Poverty in HD: What Does High-Resolution Satellite Imagery Reveal About Poverty?

Ryan Engstrom
GWU
Dept. of Geography
rengstro@gwu.edu

Jonathan Hersh
Boston University
Department of Economics
jhersh@bu.edu

David Newhouse
World Bank
Poverty Global Practice
dnewhouse@worldbank.org

WORLD BANK GROUP

World Bank ABCDE Conference

June 20, 2016

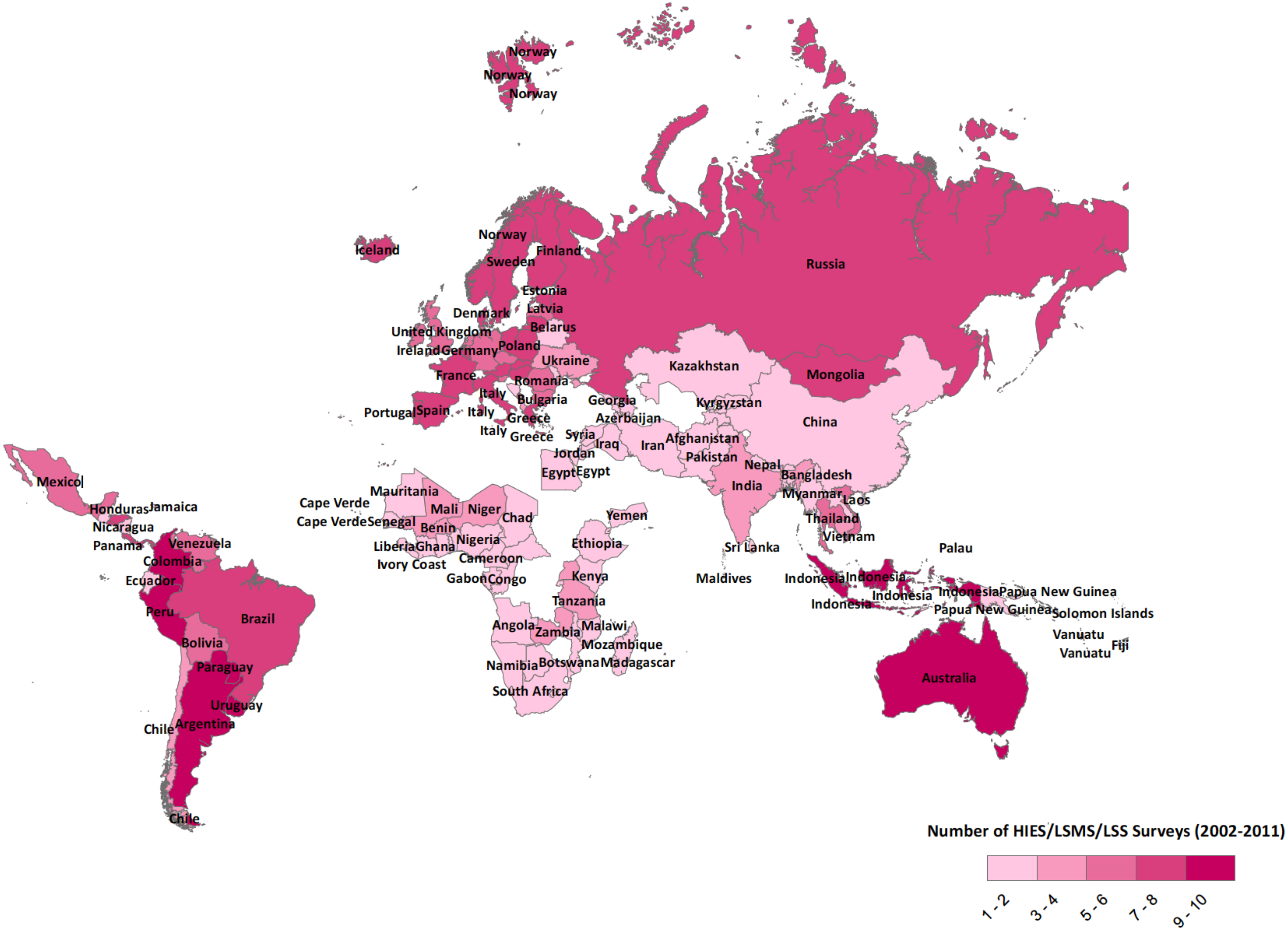
WORLD BANK GROUP

Motivation

1. Poverty data on the poverty rates of local areas are in scarce supply

2. Even when countries collect poverty data, they often can’t collect it in areas where it’s needed most

Number of Poverty Data Points, 2002 - 2011



Motivation

1. Poverty data, on the poverty rates of local areas, are in scarce supply
2. Even when countries collect poverty data, they often can’t collect it in areas where it’s needed most

One Overlooked Piece of Data: Very High Resolution Satellite Imagery

Paper Overview

Examine potential of features derived from very high resolution satellite imagery (VHRSI) to:

1. Estimate poverty at local areas using only VHRSI features as explanatory variables

2. Extrapolate poverty estimates into areas not covered by surveys

Results Preview

- Features from VHRSI explain 40-70% of variation in small area poverty.

