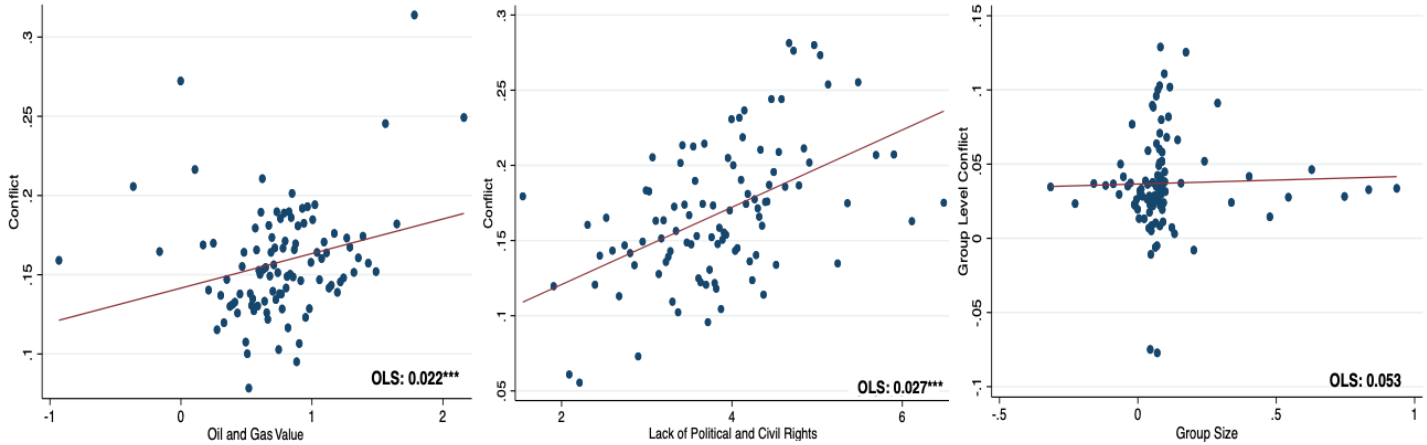
- Ethnic divisions matter
- Ethnically polarized countries more at risk when prize is public
- Ethnically fractionalized countries more at risk when prize is private
- Instrumental use of ethnic divisions

3.2. What are the characteristics of groups that fight?

- Difficult question in general
- We focus on one characteristic: Group size
- Group size: will large or small groups be more likely to be involved in conflict?
- Mayoral and Ray, JDE 2022

Group size and Conflict, II

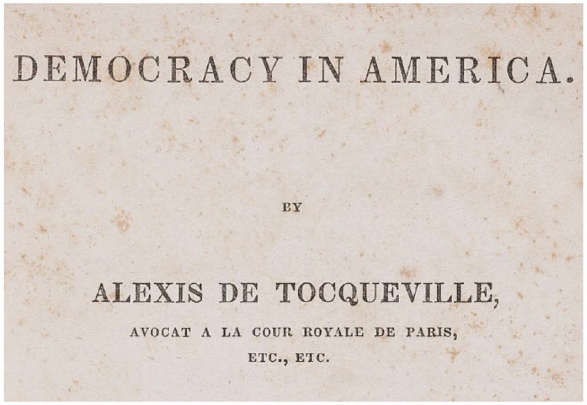
- Conflict (ethnic group level) vs.
  - oil reserves
  - Lack of political&civil rights
  - ethnic group size
- Conflict and (ethnic) group size are uncorrelated

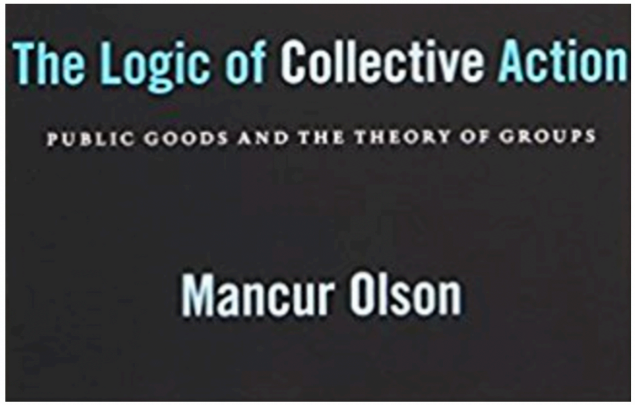


(a) Conflict and Oil Reserves (b) Conflict and Lack of rights (c) Conflict and Group Size

Figure 1. Panel **(a)** shows a binned scatterplot relating civil conflict to the country-level values of oil and gas. Panel **(b)** does the same for civil conflict and an index capturing the lack of political and civil rights. Panel **(c)** shows a binned scatterplot relating civil conflict at the ethnic group level to group size. All graphs control for log GDP per capita, log population and country and year dummies; panel **(b)** also controls for the value of oil and gas. OLS coefficients are noted. *** denotes significance at the 1% level. Details of the construction of these graphs are in Section B.1 of the Online Appendix.

Group size and Conflict, III





- Tyranny of the majority (Tocqueville 1835, Mill 1959) "Society . . . practices a social tyranny more formidable than many kinds of political oppression . . . [imposing] its own ideas and practices as rules of conduct on those who dissent from them . . . " Mill 1859

Tyranny of the minority (Pareto 1927, Olson 1965): "[A] protectionist measure provides large benefits to a small number of people, and causes a very great number of consumers a slight loss. This circumstance makes it easier to put a protection measure into practice." Pareto 1927

Understanding the lack of correlation between group size and conflict

- Two opposite forces:
  - Large groups are stronger$\rightarrow$ higher probability of winning)
  - Small groups might be more motivated (at least in certain types of conflict)

Understanding the lack of correlation between group size and conflict, II

- Consider an additional element, the nature of the prize at stake
  - Conflict over public goods:
    1. large groups are stronger and
    2. as motivated as small groups (as the value of the prize doesn't get diluted with group size)
  - Conflict over private goods:
    1. Large groups are stronger BUT
    2. they are less motivated, as the prize gets diluted as group size increases

Group size: theoretical prediction

- Game theoretical model of conflict initiation.
- Main theoretical prediction:
  - Public prize⇒ Large groups more likely to initiate conflict
  - Private prize⇒ Small groups more likely to initiate conflict

Group size: empirics

Unit of analysis: Ethnic group

$$\text{CONFLICT}_{c,g,t} = \beta_1 \text{SIZE}_{c,g} + \beta_2 \text{PRIV}_{c,g,t} + \beta_3 \text{SIZE}_{c,g} \times \text{PRIV}_{c,g,t} + X'_{c,g,t}\alpha + Y'_{c,t}\delta + Z'\gamma + W'_t\eta + \epsilon_{c,g,t}$$

- Different proxies for PUB: lack of political and civil rights, whether a group is excluded from power, level of autocray . . .
- Different proxies for PRIV: oil reserves, presence of mines, size of homeland. . .

Group size: empirics, II

	Dependent Variable: **Conflict Incidence**								
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
SIZE	-0.002	0.011	0.022	0.004	0.110	0.084**	0.014	0.031	0.012
	(0.915)	(0.656)	(0.412)	(0.893)	(0.118)	(0.019)	(0.587)	(0.305)	(0.656)
SIZE× LACK RIGHTS	0.068*	0.083**	0.083*	0.086*	0.067				
	(0.062)	(0.035)	(0.050)	(0.086)	(0.134)				
SIZE× OIL	-14.455**	-12.836**					-12.554**	-13.350***	-11.696**
	(0.036)	(0.026)					(0.025)	(0.001)	(0.046)
SIZE × OIL_{0-25}			0.033						
			(0.761)						
SIZE × OIL_{25-50}			0.182						
			(0.551)						
SIZE × OIL_{50-75}			-0.166**						
			(0.027)						
SIZE × $OIL_{>75}$			-0.118***						
			(0.005)						
SIZE× PRIVIND						-0.046**			
						(0.016)			
SIZE× PUBINDEX						0.023*			
						(0.080)			
SIZE × MINES				-0.013*					
				(0.099)					
SIZE × HOME					-0.331**				
					(0.034)				
SIZE × AUTOC							0.097**		
							(0.013)		
SIZE × EXCLUDED								0.098**	
								(0.015)	
SIZE × CHILD MORT.									0.004
									(0.137)
OIL	0.887*	0.828*		0.482	0.515		0.841*	0.710**	0.796*
	(0.062)	(0.069)		(0.251)	(0.165)		(0.052)	(0.018)	(0.074)
$OIL_{>75}$			0.007**						
			(0.029)						
OIL_{50-75}			0.005*						
			(0.064)						
OIL_{25-50}			-0.002						
			(0.600)						
OIL_{0-25}			-0.003						
			(0.393)						

Fractionalization, Polarization and the size of the group

Highly consistent results at country and ethnic group level:

- Public prize$\Rightarrow$ Large groups more likely to initiate conflict $\Rightarrow$ Polarized societies are more at risk
- Private prize$\Rightarrow$ Small groups more likely to initiate conflict$\Rightarrow$ Fractionalized societies are more at risk

Takeaway

So far:

- Inequality (≈Gini) not related to conflict (≈ armed civil conflict)
- Ethnic divisions do matter
  - Different indices of ethnic divisions capture different dimensions that are important to understand conflict
  - The nature of the prize at stake is key
  - Large groups fight for public goods, small groups fight for private prizes
- Combination of theoretical and empirical work is key to identify and interpret the results

4. Inequality, Ethnicity and Conflict

Main Question:

- Are there interactions between the ethnic and the income distributions?
- More specifically: Is there any role for
  - Between-group inequality (Horizontal inequality): Differences of group-level average incomes
  - Within-group inequality: heterogeneity of incomes within ethnic groups

Within-Group Inequality and Conflict Intensity

Esteban and Ray (2011), Huber and Mayoral (2019)

- Effective conflict requires two inputs: financial resources and labor (i.e., fighters).
- High-intensity conflict has two opportunity costs:
  - Cost of contributing resources
  - Cost of contributing labor to fight
- Within-group inequality decreases both opportunity costs $\Rightarrow$ facilitates the mobilisation of combatants$\Rightarrow$ Higher conflict intensity.

Within-Group Inequality and Conflict Onset

- Potentially more ambiguous relationship
  - Within-group inequality has the potential of increasing conflict intensity
  - The threat of a highly destructive conflict could also deter conflict onset by encouraging negotiation and compromise by the government.

Horizontal Inequality and Conflict Onset

Wintrobe 1995; Stewart 2002; Cramer 2003, Cederman et al. 2011

- Hypothesis: A positive relationship horizontal inequality and conflict onset (both relatively poor and relatively rich groups are more likely to precipate conflict)
  - Large gaps in average income between groups generate grievances
  - Ethnic markers make it easier to solve the collective action problem

Horizontal Inequality and Conflict Onset, II

- The relationship is, however, more ambiguous:
  - Ceteris paribus, a relatively poor group is less likely to succeed in conflict, then why to start one?
  - A rise in the income of a group might enhance its capacity to fund militants and thus increase its probability of success in conflict.
  - As a result, the closing of the income gap between two groups – rather than its widening – might ignite conflict.
  - Thucydides's Trap (Allison 2017), which states that when a disadvantaged group becomes more powerful and threatens to displace a ruling one, war becomes more likely Mitra and Ray (JPE, 2014)

Predictions in a nutshell

Table 1. The Relation between Ethnic Inequality and Conflict

	Conflict Onset	Conflict Intensity
Horizontal inequality	**Positive** Poor (Rich) groups start conflicts to gain (preserve) resources (Cederman et al., 2011)	Ambiguous
Within-group inequality	Ambiguous	**Positive** Unequal groups have labor and capital necessary to sustain intense conflicts (Esteban and Ray, 2011)

Note. This table summarizes the two most prominent strands of research on group-based inequality and its connections with conflict initiation and intensity.

