Chapter 1

Estimating Regional Inequality and Growth Using Satellite Data

Abstract

A new panel data set on per-capita economic activity and inequality is developed for India. Population density is estimated using a random forest algorithm that is trained on district-level, remotely sensed data. The algorithm is then used to predict population at a 1×1 kilometer resolution for the entire country. These predictions are then combined with luminosity data to obtain an estimate of per-capita economic activity for each square kilometer. I find a negative and significant Kuznets relationship between per-capita activity and inequality, but a positive relationship between district-level inequality and per-capita growth.

1.1 Introduction

Inequality has received considerable recent attention in the economics literature. Led by the efforts of Thomas Piketty and Emmanuel Saez, datasets that describe the distribution of wealth and income have been developed for much of the developed world. These data have illuminated recent and long-term trends in inequality dynamics, and renewed interest in the effects of income distribution on real economic outcomes.

Despite this renewed focus, however, very little can be said about inequality in the developing world. Survey and administrative data are often sparse and unreliable, making both time-series and cross-sectional analyses difficult to undertake. A recent effort by Alvaredo et al. (2017) seeks to fill this gap by compiling administrative data from a variety of nations, with a particular emphasis on the United States, France, the United Kingdom, and China. However, the authors "stress the need for more democratic transparency on income and wealth dynamics and better access to administrative and financial data" (ibid). The reliance on official statistics severely constrains the set of questions that can be answered in the developing world and inequality research as a whole.

As part of this research I construct a panel dataset on the distribution of economic activity for India. Using results from the geography and machine-learning literatures, I develop a methodology that, in principle, allows for annual estimates of per-capita income from 1992 to the present. This methodology uses open-source and freely available data from multiple satellite sources, and is easily adaptable to any region on the Earth's surface for which census population counts are available. The resulting dataset consists of annual, gridded estimates of per-capita economic activity at a 1×1-kilometer resolution.

I use these data to construct district-level statistics of per-capita economic activity and inequality in India. I then estimate the empirical relationships between per-capita growth, activity, and inequality. I find that inequality is negatively associated with per-capita activity, but it has a strong positive relationship with growth. Districts with higher inequality of economic activity tend to grow

faster than more-equal districts, largely because of their greater increase in rural economic activity. These findings are consistent with a rural convergence hypothesis, similar to the mechanism described by Kuznets (1955).

1.2 Data

The algorithm that develop to estimate population is trained using population counts from India's decennial census. District-level counts for the years 2001 and 2011 and shapefiles for district-boundaries are used to construct chloropleth population maps for each year.[1]

There are four sources of remotely sensed data. Reflectance data are obtained from Landsat 5 and Landsat 7. Each satellite observes the reflectance of six spectral bands at 30×30-meter resolution—three bands in the visible spectrum (red, blue, and green light) and three bands in the infrared spectrum. Before it is released, the USGS processes the data into orthorectified, top-of-atmosphere reflectance. These corrections account for differences in terrain, angle of the sun, satellite altitude, and sensor degradation. Every available scene covering the sample area is collected for each of the census years. Pixels covered by clouds are discarded and a "greenest-pixel" composite is produced on a pixel-by-pixel basis for each of the census years.[2] The resulting dataset is a collection of six raster layers for each census year, one for every spectral band of the Landsat data.

Pixels that cover permanent bodies of water are removed the Landsat data. Pekel et al. (2016) describe a process for identifying surface water from satellite imagery. The authors compile raster datasets for each of the census years and are made available to the public at a 30×30-meter resolution. In addition, a binary raster layer is created that codes each pixel as equal to one if it is within a kilometer of a permanent body of water. This water-proximity raster captures the tendency of

[1]The island states of Lakshwadeep and Andaman & Nicobar as well as the disputed region of Jammu & Kashmir are dropped from the sample. Combined, these areas contain less than 1% of the total population.

[2]The greenest pixel is determined by first calculating the normalized difference vegetation index (NDVI) for each pixel, then choosing the pixel from the scene with the highest NDVI.

human settlements to locate near sources of water.

Elevation data are obtained from the Shuttle Radar Topography Mission which flew aboard the Space Shuttle Endeavour in the year 2000. Two radar antenna scanned most of the earth's surface to obtain average elevation estimates at a 30×30-meter resolution.

The final source of satellite data is the Defense Meteorological Satellite Program (DMSP), a set of satellites that collectively observe night-light luminosity at a 1000×1000-meter resolution. A well-known drawback of the DMSP data is the high incidence of sensor saturation that occurs in brightly lit areas. The brightness of many large urban areas is therefore censored. To account for this, I utilize a transformed version of the DMSP data compiled by Hsu et al. (2015), who use data from multiple DMSP sensors and sensor settings to obtain luminosity measurements that have a larger dynamic range than the traditionally used DMSP data. In addition to solving the saturation issue, these data are also much more sensitive to low-light areas.

For each year, the final satellite data are then combined into a single raster file covering the geospatial extent of India. Each raster file contains nine layers: six reflective bands from the Landsat data, the proximity-to-water band, elevation, and the luminosity band.

1.3 Night Lights and Economic Activity

The close relationship between GDP and night-light luminosity (as measured by the DMSP satellites) is established by Henderson, Storeygard, and Weil (2012). Using the raw, top-coded DMSP data, they find a near log-linear relationship between GDP and luminosity in a panel of countries since 1992, when the DMSP program began. They are unable to identify the source of the non-linearity, but it is likely that top-coding and sensor degradation are partially responsible. The sensitivity of the sensors change tend to change as a result of this natural degradation process, adding a source of non-random noise to the data.

Figure 1.1 highlights the implications of sensor degradation. Aggregate luminosity for low and

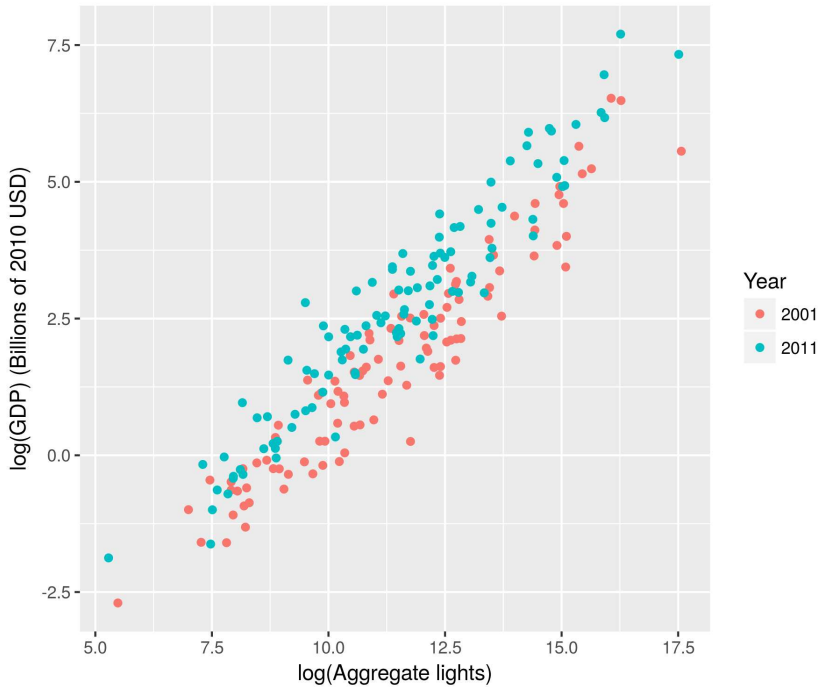


Figure 1.1: The Time-Varying Relationship Between Luminosity and GDP

middle-income countries is, as a group, linearly related to GDP, but the relationship appears to be different between years.[3] As the sensors degrade, a given level of aggregate light might corresponds to a higher level of GDP in 2011 than in 2001. Additional light activity appears to have a greater effect on GDP in 2011, reflecting changes in the sensitivity of the sensors.