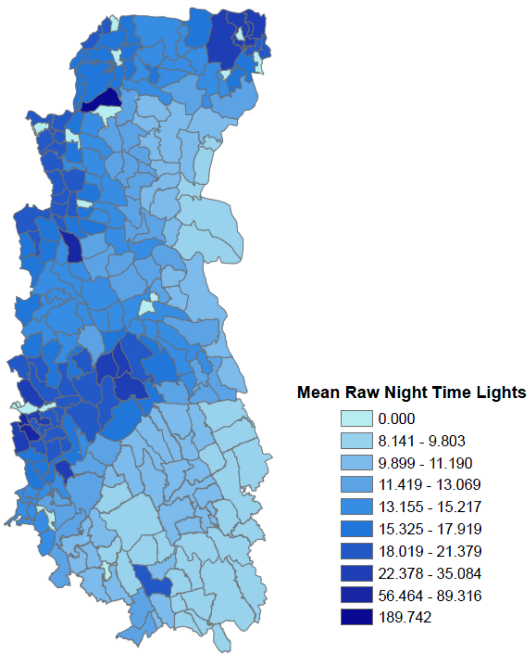
- Extrapolations are less precise, but we can generate fairly accurate rank order

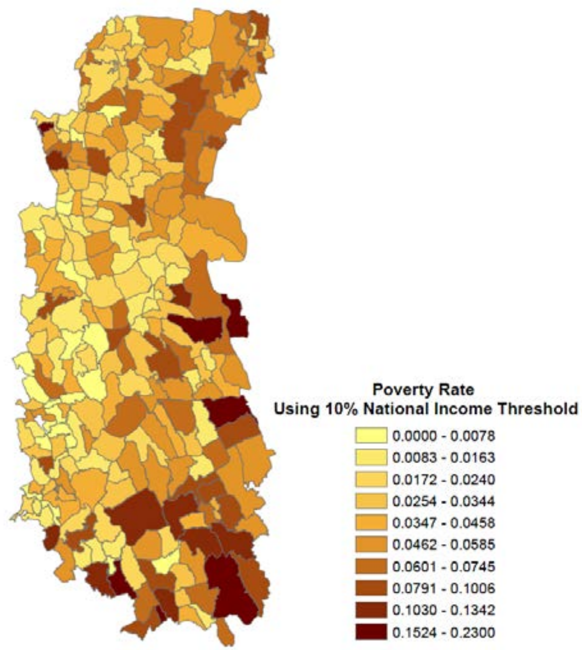
Related Literature on Remotely Sensing Human Welfare

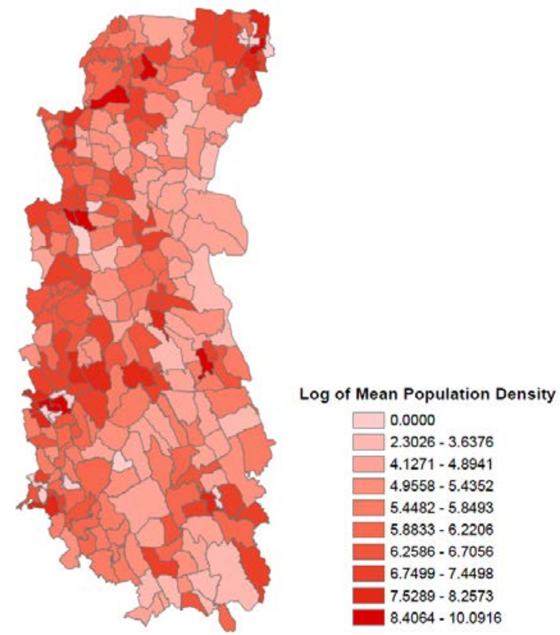
- **Night Time Lights -** Henderson, Storeygard, and Weil (2012)
- **Transfer Learning -** Xie, Jean, Burke, Lobell, and Ermon (2016)
- **Bayesian Geostatistical Modeling -** Tatem, Gething, Pezzulo, Weiss, Bhatt (2014)
- **Google Street View Imagery to Predict Housing Prices** - Glaeser, Kominers, Luca, Naik (2015)

Our project: First to use very high resolution imagery, use census based poverty estimates, and to measure through classification of correlates of poverty. .

Why Not Just Use Night Lights?

