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

Laura Mayoral

IAE, Barcelona School of Economics

with John Huber, U. Columbia

Alghero, July 4 2024

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$\longrightarrow$richer)
    - Spatially fine-grained ($\approx$ 1km^2 equator)
    - Whole world since 1992

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)

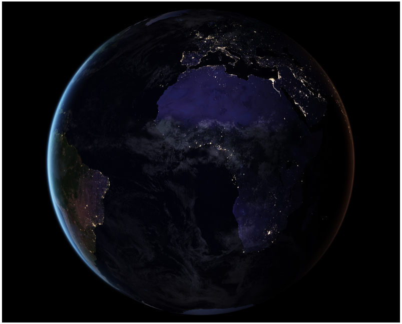
Nightlights: the "go to" spatially disaggregated measure of economic development, II

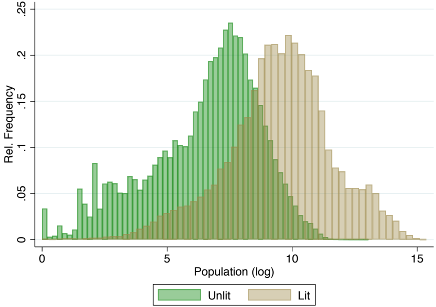
- Limitations

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
    - The "problem of darkness": lack of sensitivity to low lights

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)

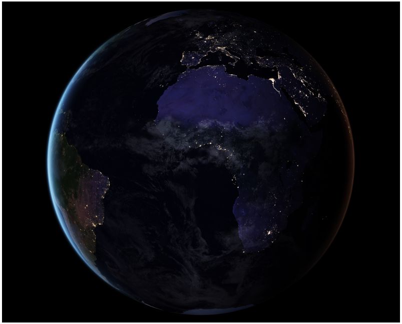


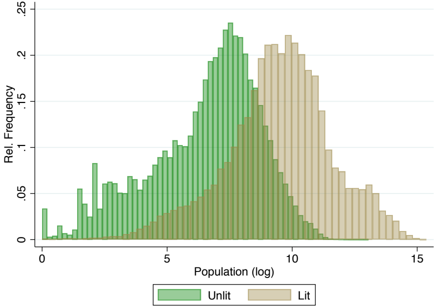


Population in SS Africa at the cell level (10×10 km) in lit vs unlit pixels 2006–2018

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)

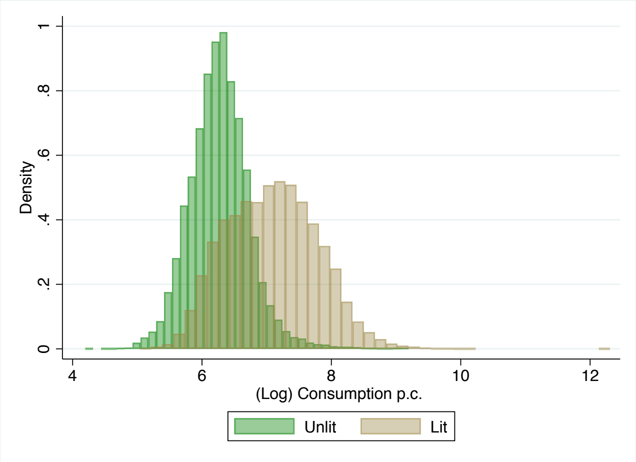




Population in SS Africa at the cell level (10$\times$10 km) in lit vs unlit pixels 2006–2018

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
  - Error is large: NLs are a poor descriptor of economic development
  - 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

This paper: Two 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)
  - Solves the main problems in NLs:
    - Easy to interpret: measured in (consumption) dollars
    - Large accuracy improvement

This paper: Two 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)
  - Solves the main problems in NLs:
    - Easy to interpret: measured in (consumption) dollars
    - Large accuracy improvement

NLs versus SED, Tanzania 2017

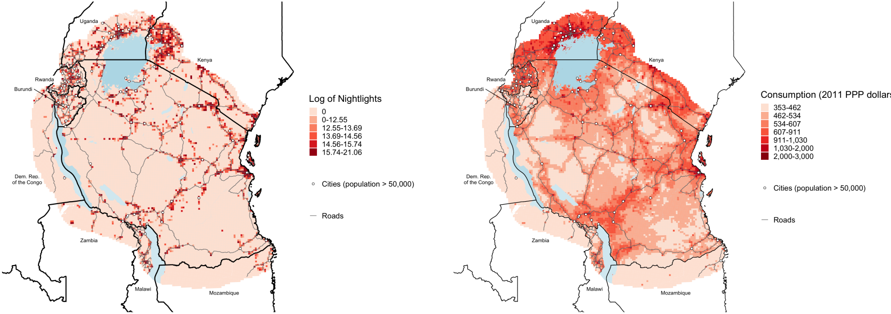


Figure: Nightlights (VIIRS) vs consumption in Tanzania (2017)