Empirical Evidence

Source: Huber and Mayoral, JOEG 2019

- Unit of analysis: ethnic group-year
- country FE
- WGI: within-group gini
- BGI: (several measures), log of ratio of group income and average income

Table 2. Within group inequality and the intensity of conflict

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
G^R	5.885**	5.575**	5.812**	6.230**	7.234**	7.756**	5.592**
	(0.019)	(0.027)	(0.014)	(0.042)	(0.028)	(0.030)	(0.016)
HI(LN)		0.103	0.002	-0.005	0.025	-0.010	0.013
		(0.531)	(0.991)	(0.978)	(0.893)	(0.961)	(0.955)
GROUP SIZE			-1.715	-1.842	-2.048	-1.140	-0.658
			(0.415)	(0.367)	(0.366)	(0.656)	(0.778)
GROUP ELEV. (SD)			-0.001	-0.001	-0.001	-0.001	-0.000
			(0.553)	(0.494)	(0.411)	(0.549)	(0.771)
GROUP DIAMONDS			1.055	1.046	1.166	0.925	0.927
			(0.344)	(0.342)	(0.290)	(0.407)	(0.416)
GROUP OIL			0.429	0.425	0.474	0.310	0.303
			(0.463)	(0.465)	(0.434)	(0.634)	(0.639)
GROUP GDP				0.177	0.660	0.900	
				(0.782)	(0.348)	(0.250)	
POP					2.874	-1.653	-1.603
					(0.456)	(0.477)	(0.491)
GDP					-1.215	0.098	0.938
					(0.328)	(0.955)	(0.520)
XPOLITY						-0.123*	-0.123*
						(0.062)	(0.062)
EXCLUDED GROUP						0.892	0.832
						(0.196)	(0.241)
INTENSITY(LAG)	3.938***	3.932***	4.120***	4.124***	4.088***	4.292***	4.312***

Inequality, ethnicity and conflict: Summary

- The distribution economic and non-economic variables, such as income or ethnicity have an impact on conflict
- Understanding the channels require both of careful theoretical and empirical analysis
- Data limitations, but new data on economic characteristics at the ethnic group level are becoming available

Economic development in pixels: New spatially disaggregated measures of consumption and poverty and the limitations of nightlights

18th Winter School on Inequality and Social Welfare

Laura Mayoral

with John Huber, U. Columbia

IAE and Barcelona School of Economics

Alba de Canazei, January 10, 2025

Motivation

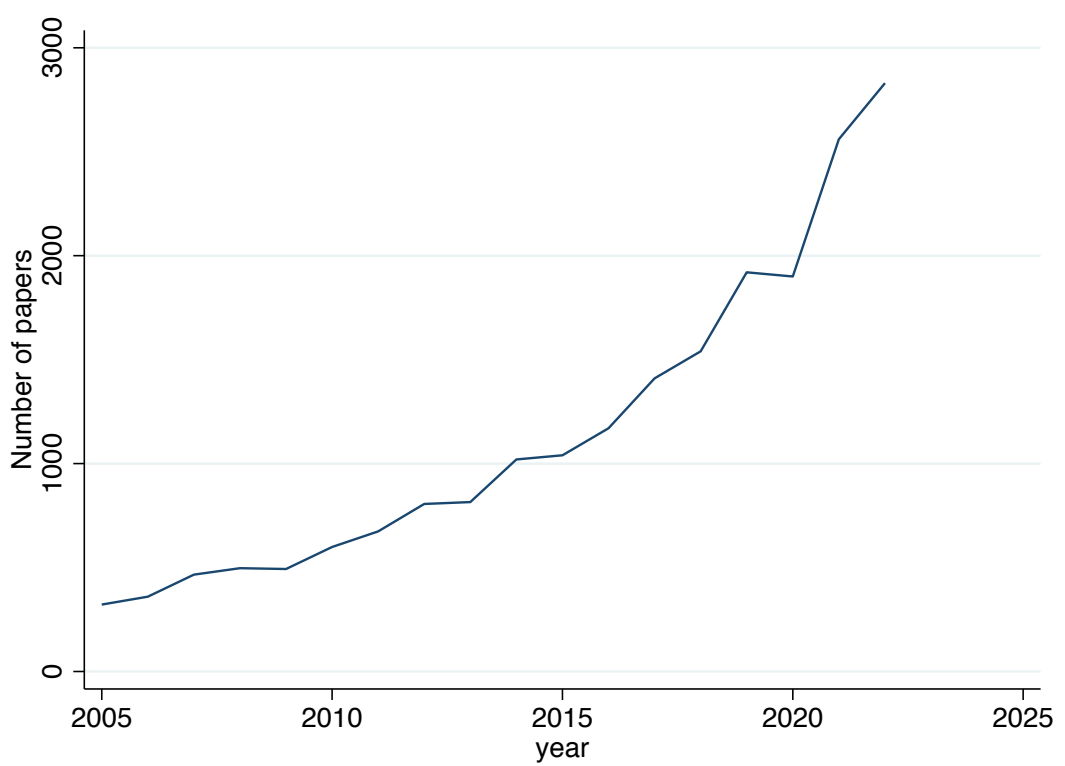
- Measures of economic development are crucial to the study of a wide-range of questions:
  - economic progress, causes and consequences of conflict, policies to alleviate poverty, impact of quality of institutions, etc
- Cross vs. within-country variation$\Rightarrow$ spatially disaggregated data is often needed.
- Problem: not available (even more so in the developing world).
- Popular solution: use nightlights as a proxy (Henderson et al (2011, 2012) and Chen and Nordhaus, 2011)

Nightlights: the "go to" spatially disaggregated measure of economic development

- Why NL?
  - Correlated with development (brighter$\rightarrow$richer)
  - Spatially fine-grained ($\approx$ 1km^2 equator)
  - Whole world since 1992

Number of papers using nightlights is increasing

Number of papers in Google Scholar obtained using the keywords "nighlights+economics" in Google Scholar from 2005 to 2022.



Many examples in Political Economy

- **Do centralized ethnic institutions affect economic development?** – Michalopoulos and Papaioannou (2014)
- **Do good national institutions affect economic development?** – Michalopoulos and Papaioannou (2013)
- Does civil conflict reduce development? – Besley and Reynol-Querol (2014)
- How does China allocate regional aid – Dreher, Fuchs, Hodler, Parks, Raschky and Tierney (2019)
- What is the geographic dispersion of benefits from the adoption of the East African Community?– Eberhard-Ruiz, and Moradi (2019)
- Do cities with railroad hubs have higher development? – Jedwab and Moradi (2016)
- How does mining activity affect local economic development? – Bhattacharyya and Moradi (2019)
- What is the effect of transport networks on development? – Storeygard (2016)

Nightlights: the "go to" spatially disaggregated measure of economic development, II

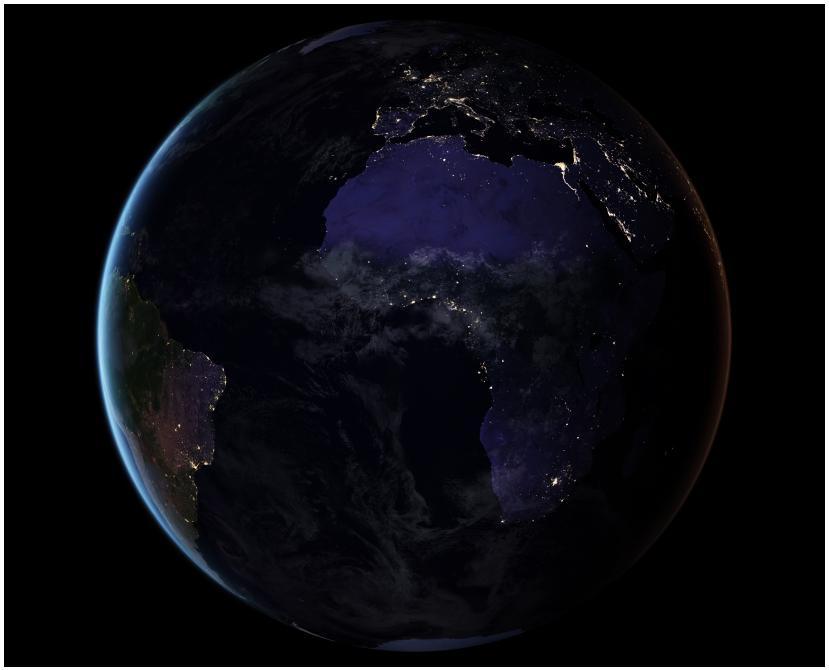
- Limitations

1. Metric (lumens at night) is difficult to interpret
   - At best, ordinality
   - Not suited for: growth, inequality, poverty, etc. . .

2. Measurement error: many well-known sources
   - Time inconsistent (changes of satellite technology, etc.), top-coding, overglow (light is wrongly attributed to areas outside where it is emitted)...
   - Lack of sensitivity to low lights: The problem of darkness

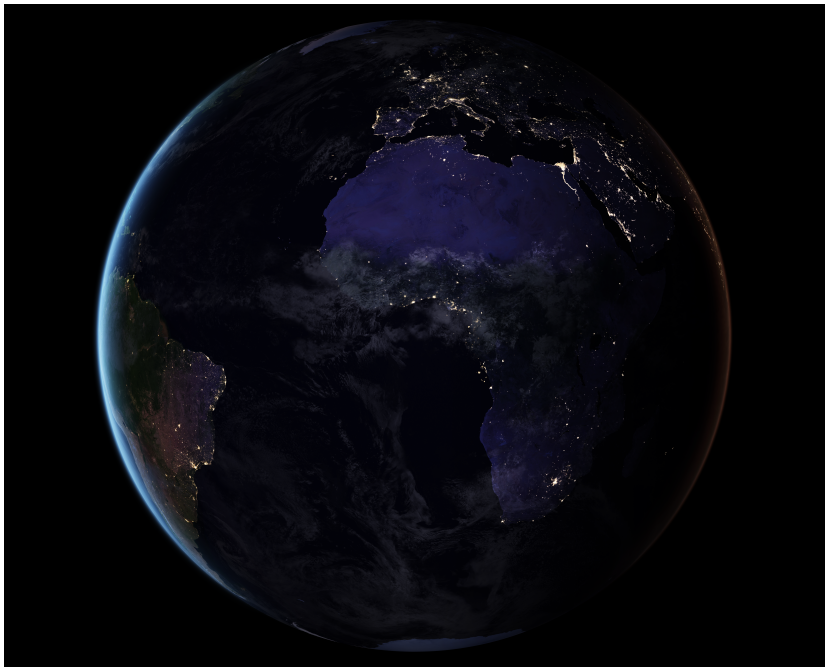
The problem of darkness

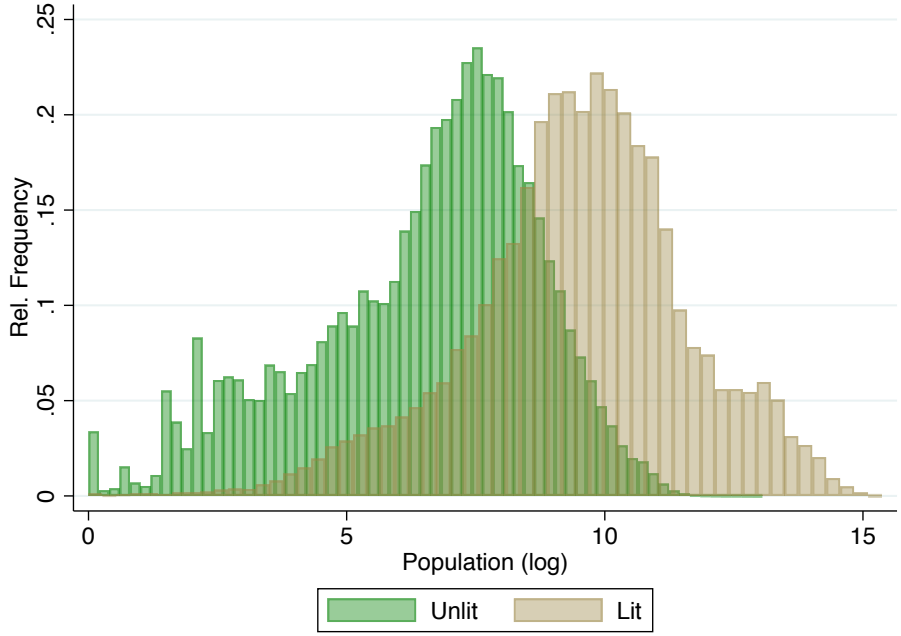
- In Africa: 85-90% of pixels 0 light
- 2018: half of the population resides in areas with 0 light (VIIRS data)