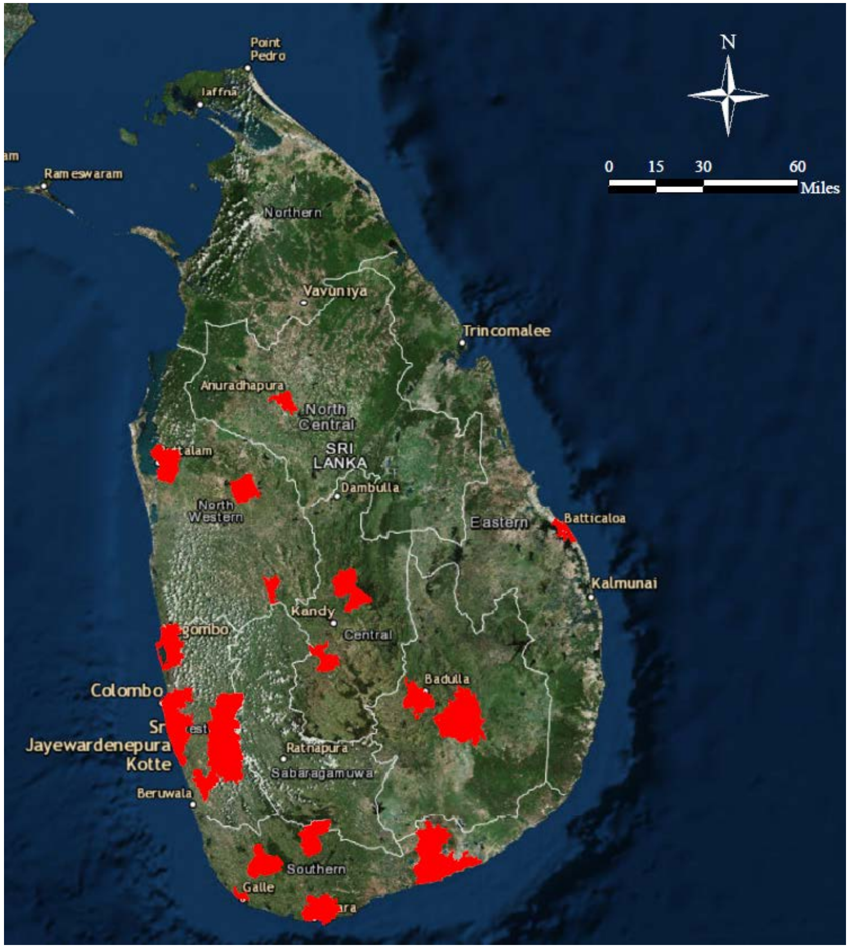






Raw Imagery Description

- High resolution (< 0.5 m pixel)
- 3,500 sq. km in Sri Lanka
- Covering 1,250 of the 13,000 Gram Niladhari (GN) Divisions
- Match to poverty data imputed into the 2011 Census



WORLD BANK GROUP

Unit of Analysis

- Unit of Analysis: GN Division

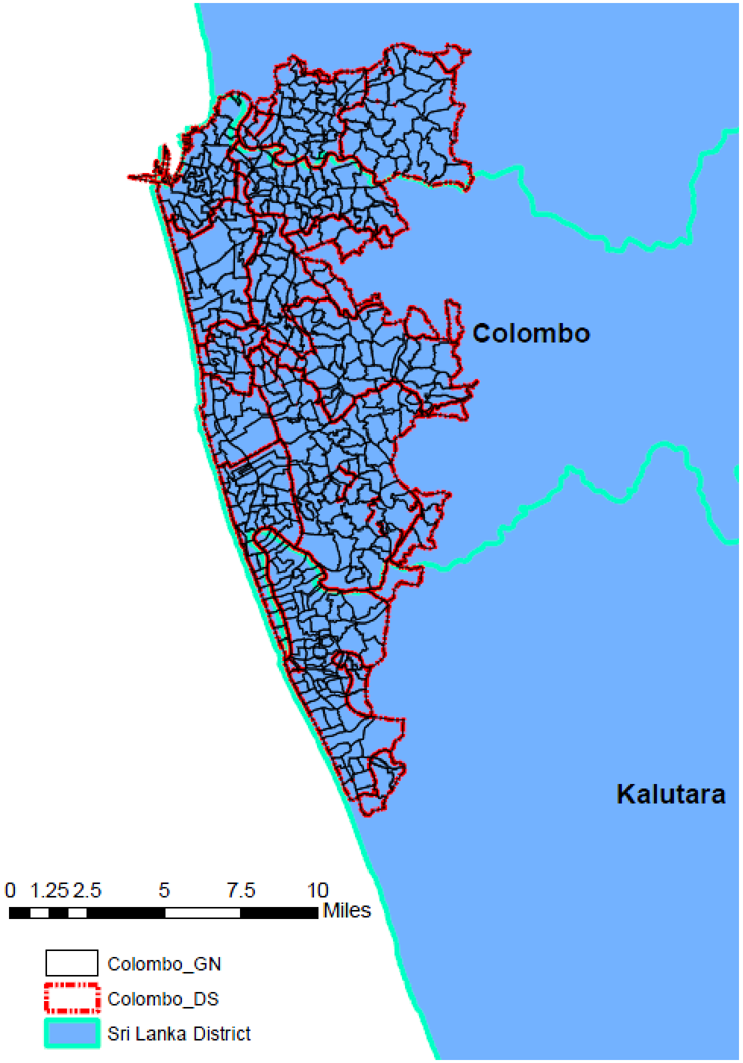
- Average Size

  ~ 10,000 persons

  ~ 2.15 sq. km.