I address this issue of sensor degradation by calibrating the luminosity data with GDP. I first estimate the equation

$$\log\left(GDP_i^y\right) = \beta_0^y + \beta_1^y \log\left(AggLuminosity_i^y\right) + \varepsilon_i^y, \tag{1.1}$$

for all low and middle-income countries for which the World Bank reports GDP (Table 1.1). The regression is repeated for each census year ($y = 2001, 2011$) and parameter estimates $\hat{\beta}_0^y$ and $\hat{\beta}_1^y$ are obtained. I then transform each pixel of luminosity data using the estimated parameters. For

[3] I define "low and middle-income countries" as having a per capita GDP of less than $20,000, measured in 2010 USD.

Table 1.1: Cross-sectional Country Regressions of GDP on Aggregate Luminosity

	Dependent variable:			
	log(GDP)			
	2001		2011	
	(1)	(2)	(3)	(4)
Intercept	-6.844^{***}	-5.642^{***}	-6.589^{***}	-6.238^{***}
	(0.284)	(1.085)	(0.287)	(1.090)
log(Aggregate lights)	0.753^{***}	0.535^{***}	0.809^{***}	0.745^{***}
	(0.024)	(0.191)	(0.024)	(0.192)
log(Aggregate lights)^2		0.009		0.003
		(0.008)		(0.008)
R^2	0.904	0.905	0.914	0.914
Note:				*p<0.1; **p<0.05; ***p<0.01

pixel p in year y, calibrated luminosity data is:

$$\ell_p^y = exp\left(\hat{\beta}_0^y + \hat{\beta}_1^y \log(Luminosity_p^y)\right) \tag{1.2}$$

$$= \widehat{GDP}_p^y. \tag{1.3}$$

Parameter estimates are given in Table 1.1. Columns 2 and 4 include a quadratic term to test for the presence of nonlinearity. The quadratic term is not significant for either of the census years. The parameter estimates in columns 1 and 3 are therefore used to calibrate the luminosity data. For both years, over 90% of the variation in estimated GDP is explained by variation in observed nighttime lights.

At the pixel level, observed luminosity may be a function of some process other than economic activity. For instance, it may be the case that government efforts to expand access to electricity to

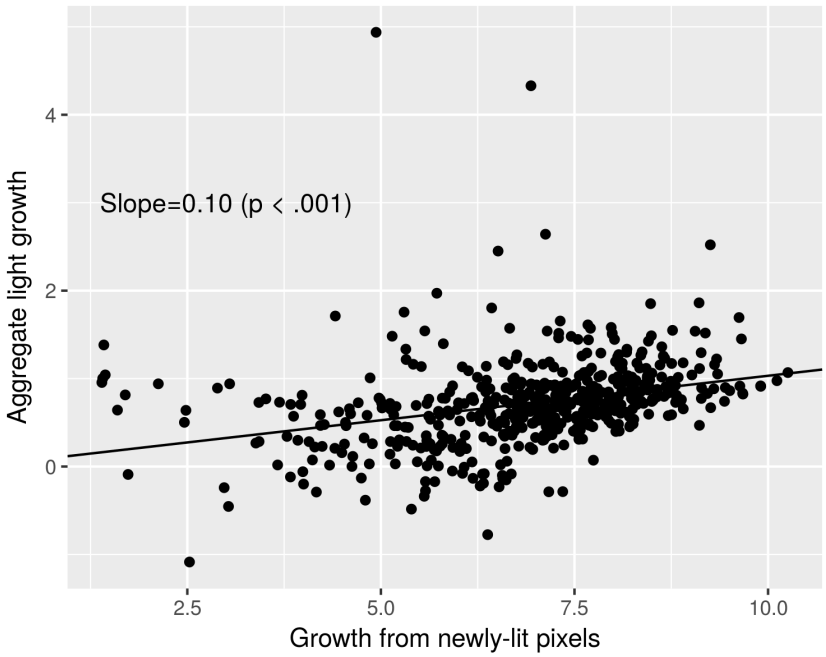


Figure 1.2: Newly-Lit Pixels are a Small Percentage of Aggregate District Light Growth

rural areas may create luminosity without increasing production or income. People living in rural areas without access to electricity may substitute electricity consumption (and therefore luminosity) with other goods available within their feasible consumption sets, such as additional household items that are not luminous.

Figure 1.2 shows the relationship between aggregate night-light growth and the light growth in previously unlit pixels at the district level. As expected, the relationship is positive—an increase in newly lit pixel luminosity is strongly associated with an increase in aggregate light growth. However, the majority of aggregate light growth occurs in already-lit pixels; newly lit pixels account for just 10% of aggregate light growth.

1.4 Population Estimation

The population estimation procedure proceeds in four steps:

1. Calculate summary statistics for the satellite data at the district level.

2. Use the district-level summary statistics to estimate a functional relationship between district-level population and the satellite data.

3. Calculate summary statistics for the satellite data for a five-kilometer circle around each 30×30-meter pixel, then resample this raster layer at 1000×1000-meter resolution.

4. Use the estimated function in Step 2 (above) to estimate population for each 1000×1000-meter pixel.

This strategy is similar to the methodology employed by Stevens et al. (2015), who use non-time-varying, proprietary data to estimate population cross-sections in three small countries. In contrast, I utilize open-source and freely available data that is available periodically since 1992, allowing for years to be compared with comparable data.

For each census year, I calculate the mean and variance for each of the eight satellite bands at the district level. In addition, I calculate the correlation between each pair of Landsat bands, for a total of 44 covariates. The (log) population density of each district is then described by the equation

$$y_{dt} = f(S_{dt}) + \varepsilon_{dt} \tag{1.4}$$

where y_{dt} is the log population density in district d in year t and S_{dt} is the vector of 44 satellite summary statistics. Due to district realignment between census years, the panel is unbalanced (there are 576 districts in 2001 and 614 in 2011).[4]

There is no theory that informs the functional form of f, though it is likely to be highly non-linear with high-level interactions between covariates. I therefore estimate f with a random forest

[4]The primary cause of the difference in the number of districts between 2001 and 2011 is due to larger districts being split into smaller districts. It is common in the literature to simply recombine the split districts, but doing so in this context decreases the number of observations with which to train the algorithm. For the main regression results, I use the 2001 district boundaries for all specfications.

algorithm (Breiman 2001). The algorithm assumes high-level interactions and non-linearity of covariates, and offers substantial improvements to predictive accuracy over least-squares estimation (Mullainathan and Spiess 2017).

1.4.1 Random Forests

A random forest is a collection of regression trees. The purpose of a regression tree is to partition continuous data according to a predefined decision rule. A regression tree is "grown" as follows:

1. Choose a covariate and partition the data into two subsets. Calculate the variance of the dependent variable for each subset.

2. Choose the covariate and observation that minimizes the sum of the variances of the two partitions.

3. For each, repeat steps 1 and 2 until every partition has one observation.

Each partition serves as a decision rule in the tree. Predictions are then performed by sending new observations "down" the tree, starting from the first decision rule and ending when there are no more decision rules, then assigning a prediction based on the value of the dependent variable in the final partition.

A single regression tree overfits the data, as every predicted value is identical to some observation in the original dataset. It is therefore useful to randomize the way in which trees are grown and collect multiple trees into a forest. Predicted values therefore become the average predicted value of each tree in the forest. Each tree in a random forest is randomized along two dimensions to avoid overfitting. First, each tree is grown on a bootstrapped sample of the original data. Second, a random subset of the covariates is used to create the decision rule at each partition.[5] These procedures introduce a small amount of bias into each tree, so overfitting is prevented.