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

This paper: Two 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
- Proposes a simple method to deal with non-classical m.e. in regressors generated with supervised machine learning methods
- Revisits two well-known papers on institutions and development, illustrates both types of bias, shows that results are reversed

This paper: Two 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
- Proposes a simple method to deal with non-classical m.e. in regressors generated with supervised machine learning methods
- Revisits two well-known papers on institutions and development, illustrates both types of bias, shows that results are reversed

This paper: Two 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
- Proposes a simple method to deal with non-classical m.e. in regressors generated with supervised machine learning methods
- Revisits two well-known papers on institutions and development, illustrates both types of bias, shows that results are reversed

Summary: Two main takeaways

1. Negative message: the paper warns against the use of NLs in development studies
   - 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 (Nightlights included!): more accurate proxies for development
   - Easy to compute/replicate (STATA)
   - We show how these new indicators can be used in regression to avoid biases

Summary: Two main takeaways

1. Negative message: the paper warns against the use of NLs in development studies
   - 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 (Nightlights included!): more accurate proxies for development
   - Easy to compute/replicate (STATA)
   - 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 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)

- 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

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:

- spatially disaggregated, geolocated, information on consumption, income, etc.
- 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
- 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: allows us to combine different types of datasets; based on (testable) assumptions
2. Combine DHS (individual-level) asset data and WB (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)
- y^C_{ict} is an index of consumption, but 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.

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, but 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.

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, but 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^C_{ict} and y^A_{ict} are related to y^*_{ict} as follows:

$$y^C_{ict} = y^*_{ict} + \epsilon^C_{ict}, \tag{1}$$

$$y^A_{ict} = \alpha_{ct} + \beta_{ct} y^*_{ict} + \epsilon^A_{ict}, \quad \beta_{ct} > 0, \tag{2}$$

The errors ϵ^C_{ict} and ϵ^A_{ict} 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 (2) it can be written as

$$\widetilde{y^*_{ict}} = y^*_{ict} + \widetilde{\epsilon_{ict}}, \text{ where } \widetilde{\epsilon_{ict}} = \epsilon^A_{ict}/\beta_{ct}. \tag{3}$$

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

- 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....

$$E(y_{ct}^{C}) = \underbrace{\mu_{y_{ct}^{C}}}_{\text{observable}} = \frac{-\alpha_{ct}}{\beta_{ct}}, \quad \text{and} \tag{4}$$

$$Var(y_{ct}^{C}) = \underbrace{\sigma^2_{y_{ct}^{C}}}_{\text{observable}} = 1/\beta_{ct}^2 + (\sigma^2_{\epsilon_{ct}^{C}} - \sigma^2_{\epsilon_{ct}^{A}}/\beta_{ct}^2), \tag{5}$$

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:

The variances of the measurement error in y^C_{ct} (the consumption variable) and in $\widetilde{y^*_{ct}}$ (the asset variable transformed by α_{ct} and β_{ct} into a measure of consumption) are similar in magnitude.

Under assumption B:

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

where

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

Computing $\widetilde{y_{ict}^*}$

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_{ict}^A on assets from DHS.
   - country-level data on μ_{ct} and σ_{ct} from Povcalnet

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
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:

$$\bar{y}^*_{ict} = y^*_{ict} + \epsilon^{'}_{ict}. \tag{6}$$

But if it fails

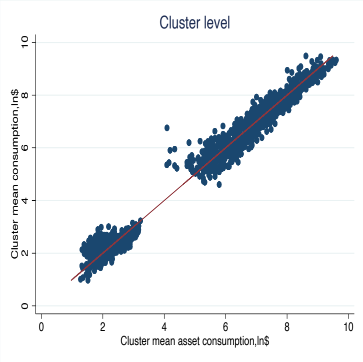
$$\bar{y}^*_{ict} = \overbrace{(\alpha_{ct}\sigma^C_{ct} + \mu^*_{ct})}^{\neq 0} + \overbrace{\sigma^C_{ct}\beta_{ct}}^{\neq 1} y^*_{ict} + \epsilon^A_{ict}\sigma^C_{ct}. \tag{7}$$