  ~ 1/60th size US Census Tract

  ~ 2.5 times the size of Census Block

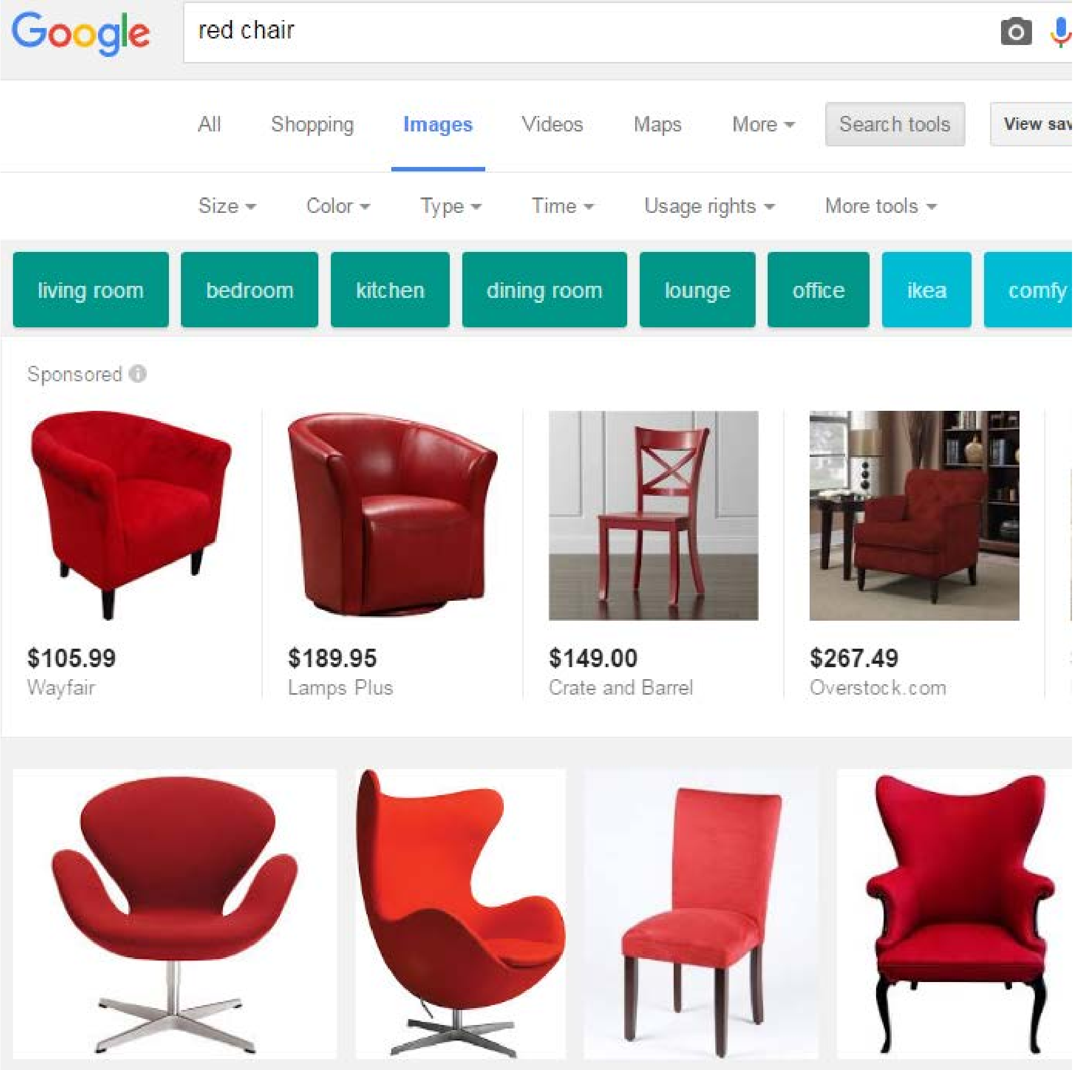


WORLD BANK GROUP

What “features” Do We Derive From Satellite Imagery?

- Machine vision algorithms extract meaning from raw images
- Two types:
  - Identify Objects
  - Identify Texture & Spectral Characteristics

Example Identifying Objects



WORLD BANK GROUP

Features Extracted from High Resolution Imagery

Object Identified Features

- Number of Buildings
- Number of Cars
- Fraction Roads Paved
- Shadow Pixels (Building Height)
- Crop Type/Extent
- Roof Type

Texture and Spectral Features

- Vegetation Index (NDVI)
- PanTex (settlement density)
- HoG
- Local Binary Pattern Moments
- Line Support Region
- Gabor Filter
- Fourier Transform
- SURF