The problem of darkness

- In Africa: 85-90% of pixels 0 light
- 2018: half of the population resides in areas with 0 light (VIIRS data)

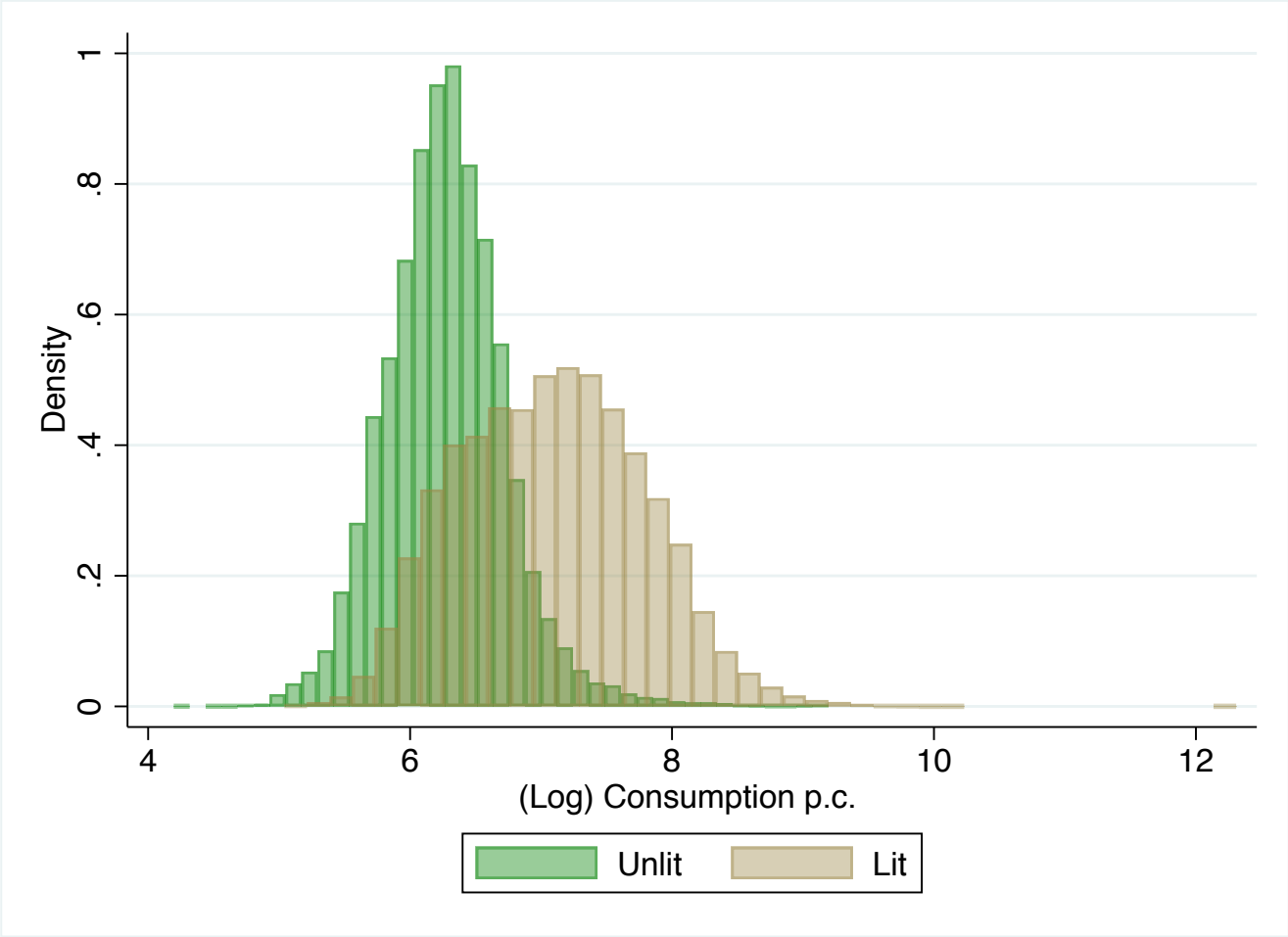




Population in SS Africa at the cell level (10×10 km) in lit vs unlit pixels 2006–2018

Distribution of mean consumption: lit and unlit pixels

Consumption per capita: 34,000+ locations in Africa, survey data (–to be explained–) in lit vs unlit pixels



Distribution of mean consumption and population in lit and unlit pixels

It follows that

- There's correlation between lit/not lit and consumption/pop. BUT
  - The problem of darkness is not data censoring, it's misclassification/measurement error

Distribution of mean consumption and population in lit and unlit pixels

It follows that

- There's correlation between lit/not lit and consumption/pop. BUT
  - The problem of darkness is not data censoring, it's misclassification/measurement error
  - Measurement error is large: NLs are a poor descriptor of economic development

Distribution of mean consumption and population in lit and unlit pixels

It follows that

- There's correlation between lit/not lit and consumption/pop. BUT
  - The problem of darkness is not data censoring, it's misclassification/measurement error
  - Measurement error is large: NLs are a poor descriptor of economic development
  - Measurement error is Non-classical ($\equiv$ systematically correlated with development)

Distribution of mean consumption and population in lit and unlit pixels

It follows that

- There's correlation between lit/not lit and consumption/pop. BUT
  - The problem of darkness is not data censoring, it's misclassification/measurement error
  - Measurement error is large: NLs are a poor descriptor of economic development
  - Measurement error is Non-classical ($\equiv$ systematically correlated with development)
    - Biased coefficients in OLS regressions when NLs is used as independent or dependent variable
    - Attenuation or Amplification bias

Overview: Three main contributions

First contribution

- Spatial Economic Development (SED) dataset: Use machine learning to create a new dataset of spatially disaggregated measures of consumption p.c. and poverty in Africa, cell level, 10×10 Km, over time (2003-2018)

Overview: Three main contributions

First contribution

- Spatial Economic Development (SED) dataset: Use machine learning to create a new dataset of spatially disaggregated measures of consumption p.c. and poverty in Africa, cell level, 10×10 Km, over time (2003-2018)
  - Provides solutions to the main problems in NLs:
    - Easy to interpret: measured in (consumption) dollars
    - Large accuracy improvement

NLs versus SED, Tanzania 2017

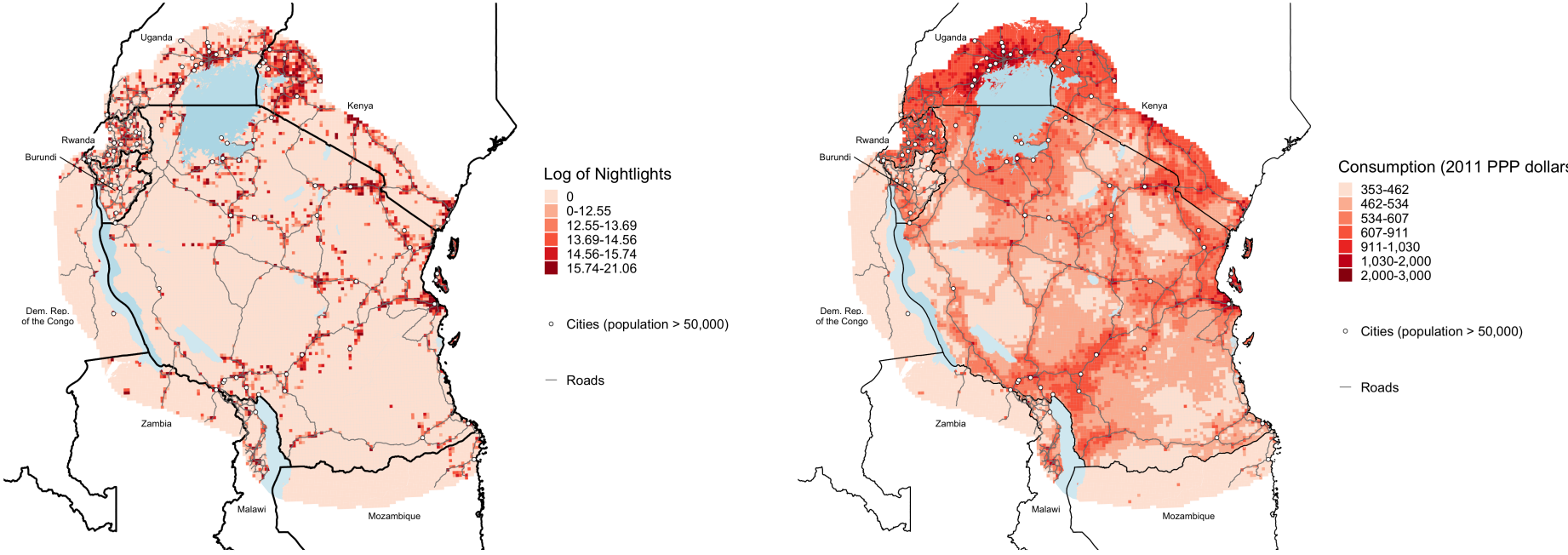


Figure: Nightlights (VIIRS) vs consumption in Tanzania (2017)

Tanzania: size is $\sim$ 950,000 Km^2, population (in 2017) is around 60,000,000

This paper: Three main contributions, II

Second contribution

- Non-classical measurement error in NLs: problems in regression analysis
  - Biased coefficients in models where NL is RHS or LHS variable
  - Amplification or attenuation bias
- Our ML-generated proxy: large accuracy improvement over NLs
- but still contaminated by NCME!
- We propose a simple method to get rid of the non-classical part of the measurement error

This paper: Three main contributions, III

Third contribution

- We revisit two well-known papers on institutions and development:
  - Do centralized ethnic institutions affect economic development? – Michalopoulos and Papaioannou (2014)
  - Do good national institutions affect economic development? – Michalopoulos and Papaioannou (2013)
- These applications illustrate both types of bias (i.e, attenuation and amplification bias)
- When the new data is employed, results are reversed

Summary: Two main takeaways

1. Negative message: the paper warns against the use of NLs in development studies using spatially disaggretated designs
    - Non-classical measurement error in NLs$\rightarrow$severe biases, attenuation or amplification bias
    - The bias can be so large that conclusions of analysis can be reversed

Summary: Two main takeaways

1. Negative message: the paper warns against the use of NLs in development studies using spatially disaggretated designs
   - Non-classical measurement error in NLs→severe biases, attenuation or amplification bias
   - The bias can be so large that conclusions of analysis can be reversed
2. Positive message: much better proxies can be available
   - machine learning & new geolocated data (NL included): more accurate proxies for development
   - Easy to compute/replicate (STATA)
   - Once a good number of features is considered, predictions are very robust to including/excluding a specific feature
   - We show how these new indicators can be used in regression to avoid biases

Outline of paper/talk

Part I:

Creation of SED (Spatial Economic Development dataset)

Part II:

Non-classical measurement error in NLs, Regressions with machine learning predictors

Part III:

Illustrations

Part I

Creation of Spatial Economic Development (SED) dataset

Supervised Machine Learning (SML)

Goal: produce new spatially disaggregated data (cell-level) on economic well-being in Africa

Supervised Machine Learning (SML)

Goal: produce new spatially disaggregated data (cell-level) on economic well-being in Africa

Supervised Machine Learning (SML)

- SML: learns the (potentially highly non-linear) relationship between the input variables ("features") and the output ("training variable").

- Three elements
  - Training variable
  - Features (predictors)
  - Algorithm: Random Forest

The Training Variable

Training variable, ideally:

- Information on consumption/income, spatially disaggregated, and geolocated.
  - (Why? The importance of cardinal interpretability.)
- Problem: no such data exists in Africa (and in much of the developing world).