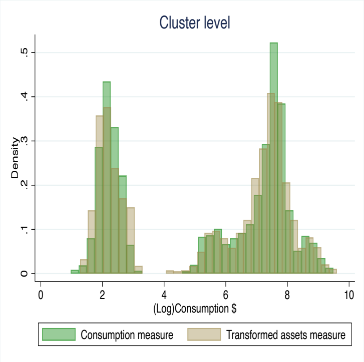
Step 2: Validation using LSMS surveys

- 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, cluster level ($\widehat{y_{ict}^*}$)

Implications I and II: linearity, similar distributions





Validation I: LSMS: Consumption$=b_0+b_1$*Asset Consumption $+\epsilon$

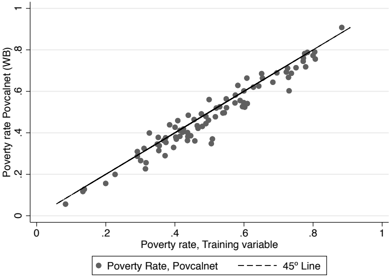
Implication III: $b_0 \approx 0$, $b_1 \approx 1$

Parameter	**Estimate**	**(Standard Error)**
$\hat{b}_0$	0.045	(0.014)
$\hat{b}_1$	0.994	(0.002)
R-squared	0.987	
Obs	2,327	

Validation II: Transformed DHS vs. Povalnet, deciles and poverty rates

1. Country-year distributions of the transformed asset index and Consumption data (povcalnet) should be similar
2. Compare:
   - Country-level deciles Povcalnet (World Bank)
   - Country-level poverty lines, Povcalnet (World Bank)

Poverty rates: Transformed DHS index vs Povcalnet



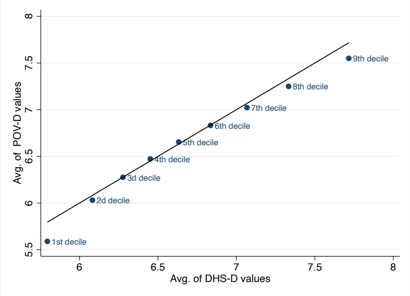


Figure: Country-year Poverty Rates (1.9$ a day threshold) and Deciles. Correlation of poverty lines is 0.97.

Validation using DHS: Mean of country-year deciles and poverty rates using DHS assets and Povcalnet

	Povcalnet (World Bank)	DHS Training Variable	RMSD (Root Mean Square Error)
Poverty rate	0.48	0.50	0.04
Decile 1	$268	$329	$24
Decile 2	$416	$438	$13
Decile 3	$532	$532	$10
Decile 4	$647	$633	$15
Decile 5	$776	$759	$22
Decile 6	$927	$930	$23
Decile 7	$1,122	$1,173	$31
Decile 8	$1,407	$1,529	$52
Decile 9	$1,900	$2,245	$115

Table: This table provides average decile and poverty rates values for the 85 country-years in our DHS sample from Povcalnet (World Bank) and from the distribution of $\widehat{y^*_{ct}}$. RMSD denotes the root of the mean square error.

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

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. ⇒ 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.
- The implications in regression of these two cases are very different, 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 classical 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) .

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 classical 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.

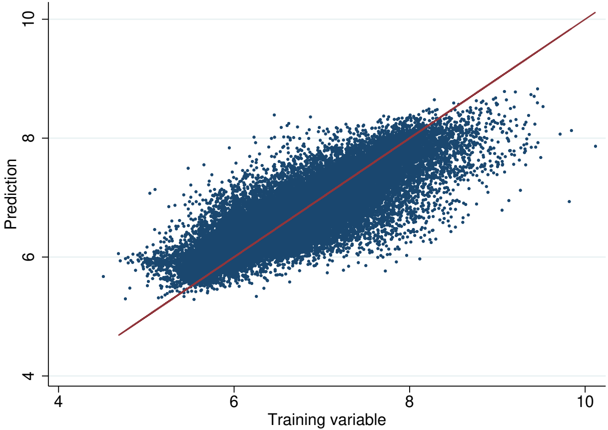
- if 'continuous NL' (given problem of darkness), same logic applies

- If X and y^* are correlated, it's reasonable to expect that X and u will be correlated as well.