[5]The size of the random subset is a model parameter chosen by the researcher. Breiman (2001) suggest a value of $K/3$, which is what I utilize here. None of the following results are sensitive to changes in the size of this subset.

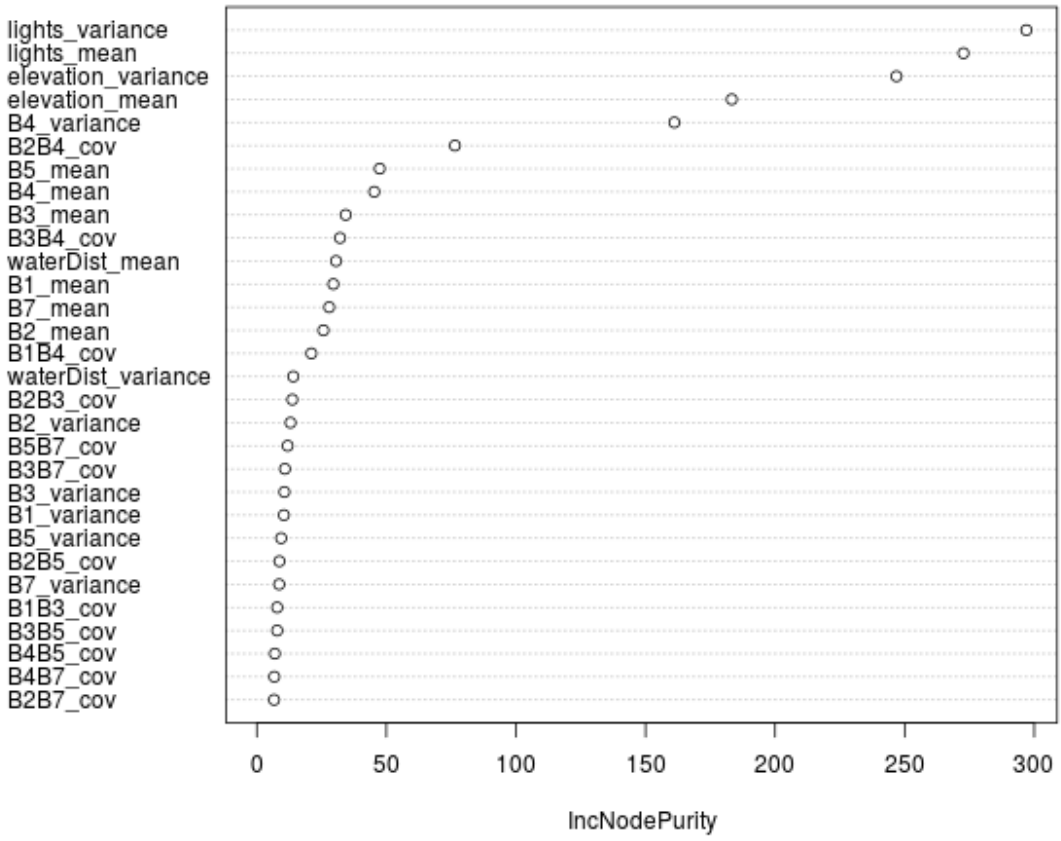


Figure 1.3: Random Forest Variable Importance (30 Most Important Covariates)

1.4.2 Predicting Population

I grow 1,000 trees, with a random subset of 11 covariates searched at each partition. The fitted forest explains 91% of the variance in district-level log population density, with a mean squared residual of 0.13. Although estimates of the marginal effects of the covariates cannot be recovered from the model, rough measures of variable importance can be recovered. Figure 1.3 ranks the covariates in terms of their "node purity," which measures the average amount of variance that is reduced when that variable is selected at any partition, in any tree. The mean and variance of night lights (*lights_mean* and *lights_variance*) are most important, followed by elevation mean and

variance, then various reflective bands from the Landsat satellites. However, covariates with low node purity scores may have substantial contributions through their interaction with other variables.

Population prediction proceeds in 2 steps:

1. Create a 5km-radius circle (approximately 28,000 pixels) around each pixel in the data raster layer, and calculate summary statistics within the circle.

2. Resample the resulting raster layer to a 1000×1000-meter (one kilometer) resolution.

Step 1 creates a new raster where each 30×30-meter pixel contains the summary statistics (mean, variance, correlation) of the satellite data within a 5km circle of the original pixel. This process allows for spatial dependence between pixels—the characteristics of neighboring pixels are allowed to influence the population estimates at each pixel. Each layer of the raster stack corresponds to one of the 44 summary statistics contained in S. The resampling procedure in Step 2 allows for the population estimates to be directly compared to the night-light data, which is available at this same resolution.

For each pixel p, the log population density is estimated using the random forest model trained on the district-level data:

$$\hat{y}_{pt} = \hat{f}(S_{pt}), \tag{1.5}$$

where $\hat{f}$ is the fitted random forest model and S_{pt} corresponds to the p^{th} pixel of the summary statistic raster S. For each year, a single-layered raster file $\hat{y}_t$ is created which maps predicted log population density at a 1000×1000 meter resolution (Figure 1.4).

1.4.3 Robustness

The accuracy of the population estimates can be checked by aggregating the fitted population estimates to district level, then comparing these estimates with known district populations from the