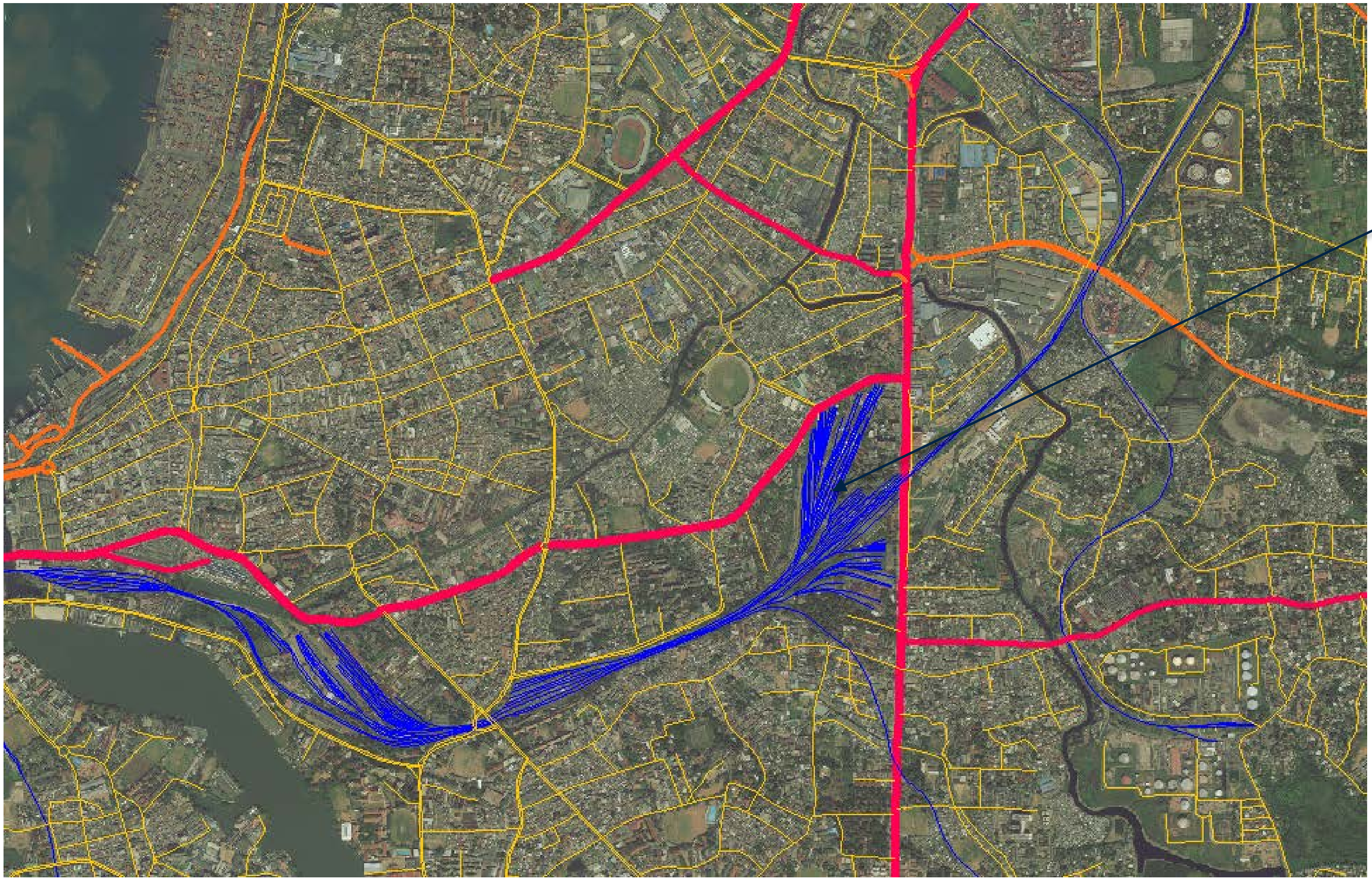
Technical Partners

ORBITAL INSIGHT

LANDinfo WORLDWIDE MAPPING LLC

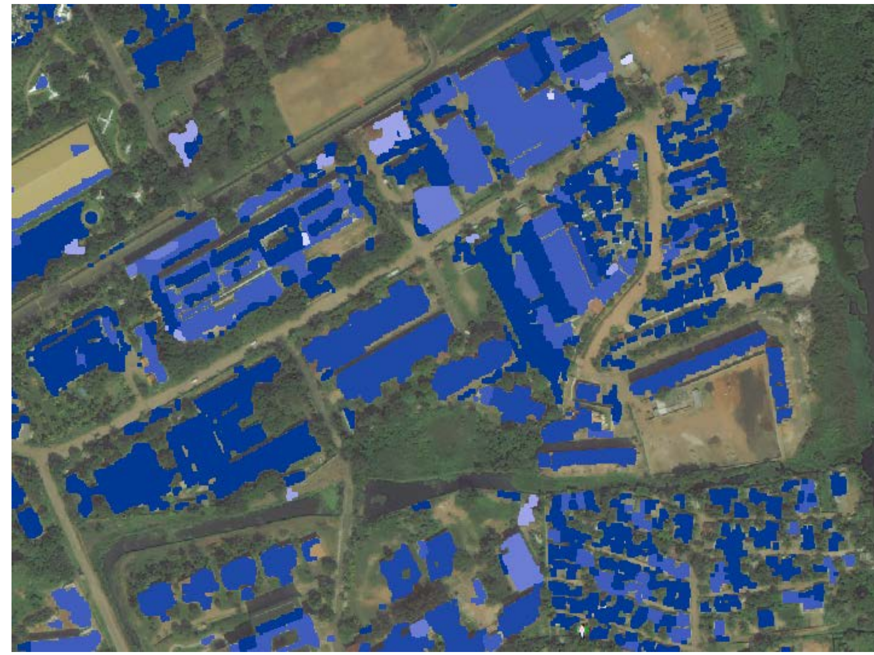
GW DEPARTMENT OF GEOGRAPHY

Example Identified Object: Road Width



WORLD BANK GROUP

Example Identified Object: Roof Type



- Aluminum White/Light Grey
- Asbestos Light Brown
- Clay Tiles Dark Brown
- Grey
- Painted Aluminum Blue
- Painted Aluminum Green

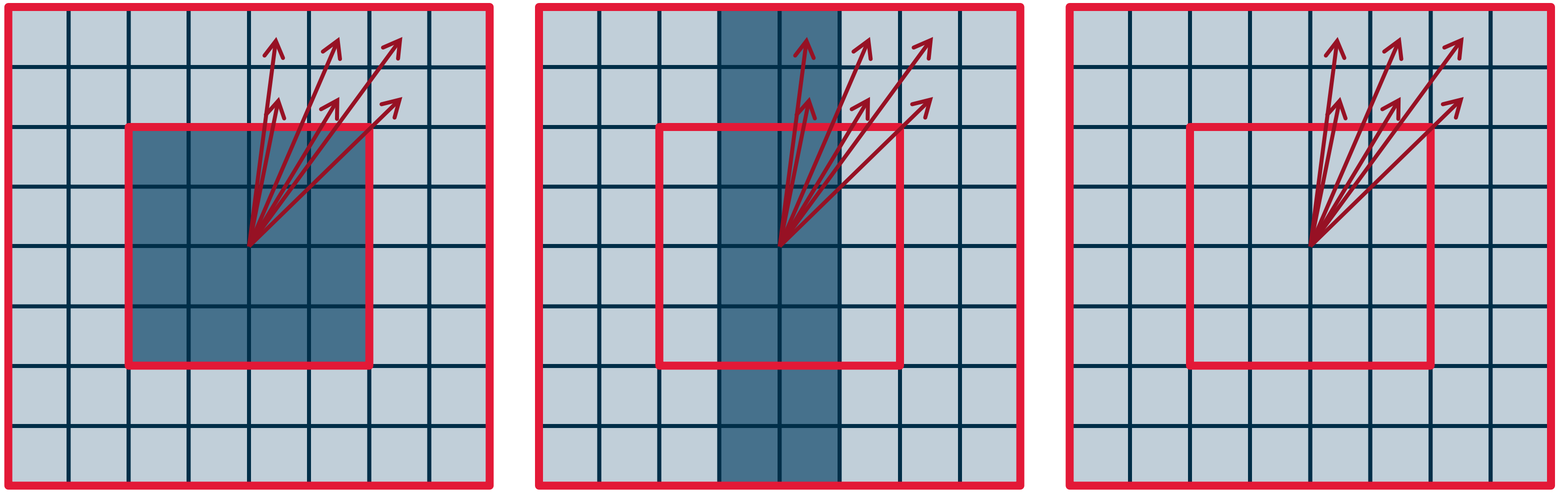
Example Identified Object: Cars

WORLD BANK GROUP

Example Texture/Spectral: PanTex

PanTex (Pesaresi et al. 2008)