Available (geolocated) data on economic well-being:

- Demographic and Health Surveys (DHS): individual level, geolocated, large coverage but only assets

Available (geolocated) data on economic well-being:

- Demographic and Health Surveys (DHS): individual level, geolocated, large coverage but only assets
- LSMS (World Bank): consumption and assets, individual level, but small coverage of geolocated surveys

Available (geolocated) data on economic well-being:

- Demographic and Health Surveys (DHS): individual level, geolocated, large coverage but only assets
- LSMS (World Bank): consumption and assets, individual level, but small coverage of geolocated surveys
- PIP-Povcalnet (World Bank): consumption p.c., country-level moments (mean, dispersion, deciles, poverty share...), large availability

Our Solution:

1. Mathematical framework: based on (testable) assumptions allows us to combine different types of datasets;
2. Combine DHS (individual-level) asset data and WB (PIP-povcalnet) country-level consumption data to produce a new training variable of individual-level consumption
3. Use LSMS as a first validation check: test the assumptions of the mathematical framework

Mathematical framework: From asset indices to consumption dollars:

To match the available data, we consider (all variables measured in logs):

- y^*_{ict} is a true measure of economic well-being (log consumption dollars in our implementation)

Mathematical framework: From asset indices to consumption dollars:

To match the available data, we consider (all variables measured in logs):

- y^*_{ict} is a true measure of economic well-being (log consumption dollars in our implementation)

- y^C_{ict} is an index of consumption. In practice, only country-level moments –mean and variance– are observed

Mathematical framework: From asset indices to consumption dollars:

To match the available data, we consider (all variables measured in logs):

- y^*_{ict} is a true measure of economic well-being (log consumption dollars in our implementation)
- y^C_{ict} is an index of consumption. In practice, only country-level moments –mean and variance– are observed
- y^A_{ict} is an asset index measuring economic well-being (WLOG has mean=0 and SD=1), observable at individual-level

i indexes individuals, c indexes countries, and t indexes time.

Assumption A

The variables y_{ict}^C and y_{ict}^A are related to y_{ict}^* as follows:

$$y_{ict}^C = y_{ict}^* + \epsilon_{ict}^C, \tag{1}$$

$$y_{ict}^A = \alpha_{ct} + \beta_{ct} y_{ict}^* + \epsilon_{ict}^A, \quad \beta_{ct} > 0, \tag{2}$$

The errors ϵ_{ict}^C and ϵ_{ict}^A have zero mean, are mutually uncorrelated and are uncorrelated with y_{ict}^*.

Assumption A

The variables y_{ict}^C and y_{ict}^A are related to y_{ict}^* as follows:

$$y_{ict}^C = y_{ict}^* + \epsilon_{ict}^C, \tag{3}$$

$$\underbrace{y_{ict}^A}_{\text{observable}} = \alpha_{ct} + \beta_{ct} y_{ict}^* + \epsilon_{ict}^A, \quad \beta_{ct} > 0, \tag{4}$$

The errors ϵ_{ict}^C and ϵ_{ict}^A have zero mean, are mutually uncorrelated and are uncorrelated with y_{ict}^*.

From asset indices to consumption dollars

Define a new proxy for y^*:

$$\widetilde{y^*_{ict}} = (y^A_{ict} - \alpha_{ct})/\beta_{ct}$$

Using equation (4) it can be written as

$$\widetilde{y^*_{ict}} = y^*_{ict} + \widetilde{\epsilon_{ict}}, \text{ where } \widetilde{\epsilon_{ict}} = \epsilon^A_{ict}/\beta_{ct}. \quad (5)$$

If we can identify β_{ct} and α_{ct}, we can obtain $\widetilde{y^*_{ict}}$, which is

- a consumption proxy, expressed in log dollars
- unbiased ($E_{ct}(\widetilde{y^*_{ict}}) = E_{ct}(y^*_{ct}) = \mu^*_{ct}$)

From asset indices to consumption dollars

Combining equations/omitting algebra/taking expectations....

$$E(y_{ct}^C) = \underbrace{\mu_{y_{ct}^C}}_{\text{observable}} = \frac{-\alpha_{ct}}{\beta_{ct}}, \quad \text{and} \tag{6}$$

$$Var(y_{ct}^C) = \underbrace{\sigma^2_{y_{ct}^C}}_{\text{observable}} = 1/\beta_{ct}^2 + (\sigma^2_{\epsilon_{ct}^C} - \sigma^2_{\epsilon_{ct}^A}/\beta_{ct}^2), \tag{7}$$

If we can eliminate $(\sigma^2_{\epsilon_{ct}^C} - \sigma^2_{\epsilon_{ct}^A}/\beta_{ct}^2)$, we can solve for α_{ct} and β_{ct}

Identifying α_{ct} and β_{ct}

Assumption B:

$$\sigma^2_{\epsilon^C_{ct}} \approx \sigma^2_{\epsilon^A_{ct}} / \beta^2_{ct}$$

Under assumptions A and B, it's possible to identify α_{ct} and β_{ct} using only country-level informaiton on y^C_{ct}:

$$\widetilde{y^*_{ict}} = (y^A_{ict} - \alpha_{ct}) / \beta_{ct} = \mu_{ct} + (y^A_{ict} * \sigma_{ct})$$

where

- μ_{ct} is the country-level mean of log consumption per capita
- σ_{ct} is the country-level standard deviation of y^*_{ict} (and can be measured using country Gini)

Computing the training variable $\widetilde{y^*_{ict}}$

Two steps: 1.) Construction; 2.) Validation

1. Construction: To compute an estimate of $\widetilde{y^*_{ict}}$, denoted as $\widehat{y^*_{ict}}$, we use:
   - Individual level data y^A_{ict} on assets from DHS.
   - country-level data on μ_{ct} and σ_{ct} from PIP-Povcalnet (WB)

Step 1: Construction of the training variable

Steps:

1. Create an asset index using DHS data at the individual level. Details
   - $\sim$ 85 surveys, 1,000,000 households in Africa, 29 countries, 2006–2018
2. Transform as described above using Povcalnet country level moments. Details
3. To match spatial predictors: aggregate at the "enumeration" area or cluster level
   - $\sim$ 35,000 locations in Africa. Details

Step 2: Validation

1. Testable implications of Assumption A:
   - Implication I: the transformed asset and the consumption indices are linearly related;
   - Implication II: their distributions are "similar" (i.e., identical, except for some random noise)

2. Testable implications of Assumption B: If assumption B holds Implication III:

$$\bar{y}^*_{ict} = y^*_{ict} + \epsilon'_{ict}. \tag{8}$$

$$= y^C_{ict} + \epsilon''_{ict}. \tag{9}$$

But if it fails

$$\bar{y}^*_{ict} = \overbrace{(\alpha_{ct}\sigma^C_{ct} + \mu^*_{ct})}^{\neq 0} + \overbrace{\sigma^C_{ct}\beta_{ct}}^{\neq 1} y^*_{ict} + \epsilon^A_{ict}\sigma^C_{ct}. \tag{10}$$

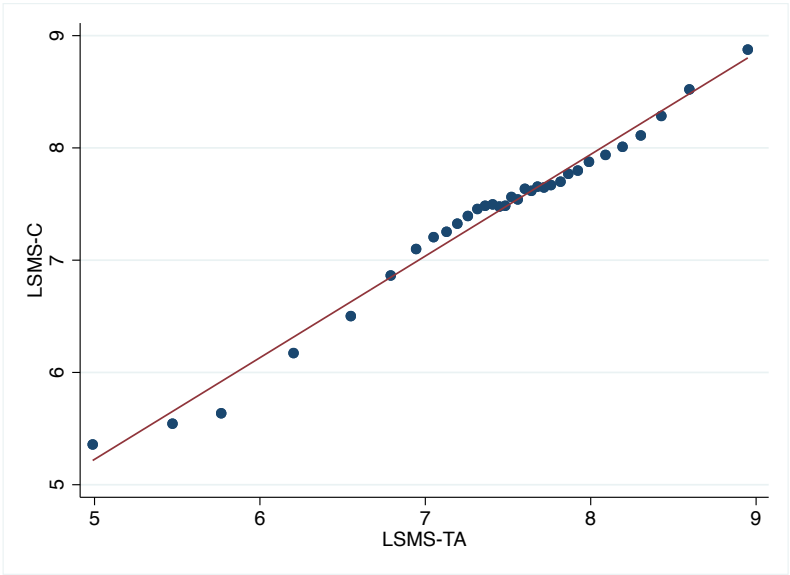
Step 2: Validation using LSMS surveys

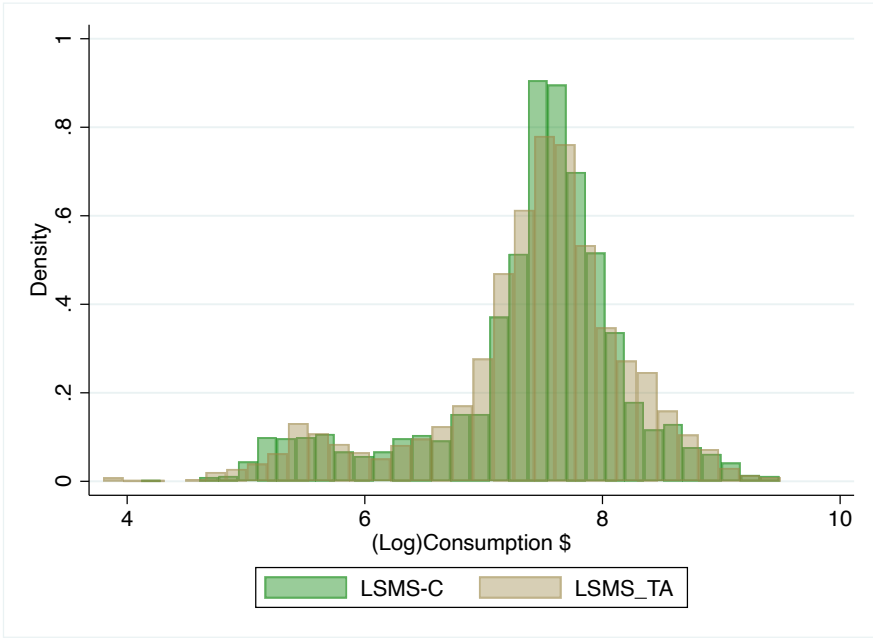
- Compare asset-based values of consumption with direct measures of consumption at individual and cluster level
- Seven country surveys measuring assets and consumption
  - Burkina Faso, Ghana, Malawi, Niger, Nigeria, Tanzania, Uganda
  - 49,062 households
- Enumeration areas to create 'cluster' data (mean values of households in clusters)

Validation using LSMS: Distributions of consumption and transformed asset indeces, cluster level

Implications I and II: similar cluster-level distributions, linearity

LSMS Consumption vs Asset-based indices: Binned scatter plot and histogram





We can't reject: constant$\sim$ 0, slope $\sim$ 1

Validation using WB-PIP data.: DHS vs. WB-PIP, country-level

1. Country-year distributions of the transformed asset index and Consumption data (povcalnet) should be similar

2. Compare:
   - Country-level deciles WB-PIP (World Bank)
   - Country-level poverty lines, WB-PIP (World Bank)

Poverty rates: Transformed DHS index vs PIP-Povcalnet

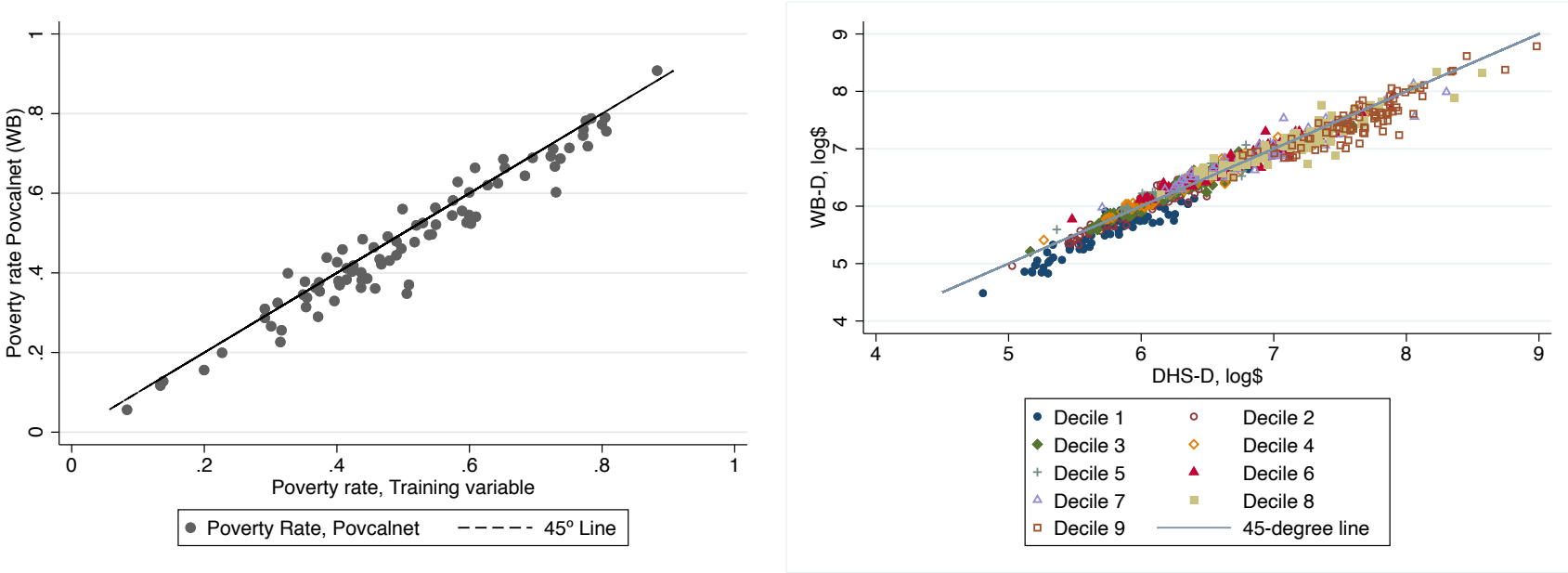


Figure: Country-year Poverty Rates (1.9$ a day threshold) and Deciles. Correlation of poverty lines is 0.97.