(a) 2001 (b) 2011

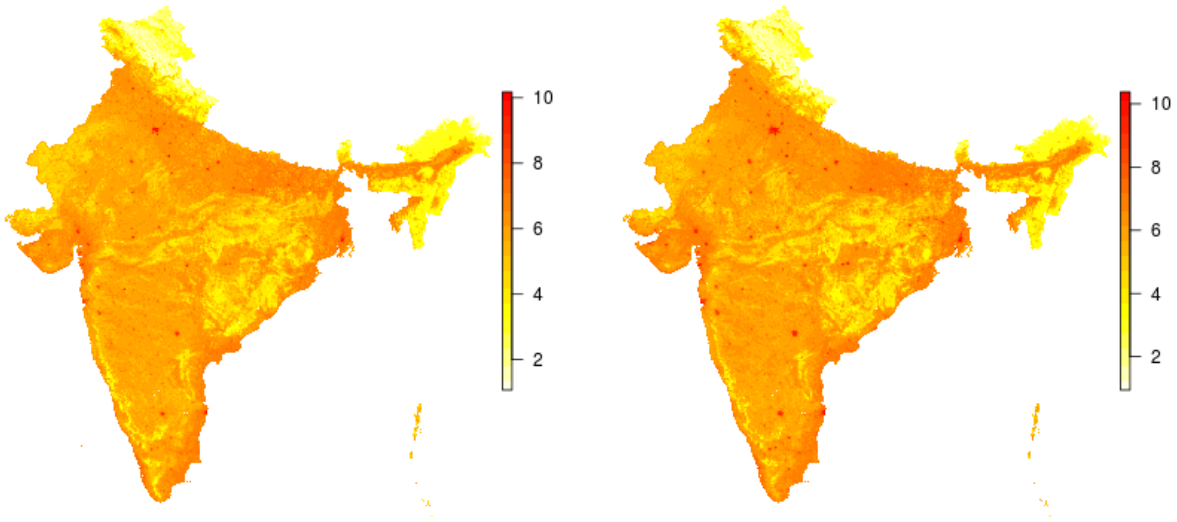


Figure 1.4: Log Population Density Estimates

(a) 2001 (b) 2011

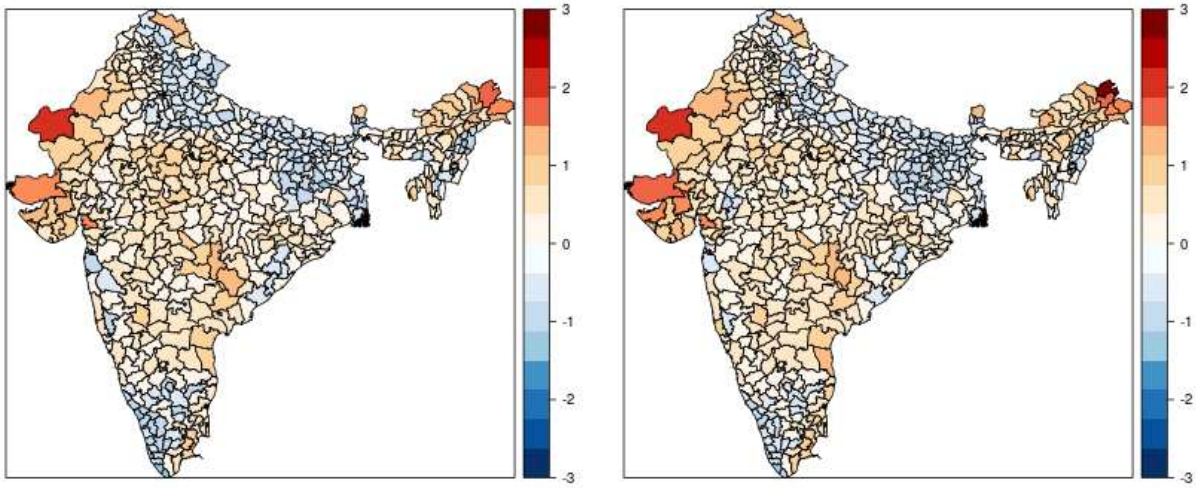


Figure 1.5: Log Density Prediction Error

census data. Figure 1.5 shows the district log-density prediction error for each year. The districts with the largest errors are concentrated along the northern, western, and eastern borders. Importantly, the districts with the largest errors are among the least populated. These errors are therefore unlikely to contaminate district-level regressions, especially when regressions are weighted by population.

The accuracy of the predicted pixel-level densities can be described by estimating the equation

$$\log\left(y_{dt}\right) = \beta_0 + \beta_1 \log\left(\hat{y}_{dt}\right) + \varepsilon_{dt}, \tag{1.6}$$

where $\hat{y}_{dt}$ is the aggregated estimate of log density for district d in year t. Results are shown in Table 1.2. If $\hat{y}_{dt}$ is an unbiased estimate of y_{dt}, the coefficient estimate $\hat{\beta}_1$ should be statistically indistinguishable from one. In other words, the density elasticity of predicted density should be equal to one—a $x\%$ change in predicted density should correspond to an $x\%$ change in actual density.

Column 1 shows the results for the pooled data from both census years, while columns 2 and 3 show the estimates for each census year separately. The R^2 for these regressions implies that more than 82% of the variance in district-level log population density is explained by the 1000×1000-meter density estimates. However, the coefficient estimates are significantly different from one, indicating that $\hat{y}_{dt}$ may be biased. To uncover the source of this bias, the models in columns 4, 5, and 6 reestimate equation 1.6 after dropping the districts in the lowest 5% of population density. In all but one of the specifications, $\hat{\beta}_1$ becomes statistically indistinguishable from one. The suggests that the source of bias is low-density districts.

Closer inspection of the map in Figure 1.5 and Table 1.2 reveals that the largest prediction errors occur due to an overestimation of population in low-density districts. This is a direct result of a primary shortcoming of the random forest estimator: the algorithm cannot predict values outside the observed range of the dependent variable. The lowest population-density district therefore

Table 1.2: Regressing Actual Population Density on Predicted Density

	Dependent variable: Log census population density					
	All Districts			Highest 95% Density Districts		
	(1)	(2)	(3)	(4)	(5)	(6)
Log predicted density	1.093	1.106	1.089	1.016	1.033	1.008
	(0.014)	(0.021)	(0.019)	(0.016)	(0.024)	(0.020)
p-value (H_0 : slope $= 1$)	<0.001	<0.001	<0.001	0.159	0.085	0.345
Year	All	2001	2011	All	2001	2011
Observations	1,190	576	614	1,130	547	583
R^2	0.835	0.827	0.845	0.788	0.775	0.807

serves as a lower bound on predicted per-pixel population density. Over large, low-density areas, the algorithm will consistently predict pixel-level densities greater than the actual densities, resulting in an upward bias for these areas. On an aggregate level, this bias is only apparent when considering districts in the lowest 5% of population density (see Table 1.2). These districts contain approximately 0.5% of the total population of India, which makes it unlikely that this bias will contaminate the results that follow.

Figure 1.6 plots the actual population density against the predicted population density for the entire sample, along with a 45-degree line. The horizontal, dotted line is drawn at the fifth percentile (observations below the dotted line are dropped in columns 4-6 of Table 1.2). As expected, the districts that lie below the dotted line are also below the 45-degree line, indicating that the predicted density is systematically higher than the actual density for these districts. However, above the dotted line, no systematic bias appears to exist. Furthermore, observations immediately above the dotted line do not appear to have any systematic deviation, implying that the algorithm becomes unbiased near the fifth percentile.

In addition to over predicting population in low-density pixels, the algorithm will symmetrically under predict population in high-density pixels. Inspecting the upper-right corner of figure

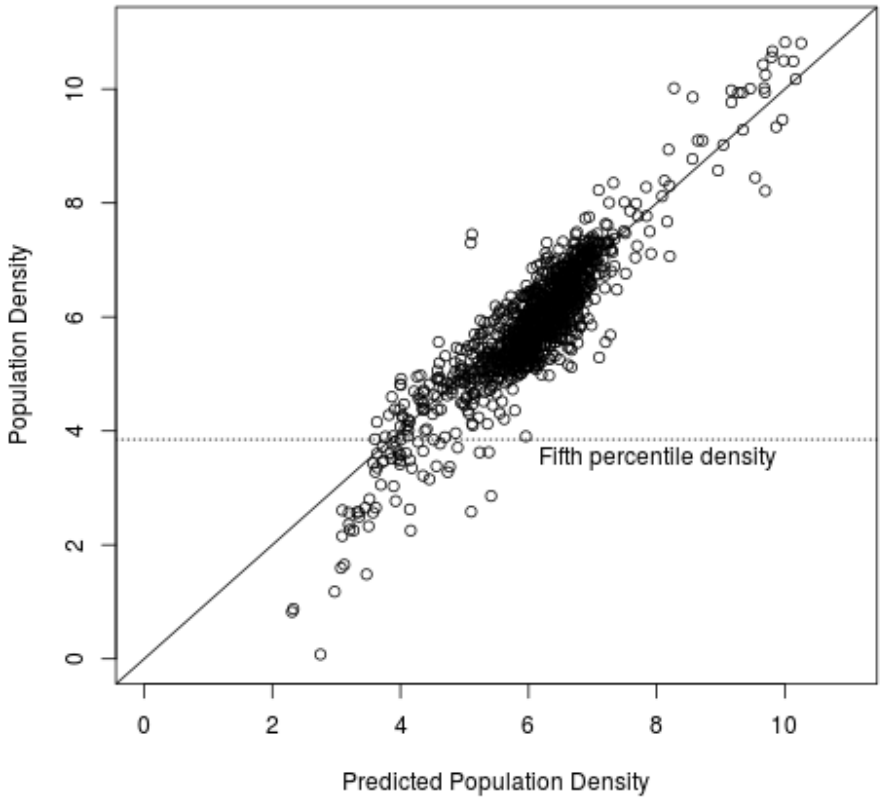


Figure 1.6: Predicted vs Actual Population Density

1.6 suggests that this problem does not affect district-level predictions to the extent that the over prediction problem does. Though the highest-density districts all lie above the 45-degree line, the deviation is minor when compared to the lowest-density districts, and similar in magnitude to the prediction errors that are common in the bulk of the distribution.