- Detects minimum contrast in every direction
- Measures density of settlements and built-up area



Building: PanTex returns a high contrast value

Road: PanTex returns a low contrast value

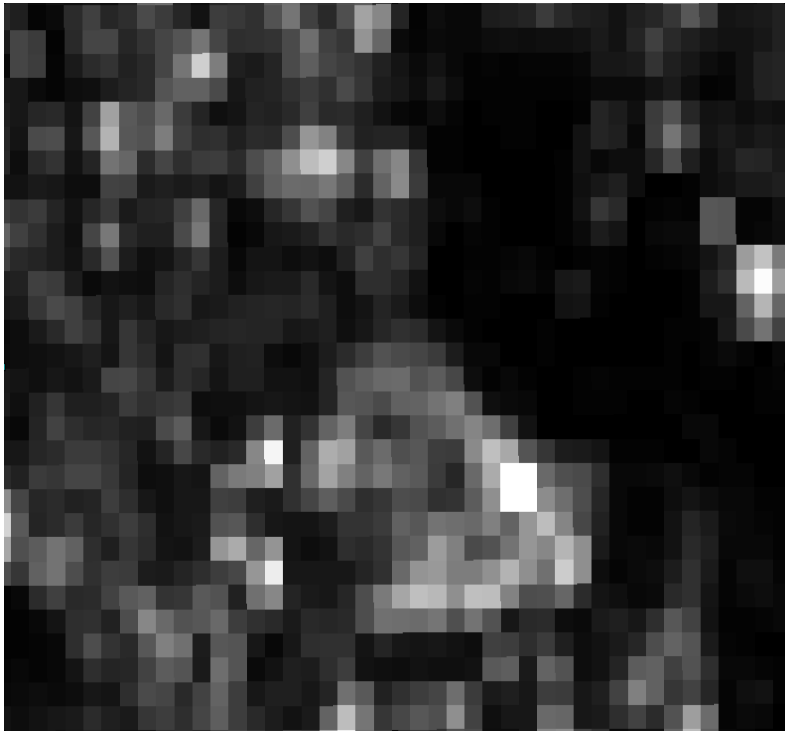
Flat Surface: PanTex low contrast value

WORLD BANK GROUP

Example Texture/Spectral: PanTex

Raw Imagery

PanTex



Wanathamulla neighborhood

WORLD BANK GROUP

Baseline Empirical Methodology

(1)
Without Scene Fixed Effects

$$\underbrace{y_j}_{poverty\ rate\ in\ GN\ j} = \underbrace{X_j{}'\beta}_{satellite\ features} + \underbrace{\Delta_j \Lambda}_{GN\ controls} + \underbrace{\varepsilon_j}_{error}$$

(2)
With Scene Fixed Effects

$$\underbrace{y_j}_{poverty\ rate\ in\ GN\ j} = \underbrace{X_j{}'\beta}_{satellite\ features} + \underbrace{\Delta_j \Lambda}_{GN\ controls} + \underbrace{Z_s'\theta}_{imagery/scene\ FEs} + \underbrace{\delta_j}_{error}$$

WORLD BANK GROUP

OLS Results, National Models (Object Features)

Variable	10% Poverty Rate		40% Poverty Rate	
	b	t	b	t
urban	-0.0020	[-0.19]	-0.050	[-1.25]
log GN Area	0.010*	[2.22]	0.029*	[2.02]
% of roads that are paved	-0.00033***	[-3.87]	-0.0013***	[-4.30]
% of GN area that is road	1.08	[1.03]	3.06	[0.98]
% of roads that are railroad	0.00015	[0.38]	-0.00022	[-0.18]
% of valid GN area that is built up	-0.0029*	[-2.24]	-0.011*	[-2.33]
% shadow pixels covering valid area (building height)	0.0024	[1.53]	0.012**	[2.78]
Fraction of total roofs that are clay	0.00021	[0.92]	0.00073	[1.04]
Fraction of total roofs that are aluminum	0.00074	[1.92]	0.0024*	[2.02]
Fraction of total roofs are asbestos	-0.00036	[-1.03]	-0.0015	[-1.56]
log number of roofs count	-0.012**	[-3.12]	-0.045***	[-3.89]
Total cars divided by total road length	-0.39	[-1.84]	-1.13	[-1.66]
Total cars divided by total GN Area	41.2	[0.91]	101.5	[0.66]
log number of cars	0.0018	[0.47]	0.0044	[0.43]
% of GN area that is agriculture	-0.062	[-1.18]	-0.064	[-0.30]
% of GN agriculture that is paddy	0.00050*	[2.16]	0.00032	[0.22]
% of GN agriculture that is plantation	0.00055*	[2.58]	0.00056	[0.39]
% of Total GN area that is paddy	0.000073	[0.12]	-0.00096	[-0.39]
% of Total GN area that is plantation	0.00042	[0.98]	0.0011	[0.62]
Constant	-0.021	[-0.19]	0.23	[0.53]
…				
Observations	1244		1244	
R Squared	0.39		0.59	

Dependent variable is log of GN Poverty Rate Defined at X% of national consumption

OLS Results, National Models (con't, Texture Features)