3. Constructing prediction models: out-of-sample prediction (of known data points) and evaluation

3.0. Training variable: consumption p.c., $\approx$ 35,000 locations in Africa, based on 1,000,000 households

3.1. Predictors Details

3.2. Prediction Models Details

3.3. Algorithm: Random forest Details

3.4 Hyperparameter Tuning Details

3.5 Evaluation: out-of-sample performance Details

3.6. Variable Importance Details

3.7. Robustness Details

4. SED: Predicting all cells in Africa over time

Two Steps: 1) Prediction; 2) Comparison with existing datasets

Step 1: Prediction

1. Predict all cells in 42 sub-Saharan African countries, 2003-18 (log consumption 2011 $)
2. Calculate poverty rates based on consumption (non-parametric method)
3. $\Rightarrow$ Spatial Economic Development ("SED") data

https://www.spatial-economic-development.com/

4. SED: Predicting all cells in Africa over time, II

Step 2: Compare the resulting data (11,000,000+ datapoints) with other existing regional-level datasets. Details

- HDI and its components (**income per capita**, education index, life expectancy), World Bank's regional poverty rates
- Large correlation

Part II:

Non-classical measurement error in NLs, biases in regression analysis and proposed solutions

5. Non-classical measurement error

y^* is "true" indicator, y is the observable variable, u error

$$y = y^* + u$$

- classical measurement error: y^* and u are uncorrelated (which implies that y and u are correlated)
- non-classical measurement error: y^* are u correlated.
- Very different implications in regression, particularly if y is used as dependent variable

Bias from non-classical measurement error, y is dependent variable

- Goal: To estimate $y^* = X\beta + \epsilon$ (assume X exogeneous, $\beta \geq 0$)
- Problem: We observe $y = y^* + u$
- Resulting model: $y = X\beta + (\epsilon + u)$
  - If measurement error is classical: $\hat{\beta}$ is consistent

Bias from non-classical measurement error, y is dependent variable

- Goal: To estimate $y^* = X\beta + \epsilon$ (assume X exogeneous, $\beta \geq 0$)
- Problem: We observe $y = y^* + u$
- Resulting model: $y = X\beta + (\epsilon + u)$
  - If measurement error is classical: $\hat{\beta}$ is consistent
- If non-classical measurement error: (asymptotic) bias is:

$$\delta = plim(X'X)^{-1}X'u$$

- The sign of δ: given by the sign of the correlation between X and u.
  - Bias can be negative (attenuation) or positive (amplification) .

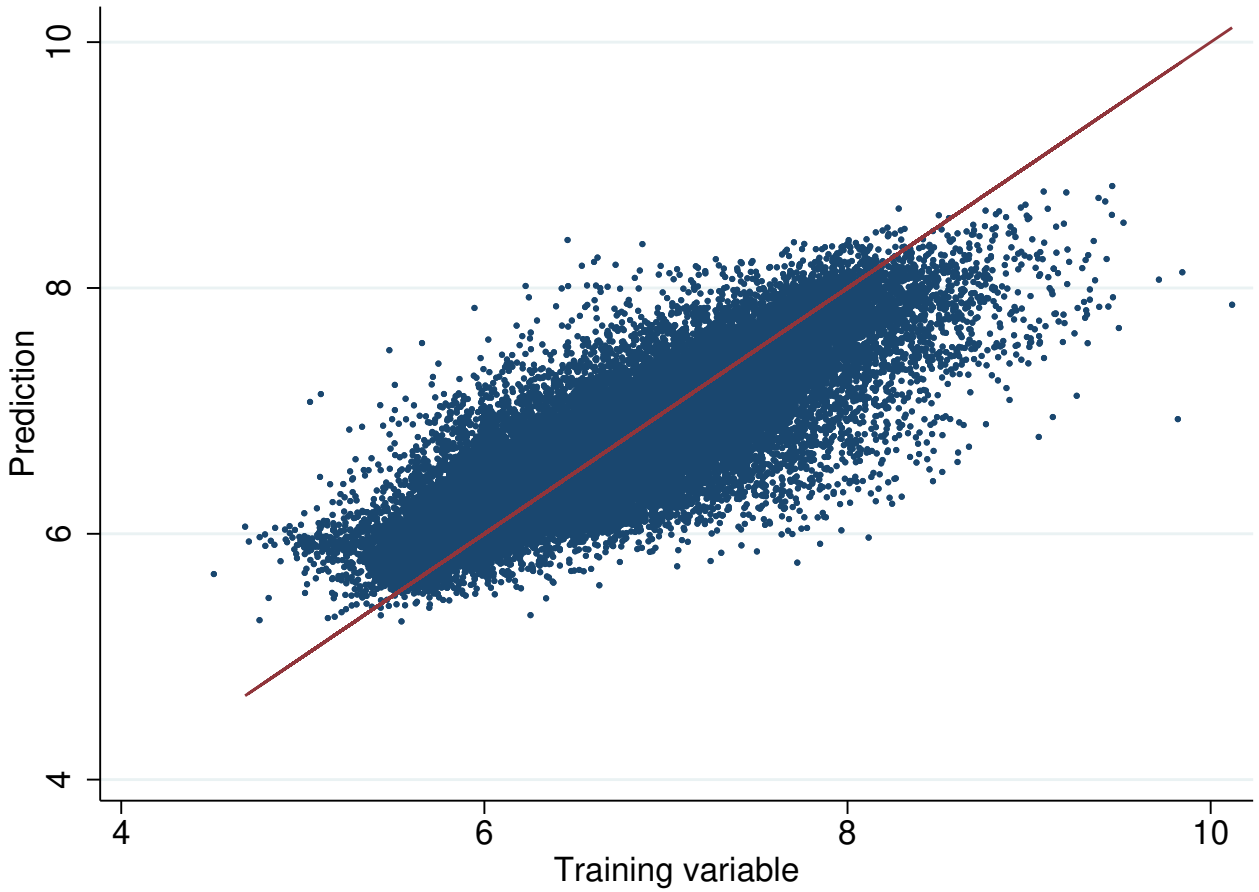
Non-classical measurement error in NLs

- Simple case: NL and y^* both binary (poor/rich)
- When $u = NL - y^*$, u can have only three values:
  - $u = 0$ (no misclassification)
  - $u = 1$ (false positive case, where $y^* = 0$ and $NL = 1$)
  - $u = -1$ (false negative case, where $y^* = 1$ and $NL = 0$)
- Misclassification implies a negative correlation between y^* and u.
- if 'continuous NL' and y^*, same logic applies: if $NL = 0 \Rightarrow u = -y^*$, for more than 85% of the observations!
- If X and y^* are correlated, it's reasonable to expect that X and u will be correlated as well.

In sum

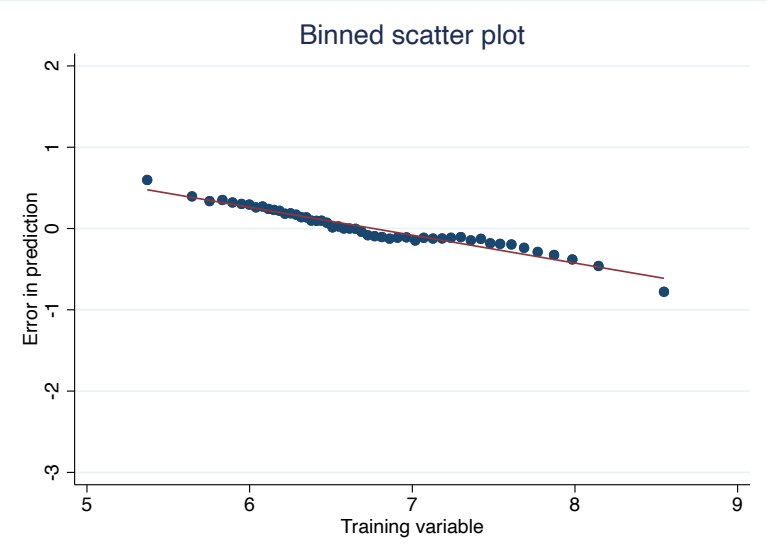
- Nightlights contains non-classical m.e.
- Leads to biases in regression coefficients
  - Both when NLs is the dependent or the independent variable
  - Biases in any direction! (not only "attenuation")
- Solution: Can we use SED instead?

SED also contains non-classical m.e.



SED also contains non-classical m.e.

Negative correlation between training variable and u, $u = \hat{y} - y^*$



- The predicted variable tends to over-predict the poor and under-predict the rich.
- The relationship between the prediction error and the training variable is remarkably linear.

Solution: use linear projections to obtain a new proxy whose prediction error is uncorrelated with y.

- Let $\tilde{y}$ be a proxy (that might contain NCME) of a target variable y. Consider the linear projection of $\tilde{y}$ on y:

$$\tilde{y} = \alpha_0 + \alpha_1 y + \epsilon. \tag{11}$$

- Define $\tilde{y}^T$ as:

$$\tilde{y}^T = \frac{\tilde{y} - \alpha_0}{\alpha_1} = y + \epsilon/\alpha_1. \tag{12}$$

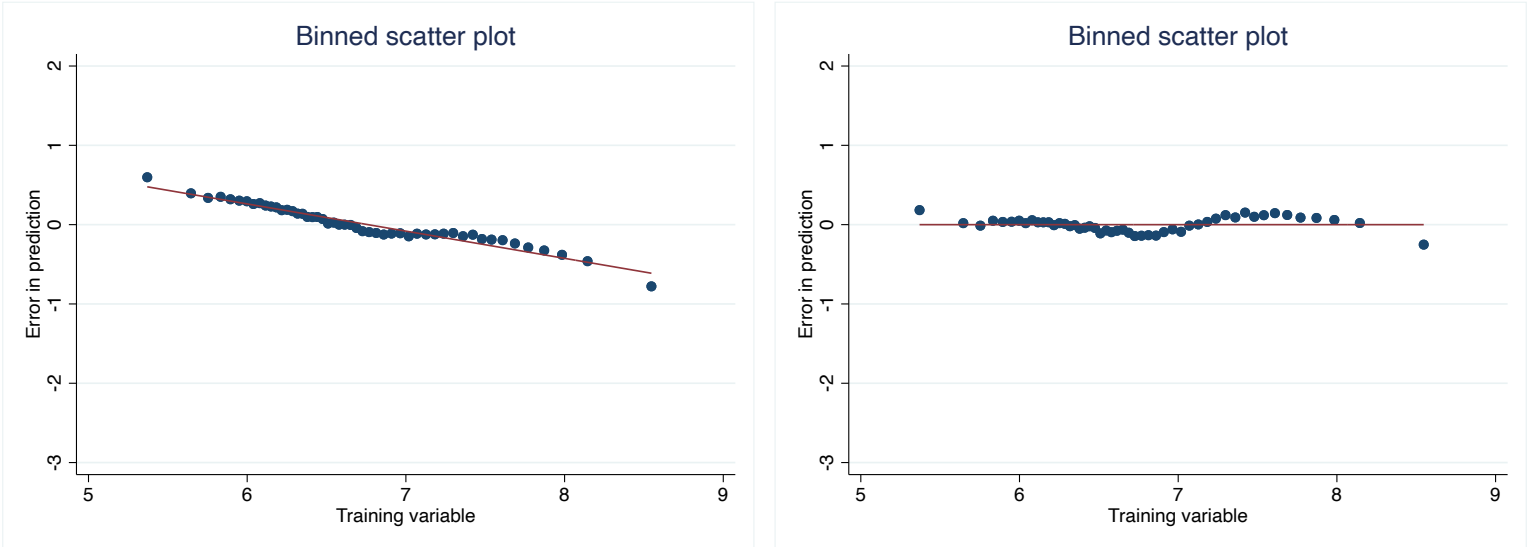
By definition of linear projection, ϵ and y are uncorrelated$\Rightarrow$ $\tilde{y}^T$ contains only classical measurement error.