Although each regression tree is grown on a bootstrapped sample of all observations, it is possible to withhold certain observations for cross-validation purposes. A natural subsample with which to perform cross validation is the observations from a specific census year. In Table 1.3, I report the results of a regression of each year's predicted density on the actual density, where the predicted densities are the result of fitting the data on the other, out-of-sample year. Columns 1

Table 1.3: Out-of-Sample Predictive Accuracy

	Dependent variable: Log census population density			
	All Districts		Highest 95% Density Districts	
	(1)	(2)	(3)	(4)
Log predicted density	1.204	1.177	1.063	1.039
	(0.033)	(0.028)	(0.034)	(0.028)
p-value (H_0 : slope $= 1$)	<0.001	<0.001	0.032	0.082
Year	2001	2011	2001	2011
Observations	576	614	547	583
R^2	0.697	0.745	0.642	0.706

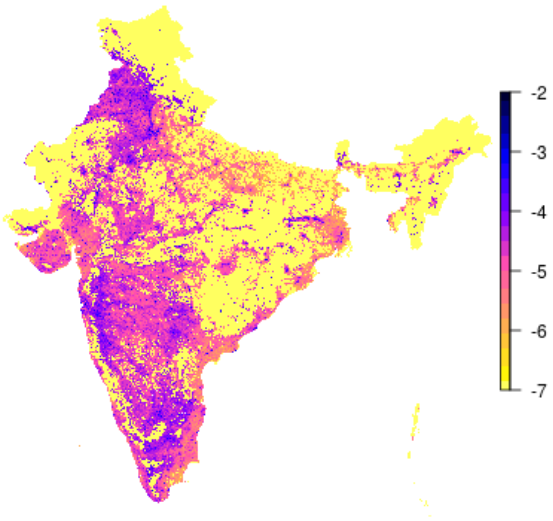
and 3 are the results of using 2011 data to predict 2001 density, and vice versa for columns 2 and 4. As in Table 1.2, the regressions are performed on a subsample of the data corresponding to the districts above the fifth percentile of population density for that year.

The results in Table 1.3 follow a similar pattern to those reported in Table 1.2. The elasticity of actual density to predicted density is greater than one in the complete sample, but becomes insignificantly different (at the 1% level) from unity when the lowest-density districts are dropped. The estimation procedure therefore retains a high degree of predictive accuracy even when predicting densities a decade removed from the data used to train the algorithm.

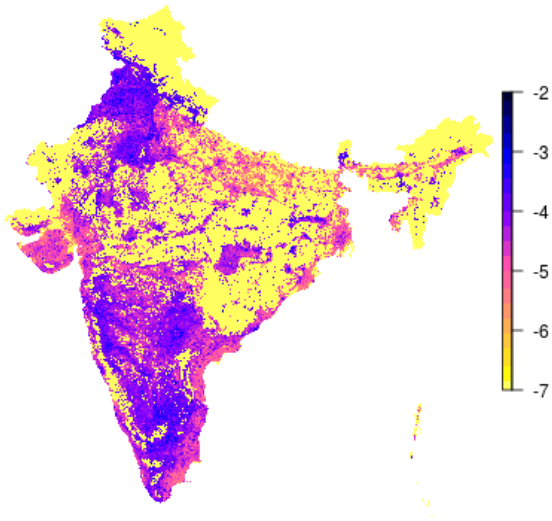
1.5 Results

1.5.1 Economic Activity

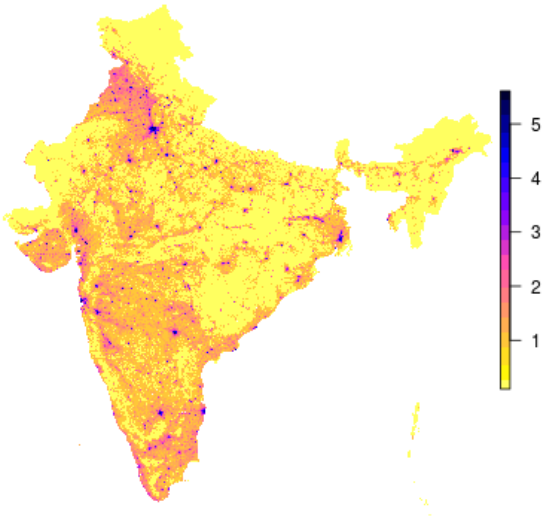
Both the calibrated night-light data and the population density estimates are available at the 1000×1000-meter resolution. By combining these two sources, a disaggregated estimate of per-capita



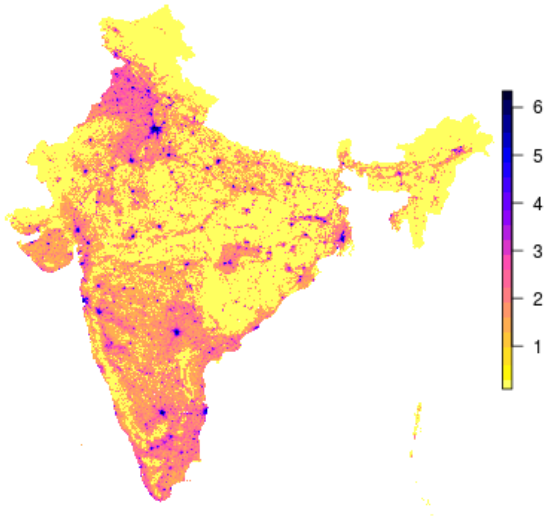
(a) Per-capita Lights, 2001



(b) Per-capita Lights, 2011



(c) Aggregate Lights, 2001



(d) Aggregate Lights, 2011

Figure 1.7: Log Predicted Per-capita and Aggregate Lights

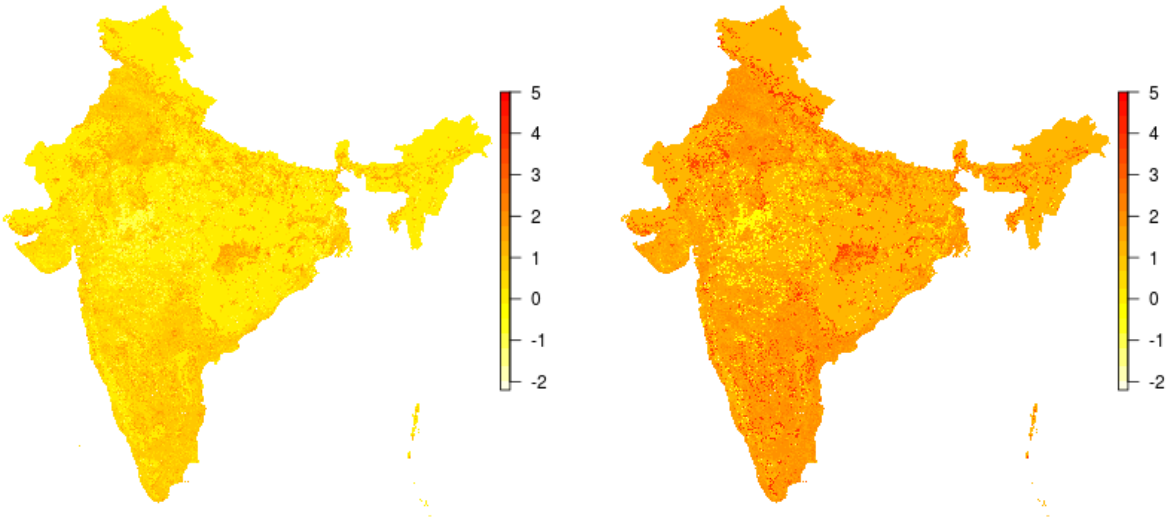
economic activity can be obtained. Denoting the calibrated lights observed in pixel p by ℓ_{pt}, the

per-pixel estimate of per-capita economic activity is

$$\hat{A}_{pt} = \frac{\ell_{pt}}{\exp\left(\hat{y}_{pt}\right)} \tag{1.7}$$

where $\hat{y}_{pt}$ is the estimated log-population density in pixel p.