In sum

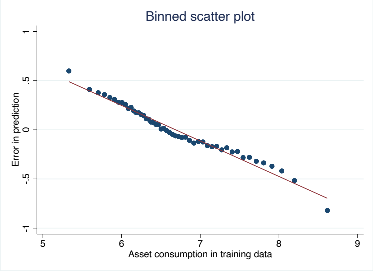
- Nightlights contains non-classical m.e.
- Leads to biases in regression coefficients
  - Both when NLs is the dependent or the independent variable
  - Biases in any direction! (not only "attenuation")
- Solution: Can we use SED instead?

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 $\hat{y}$ be a proxy for y with non-classical m.e.$\Rightarrow$ μ is a function of $\hat{y}$).

$$\hat{y} = y + \mu, \quad \mu = f(y) + \epsilon, \quad cov(\epsilon, y) = 0.$$

Assume that $f(.)$ is linear: $\mu = \alpha_0 + \alpha_1 y + \epsilon$. Then:

$$\hat{y} = \alpha_0 + (1 + \alpha_1)y + \epsilon.$$

Thus, a new proxy for y is $\hat{y_2}$:

$$\hat{y_2} = \frac{\hat{y} - \alpha_0}{(1 + \alpha_1)} = y + \epsilon/(1 + \alpha_1).$$

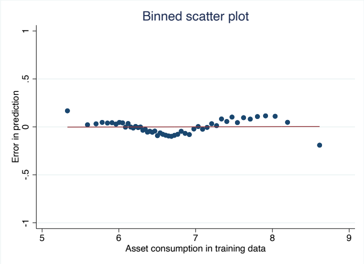
The proxy $\hat{y_2}$ only contains classical measurement error

$$\hat{y_2} = \frac{\hat{y} - \alpha_0}{(1 + \alpha_1)} = y + \epsilon / (1 + \alpha_1).$$

- Computation:
  - Regress predicted cluster consumption on actual consumption using DHS training data to obtain α_0 and α_1
  - Use these coefficients to transform $\hat{y}$ into $\hat{y_2}$
- $\hat{y_2}$
  - has the same correlation with y as does $\hat{y}$
  - 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 with only classical measurment 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, $\hat{y}$, and $\hat{y}_2$

	Mean	Std	Min	Max	MSE	Corr with y	Corr with u
y	6.72	.72	4.51	9.84	–	–	–
$\hat{y}$	6.71	.57	5.31	8.90	0.18	0.814	-0.62
$\hat{y}_2$	6.72	.89	4.52	10.15	0.27	0.814	0.0025

Table: SUMMARY STATISTICS FOR y, $\hat{y}$, AND $\hat{y}_2$. MSE DENOTES MEAN SQUARED ERROR.

A trade-off

- $\hat{y}_2$ has only classical measurement error (zero correlation of error with y)
- $\hat{y}_2$ has larger measurement error (50%)

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 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!
- 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)

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)

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)

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

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 $\hat{y}_2$)

	(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 $\hat{y}$)

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 $\hat{y}_2$)

	(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 $\hat{y}$)

	(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 → 45% increase in consumption per capita
- Going from the lowest to highest value of RULE OF LAW → 183% increase in consumption

Jurisdictional hierarchy: (ignore null effect)

- Using NL-only RF model: A one-unit increase in JH → 2.6% increase in consumption
- Using NL-only RF model: Worst to best JH → 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

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 (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

- **Nightlights.**
- *Core* variables:
  - *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**
  - *Other*: **Population, CO2 production**
- *Aggregate variables*: country-level consumption p.c. or GDP per capita (interacted with NL and core)

Time-varying predictors are in bold.

Back to Original Page

3.3. Prediction: Prediction models

RF-1	NL only
RF-2	NL + core
RF-3	NL + core + country level consumption*(NL+core)
RF-4	NL + core + country level GDP*(NL+core)
RF-5	core + country level consumption*(NL+core)
RF-6	core + country level GDP*(NL+core)

Table: The random forest models.

Back to Original Page

3.4. Prediction: Parameter tuning

	Preferred Hyperparameters Values			
	NTREES	NVARS	DEPTH	VAR
Model 1	120	3	unrestricted	.000005
Model 2	180	8	20	.0005
Model 3	180	9	unrestricted	.00005
Model 4	180	9	unrestricted	.000005
Model 5	180	8	unrestricted	.000005
Model 6	180	8	unrestricted	.000005

Table: Preferred hyperparameter values employed in Models 1–6.