The proxy $\tilde{y}^T$ only contains classical measurement error

- Computation:
  - Regress predicted cluster consumption on actual consumption using DHS training data to obtain α_0 and α_1
  - Use these coefficients to transform $\tilde{y}$ into $\tilde{y_T}$
- $\tilde{y_T}$
  - has the same correlation with y as does $\tilde{y}$
  - is unbiased
  - contains only classical measurement error (by definition, given α_1 and α_2 are computed to eliminate correlation of y with prediction error)

A proxy of consumption with only classical measurement error

Plot 1: $\hat{u}$ versus y, Plot 2: $\hat{u}_2$ versus y
$\Rightarrow$ $\hat{y}_2$ only contains classical measurement error.



Summary statistics for y, $\tilde{y}$, and $\tilde{y}^T$

	Mean	**Std**	**Min**	**Max**	**MSE**	**Corr with y**	**Corr with u**
y	6.72	.72	4.51	9.84	–	–	–
$\tilde{y}$	6.71	.57	5.31	8.90	0.18	0.814	-0.62
$\tilde{y}^T$	6.72	.89	4.52	10.15	0.27	0.814	0.0025

Table: Summary statistics for y, $\tilde{y}$, and $\tilde{y}^T$. MSE denotes Mean Squared Error.

A trade-off

- $\tilde{y}^T$ has only classical measurement error (zero correlation of error with y)
- $\tilde{y}^T$ has larger measurement error (50%)

Part III:

Illustration: Institutions and development

6. Illustrating non-classical measurement error: Institutions and economic development

Two papers that use NL to study institutions and development

- Michalopoulos and Papaioannou (Econometrica 2013, MP13): Good pre-colonial ethnic institutions (pre-colonial ethnic political centralization) increase economic development

- Michalopoulos and Papaioannou (QJE 2014, MP14): Good national institutions (rule of law and control of corruption) have no effect on economic development

6. Illustrating non-classical measurement error: Institutions and economic development

MP13 and MP14: similar identification strategy (in very broad strokes)

- Key independent variable: (1) Rule of Law/control corruption and (2) pre-colonial ethnic centralization
- Dependent variable: Measure economic development at pixel level using NL
- Compare NL on opposite sides of common border (within country ethnic border in MP13 and national border in MP14)
- Are lights brighter on side with "good institutions"?

Our exercise: Estimate MP models with SED data

- Results are reversed: strong effect of current national institutions and no effect of pre-colonial ethnic institutions!

Our exercise: Estimate MP models with SED data

- Results are reversed: strong effect of current national institutions and no effect of pre-colonial ethnic institutions!
- First case: attenuation bias (coefficients biased towards zero); second case: amplification bias (coefficients biased away from zero).

Our exercise: Estimate MP models with SED data

- Results are reversed: strong effect of current national institutions and no effect of pre-colonial ethnic institutions!
- First case: attenuation bias (coefficients biased towards zero); second case: amplification bias (coefficients biased away from zero).
- Illustrate substantive interpretation that's possible using SED data (i.e, we can interpret effects in dollars) that cannot be done when other proxies are employed.

MP14: National institutions and economic development

NLs and Random Forest models trained only with NLs.

MP14: National institutions and economic development

NLs and Random Forest models trained only with NLs.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Dep. variable is Nightlights (MP14)								
RULE OF LAW	0.0850*	0.0311	0.0759*	0.0370				
	(0.0428)	(0.0170)	(0.0369)	(0.0199)				
CONTROL OF CORRUPTION					0.1121*	0.0479	0.1025*	0.0541
					(0.0523)	(0.0270)	(0.0482)	(0.0296)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.131	.262	.149	.271	.14	.262	.156	.271
Panel B: Dep. variable is log consumption p.c., model RF-1								
RULE OF LAW	0.0810	0.0356	0.0727	0.0376				
	(0.0491)	(0.0222)	(0.0435)	(0.0258)				
CONTROL OF CORRUPTION					0.1119	0.0648	0.1054	0.0675
					(0.0618)	(0.0368)	(0.0589)	(0.0408)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.103	.236	.123	.243	.112	.237	.131	.244
Ethnicity fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Population density and area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location controls.	No	No	Yes	Yes	No	No	Yes	Yes
Geographic controls	No	No	Yes	Yes	No	No	Yes	Yes

MP14: National institutions and economic development (using $\tilde{y}^T$)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel C: Dep. variable is log consumption p.c., model RF-2							
RULE OF LAW	0.5289*	0.2427*	0.3822**	0.1717*				
	(0.2106)	(0.1077)	(0.1340)	(0.0758)				
CONTROL OF CORRUPTION					0.7152***	0.3461*	0.5348***	0.2910**
					(0.1884)	(0.1357)	(0.1250)	(0.1005)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.218	.774	.434	.798	.307	.776	.482	.803
	Panel D: Dep. variable is log consumption p.c., model RF-3							
RULE OF LAW	0.6646**	0.4426*	0.5142**	0.3713**				
	(0.2553)	(0.1738)	(0.1810)	(0.1313)				
CONTROL OF CORRUPTION					0.9019***	0.6580**	0.7259***	0.6013***
					(0.2299)	(0.2201)	(0.1685)	(0.1690)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
R-squared	.258	.753	.431	.773	.369	.763	.5	.786
Ethnicity fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Population density and area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location controls.	No	No	Yes	Yes	No	No	Yes	Yes
Geographic controls	No	No	Yes	Yes	No	No	Yes	Yes

MP14: National institutions and economic development (using $\tilde{y}$)

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel B: Dep. variable is log consumption p.c., model RF-1				
RULE OF LAW	0.0417	0.0184	0.0375	0.0194				
	(0.0253)	(0.0114)	(0.0224)	(0.0133)				
CONTROL OF CORRUPTION					0.0577	0.0334	0.0543	0.0348
					(0.0318)	(0.0189)	(0.0303)	(0.0210)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.103	.236	.123	.243	.112	.237	.131	.244
				Panel C: Dep. variable is log consumption p.c., model RF-2				
RULE OF LAW	0.3389*	0.1555*	0.2449**	0.1100*				
	(0.1349)	(0.0690)	(0.0858)	(0.0486)				
CONTROL OF CORRUPTION					0.4583***	0.2218*	0.3427***	0.1864**
					(0.1207)	(0.0869)	(0.0801)	(0.0644)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.218	.774	.434	.798	.307	.776	.482	.803
				Panel D: Dep. variable is log consumption p.c., model RF-3				
RULE OF LAW	0.4236**	0.2821*	0.3277**	0.2367**				
	(0.1628)	(0.1108)	(0.1154)	(0.0837)				
CONTROL OF CORRUPTION					0.5749***	0.4194**	0.4627***	0.3833***
					(0.1466)	(0.1403)	(0.1074)	(0.1078)
Obs	40871	40871	39250	39250	40871	40871	39250	39250
Adj. R-squared	.258	.753	.431	.773	.369	.763	.5	.786
Ethnicity fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Population density and area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location controls.	No	No	Yes	Yes	No	No	Yes	Yes
Geographic controls	No	No	Yes	Yes	No	No	Yes	Yes

* indicates p$<$.05, **indicates p$<$.01, and *** indicates p$<$.001.

MP13: Ethnic institutions and economic development (using $\tilde{y}^T$)

	(1)	(2)	(3)	(4)	(5)
	Panel A: Dep. variable is lit/unlit (MP13)				
JURISDICTIONAL HIERARCHY	0.0301	0.0349*	0.0238***	0.0256***	0.0173***
	(0.0203)	(0.0178)	(0.0088)	(0.0088)	(0.0060)
Obs	61359	61359	61359	61015	61015
Adj. R-squared	0.008	0.182	0.268	0.287	0.293
	Panel B: Dep. variable is consumption p.c., model RF-1				
JURISDICTIONAL HIERARCHY	0.0330	0.0375	0.0255**	0.0306**	0.0215**
	(0.0246)	(0.0235)	(0.0126)	(0.0132)	(0.0085)
R-squared	0.007	0.143	0.216	0.256	0.265
N	61359	61359	61359	61015	61015
	Panel C: Dep. variable is consumption p.c., model RF-2				
JURISDICTIONAL HIERARCHY	-0.0200	-0.0142	-0.0257	-0.0213	-0.0186
	(0.0672)	(0.0245)	(0.0255)	(0.0208)	(0.0205)
R-squared	0.001	0.762	0.779	0.820	0.824
N	61359	61359	61359	61015	61015
	Panel D: Dep. variable is consumption p.c., model RF-3				
JURISDICTIONAL HIERARCHY	-0.0239	-0.0120	-0.0240	-0.0191	-0.0187
	(0.0683)	(0.0219)	(0.0245)	(0.0215)	(0.0210)
	-0.35	-0.55	-0.98	-0.88	-0.89
R-squared	0.001	0.775	0.794	0.822	0.826
N	61359	61359	61359	61015	61015
Country Fixed effects	No	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes	Yes
Controls at the Pixel level	No	No	No	Yes	Yes
Controls at the Ethnic-Country level	No	No	No	No	Yes

Ethnic institutions and economic development (using $\tilde{y}$)

	(1)	(2)	(3)	(4)	(5)
	Panel B: Dep. variable is consumption p.c., model RF-1				
JURISDICTIONAL HIERARCHY	0.0170	0.0193	0.0132**	0.0158**	0.0111**
	(0.0127)	(0.0121)	(0.0065)	(0.0068)	(0.0044)
Obs	61359	61359	61359	61015	61015
Adj. R-squared	0.007	0.143	0.216	0.256	0.265
	Panel C: Dep. variable is consumption p.c., model RF-2				
JURISDICTIONAL HIERARCHY	-0.0128	-0.0091	-0.0165	-0.0137	-0.0119
	(0.0430)	(0.0157)	(0.0163)	(0.0134)	(0.0132)
Obs.	61359	61359	61359	61015	61015
Adj. R-squared	0.001	0.762	0.779	0.820	0.824
	Panel D: Dep. variable is consumption p.c., model RF-3				
JURISDICTIONAL HIERARCHY	-0.0152	-0.0077	-0.0153	-0.0121	-0.0119
	(0.0435)	(0.0140)	(0.0156)	(0.0137)	(0.0134)
Obs.	61359	61359	61359	61015	61015
Adj. R-squared	0.001	0.775	0.794	0.822	0.826
Country Fixed effects	No	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes	Yes
Controls at the Pixel level	No	No	No	Yes	Yes
Controls at the Ethnic-Country level	No	No	No	No	Yes

Interpretablility of results

National institutions:

- A one-unit increase in RULE OF LAW $\rightarrow$ 45% increase in consumption per capita
- Going from the lowest to highest value of RULE OF LAW $\rightarrow$ 183% increase in consumption

Jurisdictional hierarchy: (ignore null effect)

- Using NL-only RF model: A one-unit increase in JH $\rightarrow$ 2.6% increase in consumption
- Using NL-only RF model: Worst to best JH $\rightarrow$ 11% increase in consumption

Attenuation bias in national institutions paper

Suppose rule of law (X) causes development (y^*)

Then u and X must be negatively related (attenuation bias)

- u is negatively related to y^*
- y^* is positively related to rule of law
- So u is negatively related rule of law

The role of dark pixels in the attenuation bias

Roughly 90% of pixels are dark

The true relationship between any X and y^* is strongly affected by this relationship within the set of dark pixels.