Figures 1.7a and 1.7b show the estimated log per-capita economic activity $\log\left(\hat{A}_{pt}\right)$. Compared to aggregate lights (Figures 1.7c and 1.7d), the per-capita estimates show a much more diffuse distribution of economic activity. The raw light data most readily identifies large population centers, while the per-capita activity data show large contributions to overall economic activity from relatively sparsely populated areas, such as the inland areas of Maharashtra and Karnataka in the southeastern section of the country. In addition, urban centers such as Delhi and Bangalore are more difficult to identify in the per-capita data, reflecting the economic contributions or suburban and rural areas that are absent from the aggregate data.



(a) Aggregate Pixel Level Light Growth (b) Per-capita Pixel Level Light Growth

Figure 1.8: Aggregate and Per-capita Pixel Level Log Light Growth

Note: Per-capita growth is rescaled to facilitate comparison with aggregate light growth.

Figure 1.8 shows the growth in per-capita log lights and aggregate log lights over the sample period. Again, the per-capita estimates illustrate a higher degree of heterogeneity than the aggregate lights, reflecting population changes as well as changes in overall economic activity.

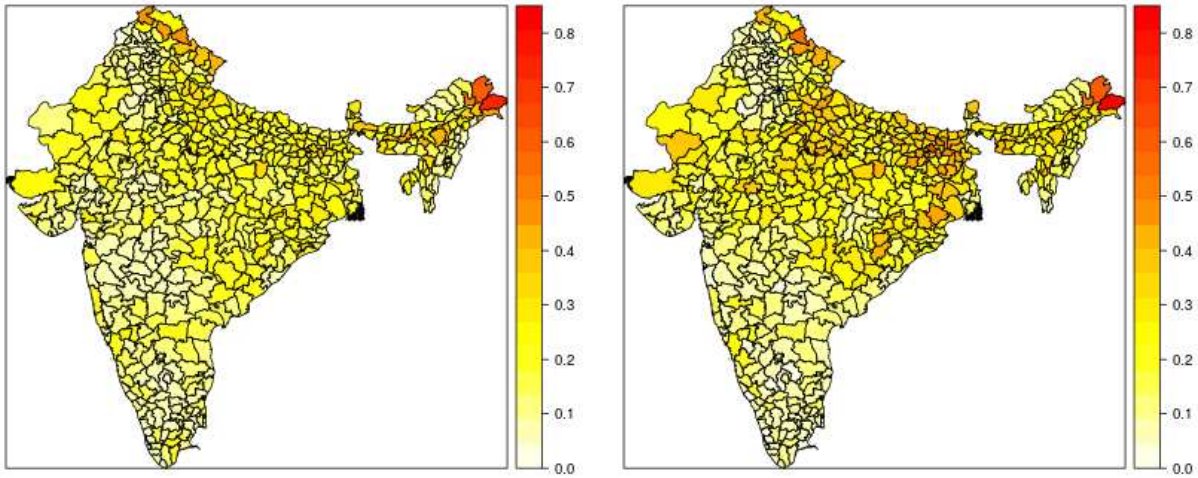
1.5.2 Inequality

The disaggregated, pixel-level measure of per-capita economic activity allows for comparisons across pixels. For instance, the distribution of income, and therefore inequality measures, can be calculated at the district level based on the distribution of income across pixels within the district. However, the spatial resolution of the population and economic activity estimates (one square kilometer) place a lower bound on the level of fineness of the per-capita estimates. Inequality measures are therefore subject to the caveat that within-pixel inequality cannot be calculated. There are two available methods for dealing with this issue. First, one can simply assume that within-pixel economic activity is homogeneous. Alternatively, an inequality measure can be selected that is agnostic toward within-pixel distributions.

Subgroup Decomposability Many inequality indices have the property of *subgroup decomposability*, which allows inequality to be separated by within-group and between-group inequality.[6] Formally, let $x = (x_1, x_2, \ldots, x_N)$ be the vector of economic activity for each agent $1, \ldots, N$. An inequality index is a function that maps from the distribution to a real number, $I : \mathbb{R}^N \to \mathbb{R}$. If the distribution x is partitioned into $K < N$ subgroups, the index I is said to be subgroup decomposable if:

$$I(x) = \underbrace{\sum_{k=1}^{K} \alpha_k I(x_k)}_{\text{within-group}} + \underbrace{I(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)}_{\text{between-group}}, \tag{1.8}$$

[6]For a detailed description of subgroup additivity, see Shorrocks (1984).



(a) 2001 (b) 2011

Figure 1.9: Between-Pixel Light Inequality

where $\bar{x}_k$ is the mean of partition k and α_k is a weight. The between-group inequality can be calculated between pixels, and changes to within pixel inequality will have no effect on the contribution of between pixel inequality to total inequality.

The commonly-used Theil index has the subgroup decomposability property. Formally,

$$T(x) = \sum_{i=1}^{N} \frac{x_i}{X} \ln\left(\frac{Nx_i}{X}\right) \tag{1.9}$$

where $X = \sum_{i=1}^{N} x_i$. This can be decomposed as

$$T(x) = \underbrace{\sum_{k=1}^{K} \frac{x_k}{X}\, T(x_k)}_{\text{within-group}} + \underbrace{\sum_{k=1}^{K} \frac{x_k}{X} \ln\left(\frac{\frac{x_k}{X}}{\frac{N_k}{N}}\right)}_{\text{between-group}}. \tag{1.10}$$

Between-pixel inequality of economic activity can therefore be calculated as

$$T_B(\ell) = \sum_{p=1}^{P} \frac{\ell_p}{L} \ln\left(\frac{\ell_p/L}{\exp\left(\hat{y}_p\right)/\hat{Y}}\right), \tag{1.11}$$

where $\ell = \ell_1, \ldots, \ell_P$ are the calibrated lights of each pixel, $L = \sum_{p=1}^{P} \ell_p$, and $\hat{Y} = \sum_{p=1}^{P} \exp\left(\hat{y}_p\right)$.

Figure 1.9 shows this inequality measure calculated at the district level for each of the census years. Two trends are immediately apparent. First, inequality has increased in most of the districts over the sample period. Second, inequality tends to be greater in the Ganges basin (the north and north-east portion of the country).

1.5.3 Inequality and Growth

The Kuznets relationship between growth and inequality has been explored frequently in the literature. Using the Deininger and Squire (1996) survey-based collection of national inequality data, Barro (2000) finds a negative relationship between inequality and growth for developing countries, but a positive relationship in developed countries. More recently, Ostry, Berg, and Tsangarides (2014) find a significant negative relationship in a panel of countries in a variety of specifications and time frames.

Despite these findings, very little known about the relationship between intra-regional inequality and growth, especially in the developing world. This is primarily owing to the paucity of data, both on regional growth rates and inequality. Recent studies have used luminosity data to proxy for income and growth (Donaldson and Storeygard 2016), but there has been no effort to disaggregate remotely sensed data on economic activity based on population.