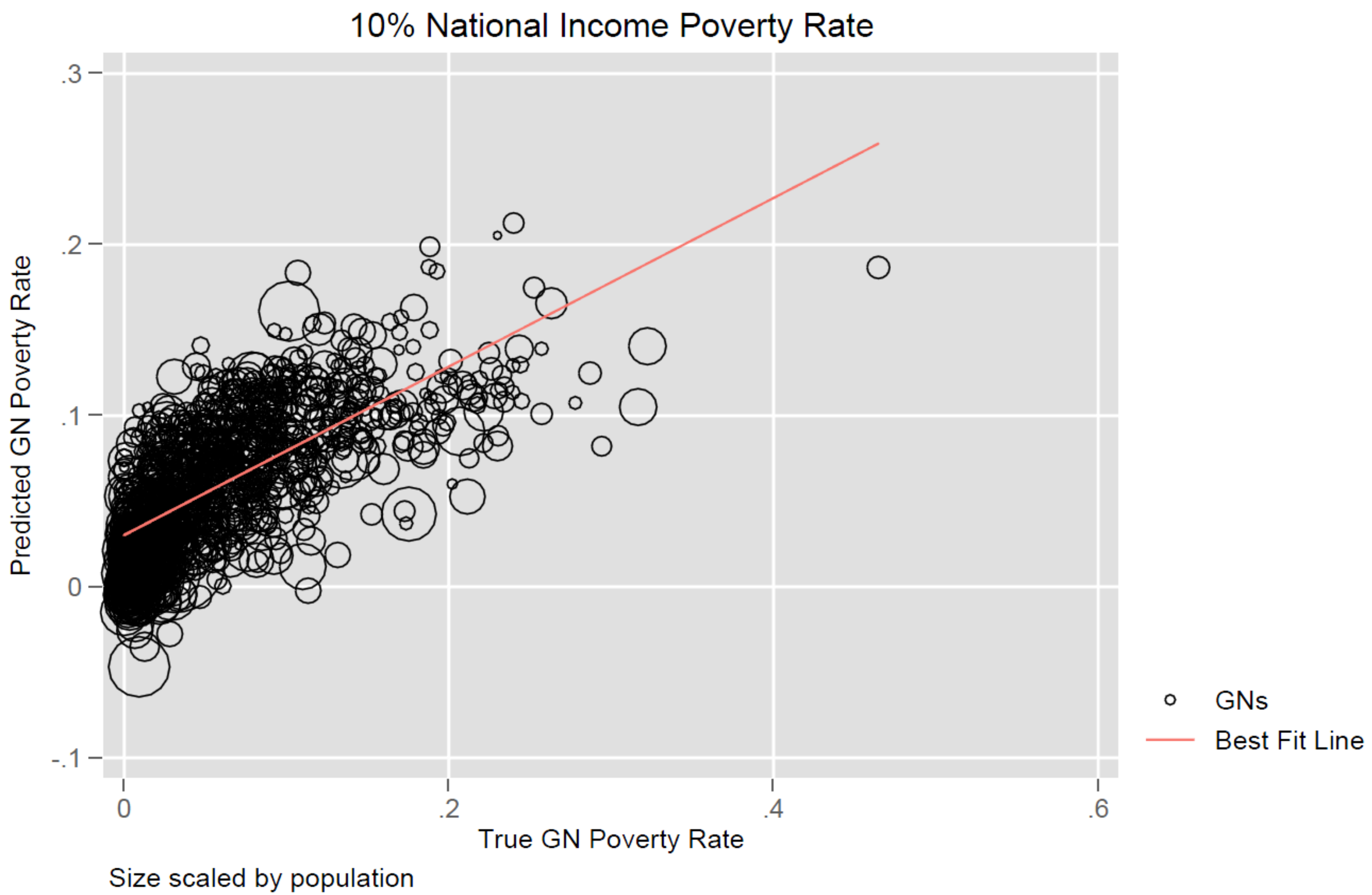
Variable	10% Poverty Rate		40% Poverty Rate	
	b	t	b	t
…				
NDVI	0.062*	[2.01]	0.22**	[2.80]
Pantex (human settlements) mean contrast	0.022	[1.78]	0.065*	[2.25]
Histogram of oriented gradients (HOG)	-0.000016*	[-2.12]	-0.000057**	[-3.39]
Local Binary Pattern (moments) skewness	-0.00032	[-0.72]	-0.00061	[-0.47]
Line support region mean - scale 8	-0.33	[-1.27]	-0.23	[-0.31]
Gabor filter mean - scale 64	0.070	[1.60]	0.19	[1.76]
Fourier transform std. dev. - scale 32	0.0034	[1.60]	0.0083	[1.24]
Surf - scale 16	-0.00013	[-1.44]	-0.00036	[-1.13]
Constant	-0.021	[-0.19]	0.23	[0.53]
Obs	1244		1244	
R Squared	0.39		0.59	

Dependent variable is log of GN Poverty Rate Defined at X% of national consumption