Estimate MP14 model (col. 4) using only dark pixels:

- RULE OF LAW coefficient: 0.333 (p=0.011) (vs, 0.37 using all data)

NL studies implicitly assume that the relationship between X and y^* is the same within dark pixels as it is across lit and dark pixels....

Amplification bias in the ethnic institutions study

Suppose there is no relationship between JURISDICTIONAL HIERARCHY and development.

X and u will be correlated if there is a third variable that is correlated with both u and X.

Population density (urbanization) is such a third variable

Population density and amplification bias in the ethnic institutions study

There is a (well-known) positive relationship between population density and NL, leading to a positive relationship between population density and u

- False negative ($u = -1$) in low density areas
- False positive ($u = 1$) in high density areas

There should be (and is) a positive relationship between JURISDICTIONAL HIERARCHY and population density

- Pre-colonial ethnic political centralization emerged from need for social organization in most populated communities (e.g., Turchin et al, 2022)

Thus, there should be positive relationship between JURISDICTIONAL HIERARCHY and u through population density, leading to amplification bias

Conclusion

- This paper highlights problems with existing proxies for economic well-being
- Proposes a way of dealing with them
  - Use existing data to predict economic well-being
  - Interpretable measures, more accurate
  - take measurement error into consideration if used in regression
- Next steps: Expand data set to all developing countries
- New possibilities for substantive research
  - inequality at different levels of aggregation (regional, ethnic...), growth
  - What types of areas grow the most (rich or poor, remoteness, ecological features, border areas, ethnic groups)
    - What types of areas benefit most from good institutions
    - Economic causes/consequences of civil conflict
  - Ethnic control of government and economic development
  - Targeting aid programs

Thank you!

Nightlights Data

Two main sources:

- DMSP (Defense Meteorological Satellite Program): 1992 to 2013;
  - Designed to detect clouds to assist with short-term weather forecasts for the Air Force.
  - worse quality data: blurring, coarse resolution, no calibration, low dynamic range, top-coding, and unrecorded variation in sensor amplification that impairs comparability over time and space

- VIIRS (Visible Infrared Imaging Radiometer Suite): 2013–;
  - Designed to measure the radiance of light coming from earth, in a wide range of lighting conditions
  - higher spatial accuracy and with temporally comparable data

Back to Original Page

Creation of Asset index

1) For each survey, estimate principal components model to generate y_{ict}^{pca}
   - Source of drinking water, type of toilet facility, flooring, wall, roof materials, presence of electricity, number of sleeping rooms, radio, television, refrigerator, motorcycle or scooter, car or truck, telephone, mobile phone.
   - Assets can vary across surveys
   - Since pca estimated separately for each survey, weights assigned to assets can vary across surveys

2) Take log of y_{ict}^{pca} and standardize it to obtain y_{ict}^{A} (mean=0 and SD=1)
   - Non-comparability across surveys at this stage
   - Working at individual level

3) Transform it using: $\widehat{y_{ict}^{*}} = \mu_{ct} + (y_{ict}^{A} * \sigma_{ct})$

Back

Creating a training variable: from survey respondents to clusters

DHS enumeration area: "cluster"

- 34,484 clusters from 29 countries, 2006-18
- Avg. respondents per cluster = 26.2 (min=16)
- Cluster are geocoded with centroid jiggered by max of 5km
- Each DHS cluster assigned to 10x10 kilometer cell with centroid=DHS cluster centroid
- Take mean of $\widehat{y^*_{ict}}$ for respondents in each cluster, yielding a measure of mean consumption per capita in each geocoded cluster

Back to Original Page

Is it important to transform first and then predict?

Short answer: YES!

- WB-PIP consumption data is sparse, i.e., it's not feasible to predict the asset variable, then predict for all country-years and then transform.
- By transforming first, and then predict, we can compare the predictions with PIP data for country-years that are not in the training sample.

Back to Original Page

3.1 Prediction: Random forest algorithm

- Supervised machine learning using ensemble approach
  - Individual trees built on bootstrap samples (about one-third of observations randomly left out)
  - Each tree built on different bootstrap sample
  - Each tree uses a fraction of predictors (this fraction is determined by researcher)
  - Specific variables employed in each tree are randomly chosen
  - Predictions from individual decision trees are combined
- Advantages
  - Avoids overfitting to the training set inherent to standard decision tree algorithms
  - Accurate performance with large number of predictors
  - **Low complexity and low computational cost**

Back to Original Page

3.2 Prediction: Predictors

Time-varying predictors are in **bold.**

- **Nightlights.**
- *Other proxies of economic activity*: **Population, CO2 production**
- *Geography.* ecosystem type, ruggedness of terrain, elevation, latitude and longitude; caloric yield of land
- *Distances* to the capital; a highway; the coast; a harbor; a river; and catholic and/or protestant missions
- *Climatic variables/disease environment*: **temperature, rainfall and malaria incidence**

Back to Original Page

3.3. Prediction: Prediction models

RF-1	NL only
RF-2	NL + All other variables
RF-3	All other variables (except NL)
Model 4	KNN with NL only
Model 5	OLS with NL only

Table: Random forest models and other prediction models (for comparison)

Back to Original Page

3.4. Prediction: Parameter tuning

	Preferred Hyperparameters Values				
	NTREES	NVARS	DEPTH	VAR	
MODEL 1	180	1	25	.0001	7
MODEL 2	180	8	35	.0001	3
MODEL 3	180	6	35	.0005	1

Table: PREFERRED HYPERPARAMETER VALUES EMPLOYED IN MODELS 1–3.

NTREES: the number of individual trees
NVARS: the maximum number of predictors included in each tree
DEPTH: maximum tree depth
VAR: the minimum proportion of the variance at a node in order for splitting to be performed

Back to Original Page

3.5. Prediction: Evaluation

- Evaluation is on predictions of held-out locations
- Steps:
  - Drop survey x
  - Estimate model on 84 other surveys
  - Predict survey x
  - Repeat for all surveys
- Measures: Mean square error (MSE) computed from the out-of-sample predictions; R^2 computed as square of the within-survey correlation between the training variable and the (out-of-sample) predictions
- Prediction performance is very good and highly competitive→outperforms existing models (i.e., Yeh et al. (2020), Nature)

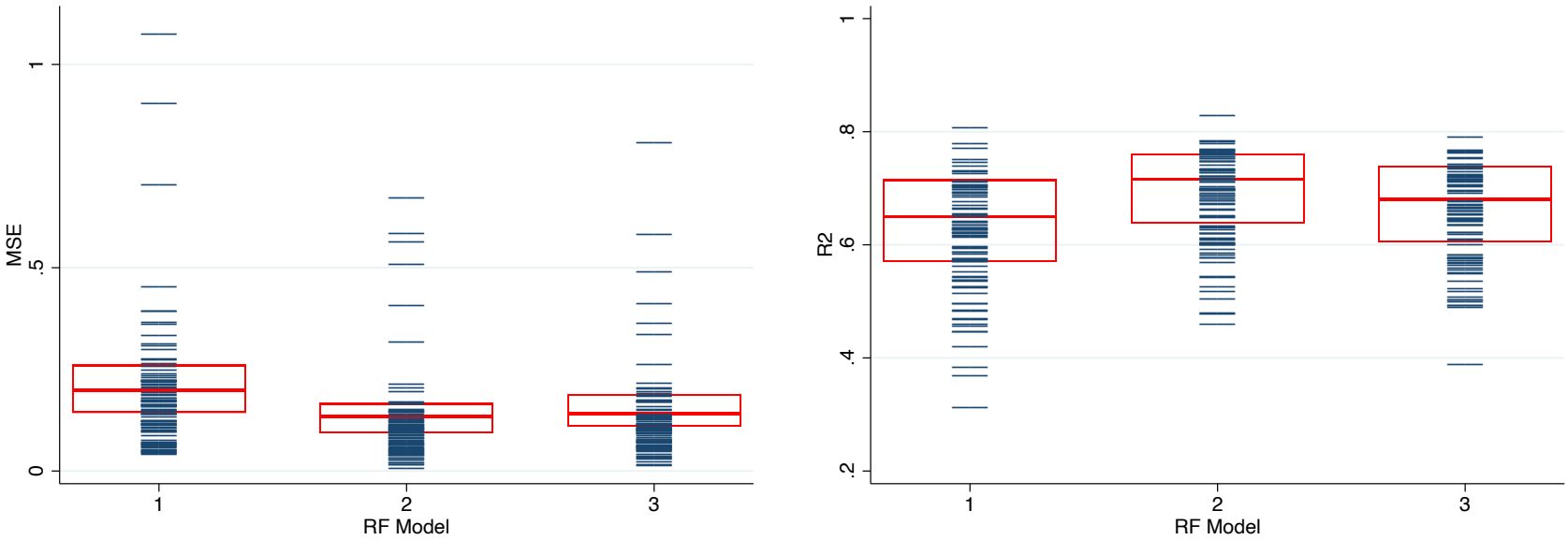
Back to Original Page

Predictive performance at the DHS cluster level

	Median MSE	Median R2
RF-1: NL	0.199	0.650
RF-2: NL, Core	0.135	0.716
RF-3: Core	0.141	0.680
Model 4: KNN with NL	0.242	0.579
Model 5: OLS with NL	0.323	0.391

Back to Original Page

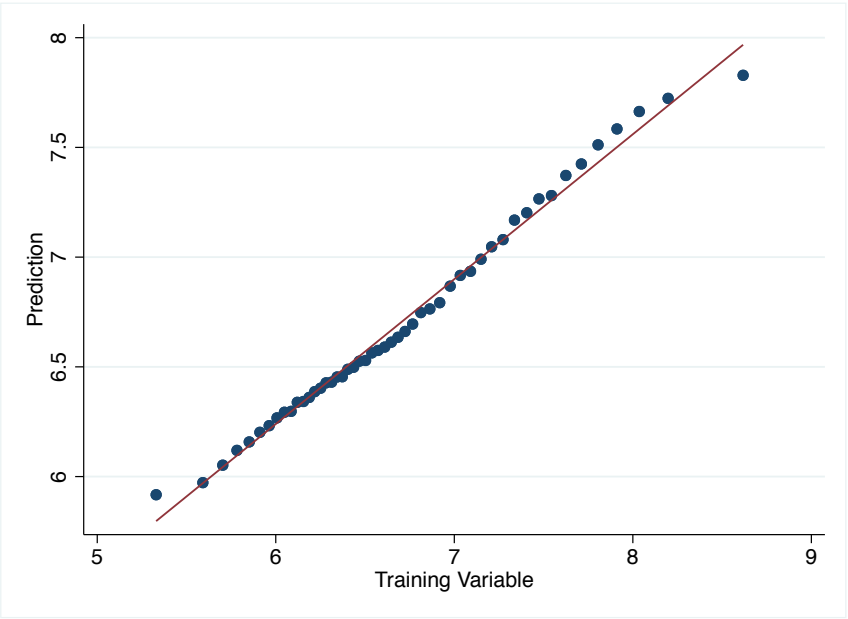
Predictive performance at the DHS cluster level

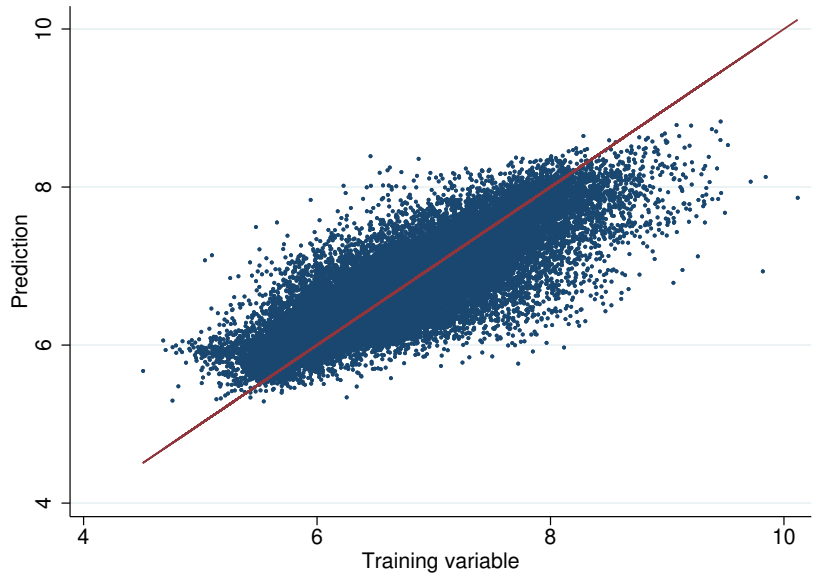


Out-of-sample prediction accuracy. This figure provides the MSE and R^2 for the 85 out-of-sample sets of predictions, corresponding to each of the surveys in our sample. Box and Whisker plots are displayed in red.