Figure 1.10 shows the cross-sectional relationship between estimated inequality and the level of economic activity (as measured by per-capita luminosity) at the district level. The relationship is striking—more-equal districts have substantially higher levels of per-capita light output. There is a possibility that the spatial resolution of the per-capita activity estimate can bias district-level esti-

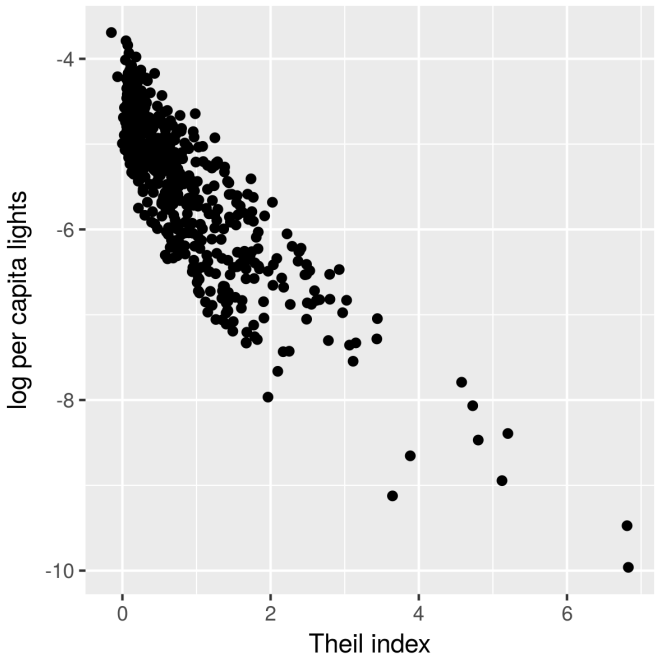


Figure 1.10: Kuznets Relationship for the Districts of India, 2001

mates of inequality in a systematic fashion. In particular, the geospatial size of districts is a function of population—heavily urbanized areas tend to be located in smaller districts than sparsely populated, rural areas. Smaller, more populated districts may therefore have less estimated inequality, since fewer pixels are used in the inequality calculation.

To account for this possibility, I run a simple cross-sectional regression of aggregate lights on inequality at the district level (Table 1.4). Column 1 corresponds to the line of best fit in Figure 1.10, and shows a negative and highly significant relationship between aggregate lights and inequality.

The algorithm to over predicts population in low-density areas. Furthermore, low-density areas are likely to have lower aggregate light activity than high density areas. It is therefore reasonable to believe that the systematically biased estimates of per-capita economic activity are biasing the Kuznets results. To account for this possibility, I weight the simple Kuznets regression by census population in column 2. This results in a larger (in magnitude) coefficient on inequality. In other words, the negative Kuznets relationship is more pronounced in districts with greater population.

Table 1.4: Kuznets Regressions

	Dependent variable: Aggregate district-level lights (2001)			
	(1)	(2)	(3)	(4)
Theil index	-0.899^{***}	-1.254^{***}	-0.746^{***}	-0.736^{***}
	(0.033)	(0.043)	(0.023)	(0.025)
Urbanization rate				0.173
				(0.225)
People per 100,000km^2				-0.321
				(0.925)
Weights	None	Population	Population^{-1}	Population^{-1}
Observations	575	575	575	575
R^2	0.725	0.705	0.909	0.909
Adjusted R^2	0.725	0.705	0.908	0.909
Note:				*p<0.1; **p<0.05; ***p<0.01

More populated districts also tend to be smaller in area (district size is not exogenous). Smaller districts contain fewer pixels, and each pixel is likely to contain more people. Since the algorithm cannot observe inter-pixel inequality, it may be the case that highly populated districts are mechanically estimated to have lower inequality than sparsely populated districts. In column 3, I account for this possibility by weighting by inverse population. The coefficient on inequality is still estimated to be negative and highly significant, even when heavily discounting the most populated districts.

Column 4 adds additional controls, including the urbanization rate (obtained from the census bureau) and population density (per 100,000 square kilometers). As in the other specifications, the coefficient on inequality is negative and highly significant.

The estimated growth rate of economic activity for each pixel can be obtained by comparing per-capita luminosity estimates across years. Letting L_{dt} denote the aggregate lights in district d

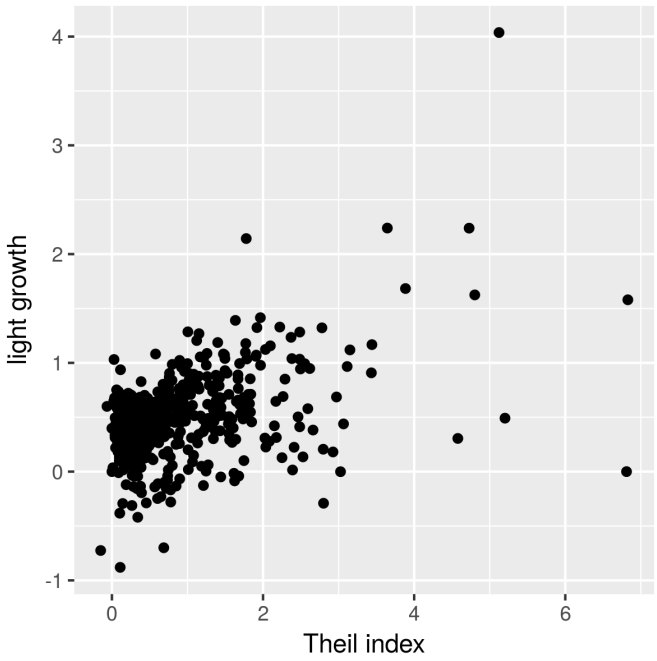


Figure 1.11: The Positive Correlation Between Inequality and Growth

and N_{dt} denote the population of district d, define the per-capita district-level economic activity as

$$\hat{A}_{dt} = \frac{L_{dt}}{N_{dt}} \tag{1.12}$$

Per-capita growth of district d is then $\hat{g}_d = \ln\left(\hat{A}_{dt}/\hat{A}_{dt-1}\right)$. Figure 1.11 shows the relationship between the light-based growth proxy $\hat{g}_d$ and inequality (column 1 of Table 1.5). In contrast to the Kuznets relationship in Figure 1.10, inequality appears to be positively related to growth.

Given the Kuznets relationship in Table 1.4, it is reasonable to expect that a large portion of the positive relationship between growth and inequality is simply due to Solow-convergence of low-income districts. To test this, I estimate a standard growth regression while controlling for inequality (Barro 1991). In the standard neoclassical growth model, the growth rate can be

expressed as

$$\hat{g}_d = \alpha - \left(1 - e^{-\gamma}\right) \ln\left(\hat{A}_{d0}\right) + u_d \tag{1.13}$$

where α is the steady-state growth rate, $\hat{A}_{d0}$ is the initial per-capita economic activity in district d, and γ is the convergence rate. I therefore estimate the equation

$$\hat{g}_d = \beta_0 + \beta_1 \ln\left(\hat{A}_{d0}\right) + \beta_2 T_{d0} + X_{d0}\delta + \varepsilon_d \tag{1.14}$$

In this specification, T_{d0} is initial inequality and X_{d0} is a vector of controls.

Regression results are shown in Table 1.5. Columns 2 and 3 are the baseline regressions, without any controls (column 3 weights by population). In both specifications, the coefficient on initial per-capita lights is negative and significant, providing evidence of neoclassical convergence. The coefficient on the Theil index is also positive and significant, suggesting a positive relationship between inequality and growth. Compared to the estimate in column 1, however, controlling for convergence does mitigate the estimated effect of inequality. Adding additional controls (column 4) has little effect on the estimated effect.

Columns 7 and 8 show the results of equation 1.14 when the sample is limited to just urban districts and just non-urban districts, respectively. Urban districts are defined as those in the top quartile of the distribution of urbanization rates. The estimated coefficient on inequality is similar in magnitude for the urban subsample to that for the complete sample. However, the coefficient estimate on the rural subsample is substantially larger. This suggests that the positive effect of inequality on growth may be driven mostly by non-urban inequality.