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 highly competitive–outperforms existing models (i.e., Yeh et al. (2020), Nature)

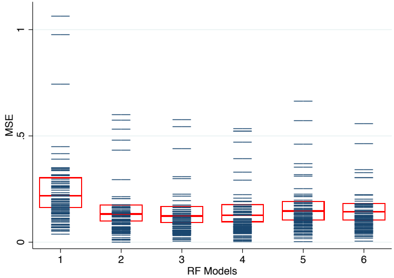
Back to Original Page

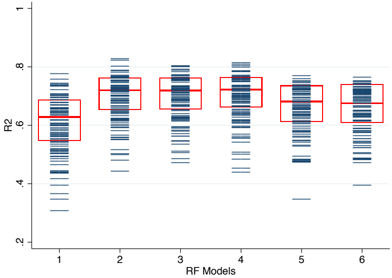
Predictive performance at the DHS cluster level

	Median MSE
RF-1: NL	.219
RF-2: NL, Core	.132
RF-3: NL, Core, Consump.×NL, Consump.×Core	.124
RF-4: NL, Core, GDP×NL, GDP×Core	.126
RF-5: Core, Consump.×Core	.143
RF-6: Core, GDP×Core	.146
Model 7: KNN with NL	.234
Model 8: OLS with NL	.320

Back to Original Page

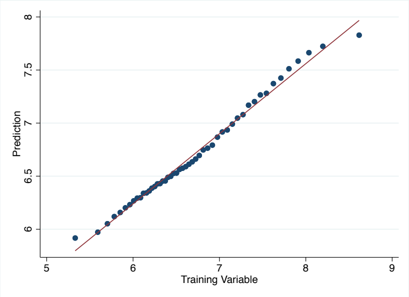
Predictive performance at the DHS cluster level

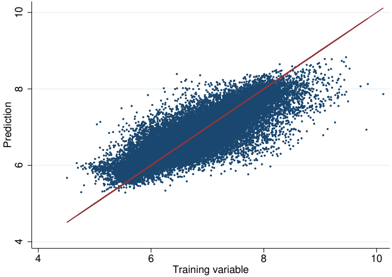




Back to Original Page

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	deserts (1)	Cons×NLs (1)	ecosystem: des
2	NLs (3 yr mean) (.48)	ecosystem: deserts (.45)	GDP×NI
3	CO_2 (.38)	Cons×Co2 (.33)	GDP×Co
4	population (.25)	Cons×population(.24)	NI
5	grassland (.17)	Cons× remoteness (.22)	Co
6	NL(VIIRS) (16)	NLs (.20)	GDP×populatio
7	disease (.15)	Co2 (.15)	ecosystem: grasslan
8	latitude (.14)	population (.11)	latitu
9	NL(DSMP, blur) (.11)	Cons×disease (.09)	populatio
10	NL(DSMP, deblur) (.11)	ecosystem: grasslands (.096)	GDP×disea

This table provides the 10 most important predictors for models RF-2, RF-3 and RF-4, together with their relative importance. Importance is relative to the most informative one (whose importance is normalized to 1).

Back to Original Page

Predictive performance at the DHS cluster level

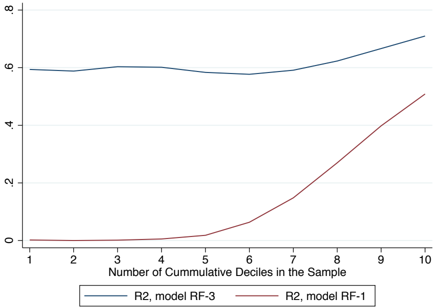


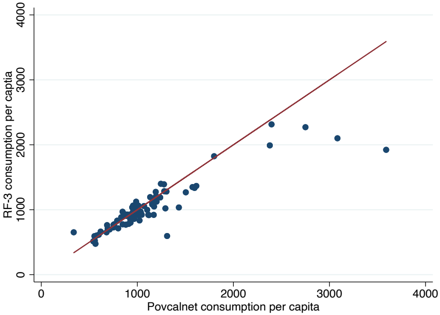
Figure: PERFORMANCE FOR INCREASING SHARES OF DATA USED IN ESTIMATION. The figure plots the R^2s from models estimated on the X smallest deciles of the training data. E.g., if X=2, estimation is carried out on the first 2 deciles of the data.

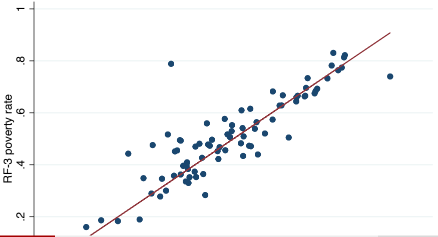
Country consumption and poverty vs Povcalnet

	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