Results Discussion

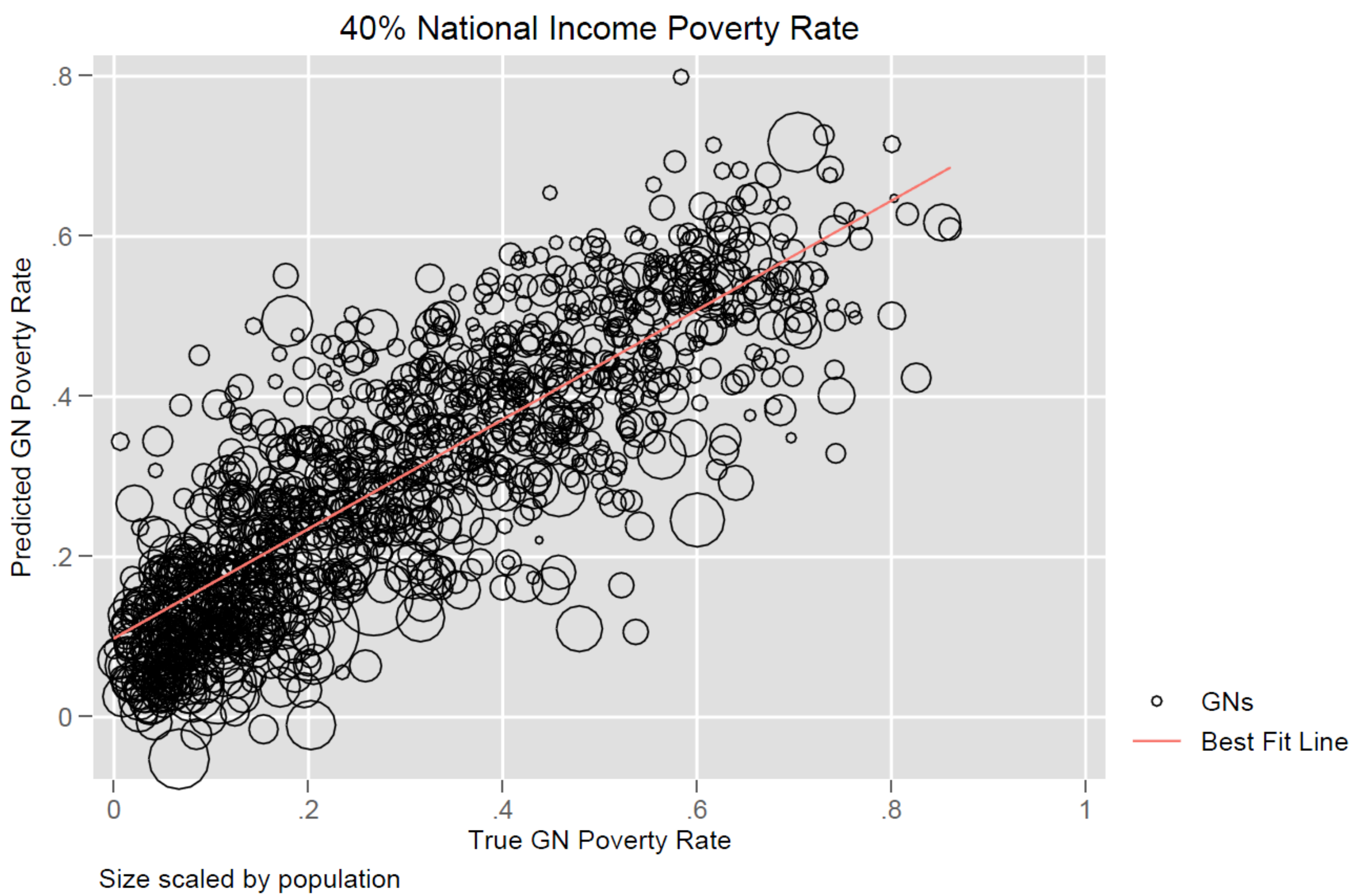
- Robust Object ID'd predictors: develop area, number of buildings, roof type, fraction roads paved
- Robust Texture/Spectral predictors: NDVI (vegetation index), PanTex (building density), HoG (gradients/straightness) of buildings
- Separate urban and rural models show different spatial patterns of poverty in urban and rural areas
  - In urban areas: NDVI negatively correlated with poverty
  - In rural areas: NDVI positively correlated

Predicted Versus True Plots – 10% National Income



WORLD BANK GROUP

Predicted Versus True Plots – 40% National Income



WORLD BANK GROUP

Shapley Decomposition of Share of Variance Explained

	Avg. Consumption in GN	10% poverty rate	20% poverty rate	30% poverty rate	40% poverty rate
Urban	8.6	2.5	3.6	4.6	5.8
Log of GN Area	6.7	7.9	7.7	7.5	7.3
Road variables	11.7	12.1	12.2	11.8	11.5
Building density variables	36.4	37.8	36.9	36.4	36.1
Of which: Built-up area	18.6	12.8	11.7	12.4	13.3
Log Number of roofs	8.8	9.4	10.8	10.4	10.0
Shadow (building index)	5.2	3.8	4.0	4.3	4.6
NDVI	3.8	11.8	10.4	9.3	8.3
Roofs	9.3	7.6	7.5	7.5	7.8
Cars	5.2	4.5	4.4	4.3	4.2
Agricultural land variables	4.9	6.3	7.3	7.5	7.3
Texture variables	17.3	21.4	20.4	20.5	20.0
Total r^2	**0.64**	**0.39**	**0.50**	**0.55**	**0.59**

WORLD BANK GROUP

Costs

- $90,000 Total Project Costs (Big Data Innovation Challenge Grant, DEC SRP)
  - $20,000 Imagery
  - $20,000 Imagery Processing (orthorecification)
  - $50,000 Processing and Deriving features

- However, business model moving towards imagery rental
  - Can analyze & extract features without paying imagery costs
  - This will scale

Conclusions

- We can explain 40-60 percent of the variation in poverty using only variables derived from high resolution satellite imagery
  - Lasso does a bit better, explaining 40-70 percent
  - Support Vector Machine (SVM) models even better
- Building density, built up area strongest predictors.
  - Vegetation index, roof type, shadow pixels (building height), and texture variables also strong predictors
- Extrapolating to out of sample areas less accurate but preserve rank

Implications and Next Steps

Implications

- May be possible for high res satellite indicators to substitute for census data in estimating poverty maps
- Understand better the tradeoffs of using more frequent higher variance poverty maps versus outdated but more accurate poverty maps for targeting
- Would this help adjust for non-response in surveys?

Next Steps

- Which features forecast changes in welfare?
- Cost/performance tradeoff of different features