Back to Original Page

Predictive performance at the DHS cluster level





Panel (a): binned scatterplot of predicted versus training data

Panel (b): scatter plot, all data points.

Back to Original Page

3.6. Variable Importance

	Relative Variable Importance	
Ranking	RF-2	RF-3
1	Nightlights (.24)	CO_2 (.22)
2	CO_2 (.13)	Population Density (.14)
3	Population Density (.09)	Distance to capital (.07)
4	Distance to capital (.05)	Distance to Christian mission (.05)
5	Distance to Christian mission (.05)	Latitude (.04)
6	Distance to highway (.04)	Distance to highway (.04)
7	Latitude (.04)	Malaria incidence (.04)
8	Malaria incidence (.03)	Longitude (.03)
9	Average Rain (.03)	Average Rain (.03)
10	Longitude (.03)	Distance to protestant mission (.03)

This table provides the 10 most important predictors for models RF-2 and RF-3, together with their relative importance. Importance is relative to the most informative one (whose importance is normalized to 1).

Back to Original Page

3.7. Robustness

Our approach could be implemented with different predictors in different contexts, once a large set of predictors is included, the impact of a specific predictor is small.

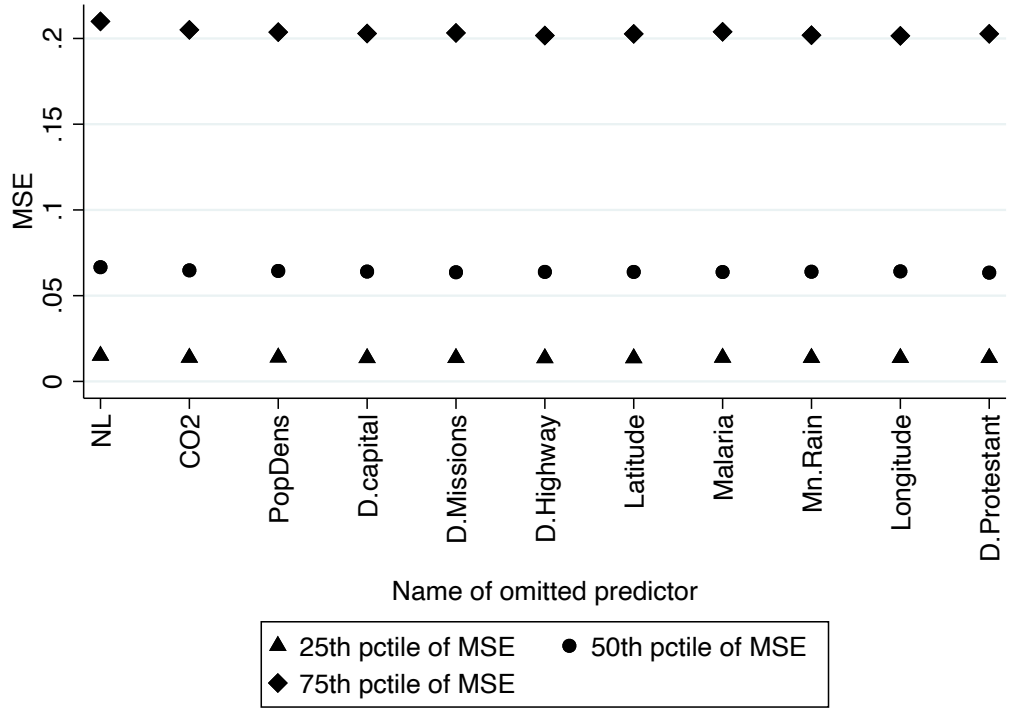


Figure: Out-of-sample prediction accuracy when one of the most important predictors is excluded. This figure provides the value of the MSE for the locations at the 25th, 50th and 75th percentiles in the distribution of MSE. The values are based on 85 out-of-sample sets of predictions when RF-2 is estimated without the variable listed on the x-axis.

Back to Original Page

Predictive performance at the DHS cluster level

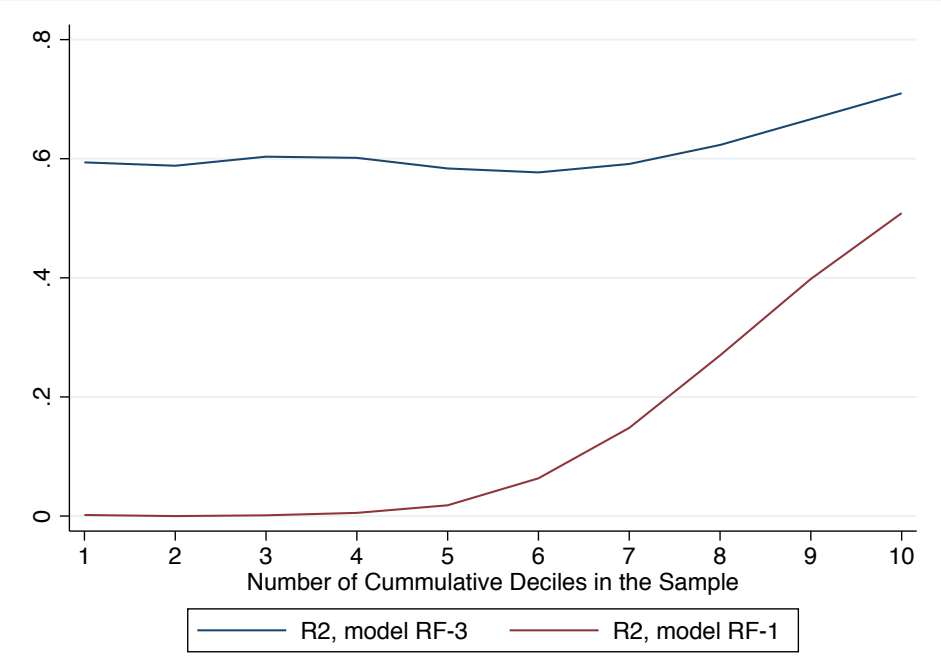


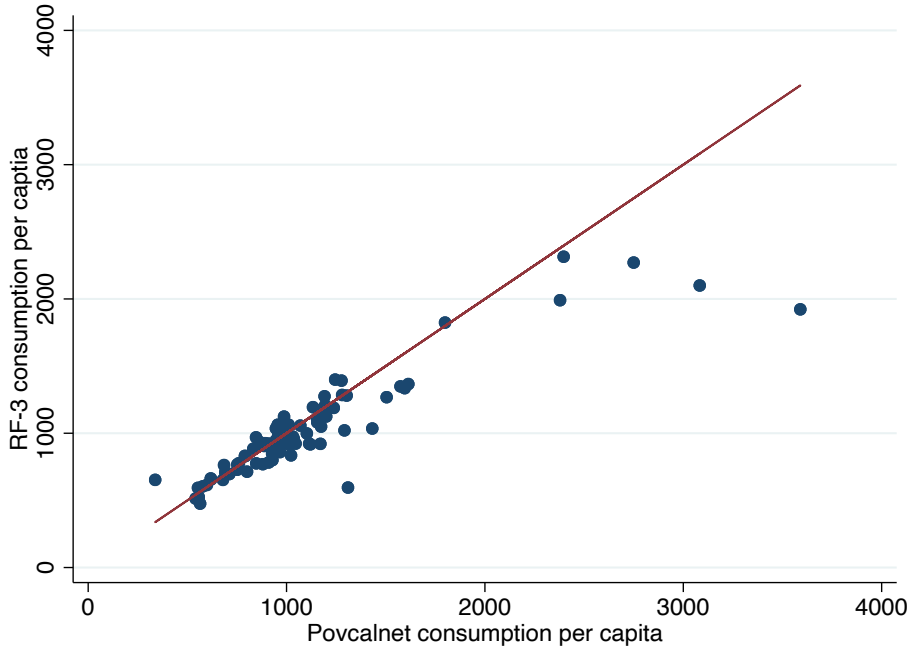
Figure: PERFORMANCE FOR INCREASING SHARES OF DATA USED IN ESTIMATION. The figure plots the R^2s from models estimated on the X smallest deciles of the training data. E.g., if X=2, estimation is carried out on the first 2 deciles of the data.

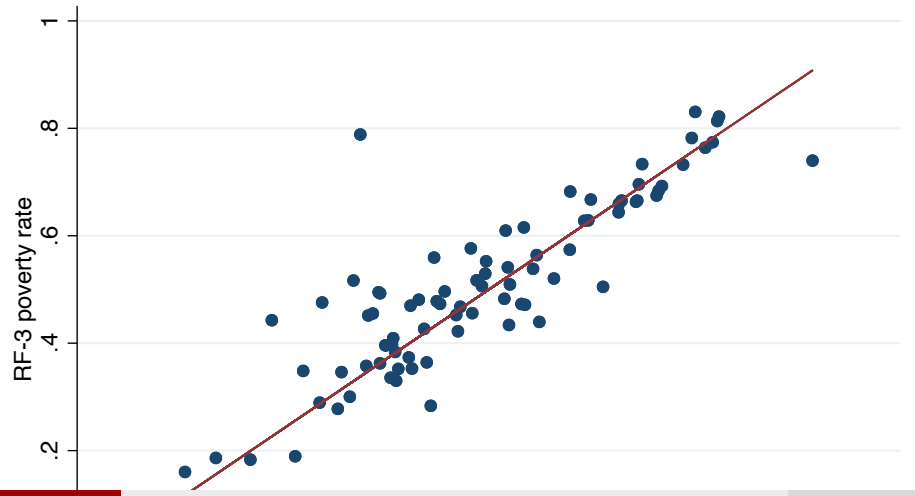
Country consumption and poverty vs Povcalnet

	Panel A: Consumption p.c.			Panel B: Poverty Rate		
	Mean	Std. dev.	Corr.	Mean	Std. dev	Corr.
WB (Povcalnet)	$1106.3	521.6	-	.484	.175	-
RF-1	$999.6	233.9	0.588	.525	.125	0.551
RF-2	$998.5	319.4	0.838	.507	.155	0.833
RF-3	$1022.2	363.8	0.905	.499	.170	0.884
RF-4	$1009.8	325.7	0.859	.502	.159	0.845
RF-5	$1013.4	350.2	0.908	.488	.173	0.885
RF-6	$1004.3	324.6	0.863	.488	.166	0.859

"Corr." is the correlation of the country-level estimate from the RF model with the country-level estimate from Povcalnet.

Country-level comparisons with Povcalnet





4.1. Validation/comparison with other datasets

- Aggregate consumption and poverty estimates at level of subnational regions
- Compare within-country estimates with external data (containing unknown measurement error)
  - HDI and its components: **income per capita**, education index, life expectancy
  - World Bank's regional poverty rates

Back to Original Page

"Validating" within-country variation in consumption and poverty

		HDI (1)	Income (2)	Education (3)	Life Exp. (4)	Poverty (5)
Consump. RF-2	Within r	0.72	0.81	0.65	0.38	
	Between r	0.58	0.64	0.58	0.02	
	Overall r	0.57	0.59	0.56	0.13	
Consump. RF-3	Within r	0.72	0.82	0.67	0.38	
	Between r	0.69	0.70	0.67	0.13	
	Overall r	0.66	0.66	0.64	0.18	
Poverty, RF-2	Within r					0.67
	Between r					0.70
	Overall r					0.70
Poverty, RF-3	Within r					0.69
	Between r					0.82
	Overall r					0.78