A similar result is estimated for the models in columns 5 and 6, which estimate the growth rate of newly-lit pixels (pixels with observed lights in 2011 but emitted no visible light in 2001) and the growth rate observed in already-lit pixels. Light growth in already-lit pixels does not appear to be influenced by district-level inequality. Furthermore, the coefficient on initial lights is positive

Table 1.5: Growth Regressions

	Dependent variable: Light growth								
						Newly lit Pixels	Already lit Pixels	Urban Districts	Rural Districts
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Theil index	0.208***	0.153***	0.133***	0.140***	0.066	0.124	0.043	0.162**	0.222***
	(0.037)	(0.054)	(0.037)	(0.038)	(0.045)	(0.218)	(0.038)	(0.080)	(0.050)
Initial per-capita lights (2001)		−0.059*	−0.075***	−0.057*	−0.150***	−0.211	0.083***	−0.105	0.011
		(0.032)	(0.029)	(0.030)	(0.039)	(0.181)	(0.030)	(0.071)	(0.043)
Urbanization rate (2001)				−0.095	0.091	−0.602	0.445***	−0.258*	−0.007
				(0.076)	(0.083)	(0.633)	(0.091)	(0.132)	(0.241)
People per 100,000km				0.173	−0.064	−80.671***	−0.428	0.434	14.140***
				(0.437)	(0.414)	(20.259)	(0.364)	(0.411)	(5.369)
Weights	None	None	Population	Population	Population	Population	Population	Population	Population
State FE	No	No	No	No	Yes	No	No	No	No
Observations	574	574	574	574	574	525	574	144	430
Adjusted R^2	0.217	0.222	0.192	0.191	0.450	0.098	0.167	0.202	0.168

Note: *p<0.1; **p<0.05; ***p<0.01

and significant, suggesting that the growth rate of already-lit pixels is close to the long-run steady state. Newly-lit pixels, on the other hand, are very responsive to district-level inequality, with a coefficient estimate similar in magnitude to the full sample estimates. The channel through which inequality affects growth therefore appears to be the growth of poor, rural areas.

The simplest explanation for this observation is intra-district convergence. If migration rates are modest, the existence of district-level inequality necessarily implies that some areas within districts have grown faster than others, perhaps due to random growth shocks. If there is a natural tendency for income growth across the distribution to converge to the same rate, poorer areas would be expected to catch up to the richer areas. Given that inter-district convergence is observed in full sample, a similar convergence is likely to be occurring on smaller levels.

This explanation is similar to the mechanism proposed by Kuznets (1955), who speculated that industrialization and urbanization will decrease the supply of rural workers and eventually increasing their wages. Kuznets additionally claimed that industrialization will eventually penetrate to the poorest levels of society, thus giving them the productive tools necessary to catch up.

An alternative mechanism that explains the above observations is policy. Previous growth in top incomes that creates initial inequality may induce demand for redistribution. Policymakers in high-growth districts may be targeting poor, rural communities for investment and electrification. If growth is autoregressive, high growth in previous periods may cause initial inequality, a demand for redistribution, and continued growth. This explanation implies that initial inequality may not be strictly exogenous, and therefore the parameter estimates in Table 1.5 may be biased. Future versions of this paper will address this issue empirically by considering instrumental variable approaches, as is common in the growth literature.

1.6 Future Research

Future versions of this paper will address three issues. First, a series of robustness checks will be performed on the results shown in Table 1.5. Second, poverty estimates can be obtained from the satellite data and utilized in various regression settings. Lastly, the political economy of inequality and growth will be addressed.

The first robustness exercise will be to perform an instrumental variable analysis of equation 1.14. Per-capita activity estimates can be obtained for years prior to 2001 (the first year in the sample), so lagged inequality can be used as an instrument for initial inequality, as is common in the growth literature. Lagged inequality is likely to suffer from the same potential endogeneity issues discussed in the previous section, though it is less correlated with error term than initial inequality. A recent literature on identification in the presence of imperfect instruments (Nevo and Rosen 2012; Conley, Hansen, and Rossi 2012) can therefore be leveraged and well-identified estimates may be obtainable.

The pixels for which growth and inequality are calculated need not be defined by political boundaries. The robustness of the results can therefore be checked by defining arbitrary regional boundaries and repeating the analysis. Additionally, regions need not consist of adjacent pixels. If the effects of inequality on growth occur due to regional redistribution or politics, defining regions as arbitrarily chosen pixels should eliminate any effect of inequality on growth. On the other hand, if the relationship between inequality and growth is driven by convergence, regions comprised of arbitrary pixels should yield similar parameter estimates.

The methods described here make it possible to estimate the entire spatial distribution of per-capita economic activity. Poverty rates can be thus be estimated for any arbitrary geographical area. Using World Bank estimates of a national headcount ratio, a poverty rate in terms of per-capita lights can be constructed. This would allow the effects of poverty on growth to be separately estimated, or allow changes in poverty rates to serve as independent variables.

Given that the method developed here is not dependent on any particular geospatial boundaries, estimates of growth, inequality, and poverty rates may be obtained for any region. Asher and Novosad (2017) show that growth rates (as measured by stock returns) in electoral districts in India that voted for ruling officials are higher than in electoral districts that voted for political opponents. Similar findings could likely be produced with the data I compile here, which would allow for a more robust measure of economic growth. Furthermore, I am able to describe the growth of any part of the distribution, including impoverished constituents. Politicians may reward some parts of the distribution more than others, and these rewards likely vary with party affiliation.

1.6.1 Village Population Data

The Census of India releases population counts at the village level, which is substantially more disaggregated than the district-level counts used thus far. The primary hurdle to using this village-level data is the lack of official documentation regarding the village boundaries. Without an ability to match population counts to the geospatial location of the village, I am unable to utilize these data in the estimation procedure.

Efforts to obtain and compile village boundaries into usable spatial datasets have been undertaken. Most successfully is the DataMeet project (datameet.org), which coordinates open-source efforts to draw village boundaries based on a multitude of sources. As of June of 2017, this community has compiled village boundaries for five of India's twenty-nine states. In addition to offering a robustness check for the population estimates, these data can also be used to train the prediction algorithm directly. However, it is unknown whether these data will allow for more robust estimation, due to the lack of geographical coverage.

1.6.2 Pollution Inequality

The population estimates developed here can be combined with other remotely sensed data sources to obtain disaggregated estimates of various outcome variables. Of particular interest is pollution exposure; annual, remotely sensed, ground-level estimates of $PM_{2.5}$ are available at the same spatial resolution as the population estimates developed here.[7] Combining these sources would therefore allow for estimates of per capita pollution exposure.

One possible application of these data would be an investigation of how pollution exposure changes with electoral outcomes. Using data from parliamentary elections, the differences in pollution exposure within districts that voted for the ruling party can be compared with those that did not. Of course, vote shares may be endogenous—unobserved characteristics within electoral districts may be correlated with voting and with pollution exposure.

To account for this source of endogeneity, the effects of electoral outcomes on changes in pollution exposure can be identified using a discontinuity design. Given the winner-take-all nature of parliamentary elections, a natural discontinuity occurs at the 50% boundary. Formally, I would be able to estimate

$$\Delta P_d = \beta_0 + \beta_1 v_d + \beta_2 D_d + \gamma X_d + \varepsilon_d \tag{1.15}$$

where ΔP_d is the change in average pollution exposure in electoral district d before and after the election, v_d is the vote share accruing to the ruling party in district d, and D_d is a dummy equal to one if the vote share is greater than 50%, and X_d is a vector of district-level controls, including night-light activity and urbanization rates.

In addition to average per-capita pollution exposure, I am also able to estimate exposure inequality. Politicians may target specific places in the pollution distribution, perhaps focusing their attention on highly polluted areas, or instead focusing on areas with lower pollution (which may

[7] See chapter 3 for more details on the pollution data.

have higher income). Since I can estimate per-capita economic activity, I can explore each of these questions in the data.