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

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

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

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

World Bank ABCDE Conference  
June 20, 2016
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


- **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?
• 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
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
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
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

[Images of technical partners logos]
Example Identified Object: Road Width

Road width (in meters)
Example Identified Object: Roof Type
Example Identified Object: Cars
**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 returns a low contrast value**
Example Texture/Spectral: PanTex

Raw Imagery

PanTex

Wanathamulla neighborhood
Baseline Empirical Methodology

(1) Without Scene Fixed Effects
\[ y_j = X_j' \beta + \Delta_j \Lambda + \varepsilon_j \]
poverty rate in GN j  satellite features  GN controls  error

(2) With Scene Fixed Effects
\[ y_j = X_j' \beta + \Delta_j \Lambda + Z_s' \theta + \delta_j \]
poverty rate in GN j  satellite features  GN controls  imagery/scene FEs  error
### OLS Results, National Models (Object Features)

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

Dependent variable is log of GN Poverty Rate Defined at X% of national consumption
## OLS Results, National Models (con’t, Texture Features)

<table>
<thead>
<tr>
<th>Variable</th>
<th>10% Poverty Rate</th>
<th>40% Poverty Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b</td>
<td>t</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.062*</td>
<td>[2.01]</td>
</tr>
<tr>
<td>Pantex (human settlements) mean contrast</td>
<td>0.022</td>
<td>[1.78]</td>
</tr>
<tr>
<td>Histogram of oriented gradients (HOG)</td>
<td>-0.000016*</td>
<td>[-2.12]</td>
</tr>
<tr>
<td>Local Binary Pattern (moments) skewness</td>
<td>-0.00032</td>
<td>[-0.72]</td>
</tr>
<tr>
<td>Line support region mean - scale 8</td>
<td>-0.33</td>
<td>[-1.27]</td>
</tr>
<tr>
<td>Gabor filter mean - scale 64</td>
<td>0.070</td>
<td>[1.60]</td>
</tr>
<tr>
<td>Fourier transform std. dev. - scale 32</td>
<td>0.0034</td>
<td>[1.60]</td>
</tr>
<tr>
<td>Surf - scale 16</td>
<td>-0.00013</td>
<td>[-1.44]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.021</td>
<td>[-0.19]</td>
</tr>
<tr>
<td>Obs</td>
<td>1244</td>
<td></td>
</tr>
<tr>
<td>R Squared</td>
<td>0.39</td>
<td></td>
</tr>
</tbody>
</table>

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

10% National Income Poverty Rate

- Predicted GN Poverty Rate
- True GN Poverty Rate

Size scaled by population

○ GNs
- Best Fit Line

WORLD BANK GROUP
Predicted Versus True Plots – 40% National Income

40% National Income Poverty Rate

Predicted GN Poverty Rate

True GN Poverty Rate

Size scaled by population
## Shapley Decomposition of Share of Variance Explained

<table>
<thead>
<tr>
<th></th>
<th>Avg. Consumption in GN</th>
<th>10% poverty rate</th>
<th>20% poverty rate</th>
<th>30% poverty rate</th>
<th>40% poverty rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>8.6</td>
<td>2.5</td>
<td>3.6</td>
<td>4.6</td>
<td>5.8</td>
</tr>
<tr>
<td>Log of GN Area</td>
<td>6.7</td>
<td>7.9</td>
<td>7.7</td>
<td>7.5</td>
<td>7.3</td>
</tr>
<tr>
<td>Road variables</td>
<td>11.7</td>
<td>12.1</td>
<td>12.2</td>
<td>11.8</td>
<td>11.5</td>
</tr>
<tr>
<td>Building density</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variables</td>
<td>36.4</td>
<td>37.8</td>
<td>36.9</td>
<td>36.4</td>
<td>36.1</td>
</tr>
<tr>
<td>Of which: Built-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>area</td>
<td>18.6</td>
<td>12.8</td>
<td>11.7</td>
<td>12.4</td>
<td>13.3</td>
</tr>
<tr>
<td>Log Number of roofs</td>
<td>8.8</td>
<td>9.4</td>
<td>10.8</td>
<td>10.4</td>
<td>10.0</td>
</tr>
<tr>
<td>Shadow (building</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>index)</td>
<td>5.2</td>
<td>3.8</td>
<td>4.0</td>
<td>4.3</td>
<td>4.6</td>
</tr>
<tr>
<td>NDVI</td>
<td>3.8</td>
<td>11.8</td>
<td>10.4</td>
<td>9.3</td>
<td>8.3</td>
</tr>
<tr>
<td>Roofs</td>
<td>9.3</td>
<td>7.6</td>
<td>7.5</td>
<td>7.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Cars</td>
<td>5.2</td>
<td>4.5</td>
<td>4.4</td>
<td>4.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Agricultural land</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variables</td>
<td>4.9</td>
<td>6.3</td>
<td>7.3</td>
<td>7.5</td>
<td>7.3</td>
</tr>
<tr>
<td>Texture variables</td>
<td>17.3</td>
<td>21.4</td>
<td>20.4</td>
<td>20.5</td>
<td>20.0</td>
</tr>
<tr>
<td>Total r²</td>
<td><strong>0.64</strong></td>
<td><strong>0.39</strong></td>
<td><strong>0.50</strong></td>
<td><strong>0.55</strong></td>
<td><strong>0.59</strong></td>
</tr>
</tbody>
</table>
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