NBER WORKING PAPER SERIES

USING NEURAL NETWORKS TO PREDICT MICRO-SPATIAL ECONOMIC GROWTH

Arman Khachiyan
Anthony Thomas
Huye Zhou
Gordon H. Hanson
Alex Cloninger
Tajana Rosing
Amit Khandelwal

Working Paper 29569
http://www.nber.org/papers/w29569

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2021

This project was funded through the support of the Russell Sage Foundation program on Computational Social Science. This funding did not involve any clearance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Neural Networks to Predict Micro-Spatial Economic Growth
Arman Khachiyan, Anthony Thomas, Huye Zhou, Gordon H. Hanson, Alex Cloninger, Tajana Rosing, and Amit Khandelwal
NBER Working Paper No. 29569
December 2021
JEL No. R0

ABSTRACT

We apply deep learning to daytime satellite imagery to predict changes in income and population at high spatial resolution in US data. For grid cells with lateral dimensions of 1.2km and 2.4km (where the average US county has dimension of 55.6km), our model predictions achieve R2 values of 0.85 to 0.91 in levels, which far exceed the accuracy of existing models, and 0.32 to 0.46 in decadal changes, which have no counterpart in the literature and are 3-4 times larger than for commonly used nighttime lights. Our network has wide application for analyzing localized shocks.

Arman Khachiyan
Economics Department
UC San Diego
arman.khachiyan@gmail.com

Anthony Thomas
Department of Computer Science
and Engineering
UC San Diego
San Diego, CA 92093
ahthomas@eng.ucsd.edu

Huye Zhou
Department of Mathematics
UC San Diego
San Diego, CA 92093
h1zhou@ucsd.edu

Gordon H. Hanson
Harvard Kennedy School
Harvard University
79 John F. Kennedy St.
Cambridge, MA 02138
and NBER
gordon_hanson@hks.harvard.edu

Alex Cloninger
Department of Mathematics
San Diego, CA 92093
acloninger@ucsd.edu

Tajana Rosing
Department of Computer Science
and Engineering
UC San Diego
San Diego, CA 92093
tajana@eng.ucsd.edu

Amit Khandelwal
Graduate School of Business
Columbia University
Uris Hall 606, 3022 Broadway
New York, NY 10027
and NBER
ak2796@columbia.edu

1 Introduction

Spatial economic analysis evaluates how localized shocks–e.g., infrastructure projects (Redding and Turner, 2015), factory openings (Greenstone et al., 2010), and natural disasters (Boustan et al., 2020)–affect the geographic distribution of economic activity. Standard approaches match administrative or survey data to the geospatial structure of these shocks. Because data tend to be released infrequently (e.g., decadally for Censuses) and for relatively coarse spatial units (e.g., counties or metro areas), this method is suitable for assessing long-run economic impacts at a broad spatial scale (e.g., Faber, 2014; Baum-Snow et al., 2017). By contrast, assessing the impact of shocks at the neighborhood level across all cities nationally would be infeasible with conventional data in most countries.

Satellite imagery offer a path forward. Recent work leverages nighttime light intensity to study regional economies where conventional data are sparse (see, e.g., Donaldson and Storeygard, 2016). Although nightlights can detect changes in economic activity across cities, states, and countries, they are problematic at smaller spatial scales. High luminosity in city centers may saturate satellite sensors, leading to top coding, while surface reflectance may cause light to bleed across space, making urban footprints appear artificially large. Aggregating imagery addresses these problems, but dampens spatial variation. To increase granularity, recent work in remote sensing and computer science uses convolutional neural networks (CNNs) to predict outcomes from multi-spectral daytime satellite imagery at high spatial resolutions. This research detects cross-sectional variation in spending and wealth for villages in Africa (Jean et al., 2016) and poverty rates across a diverse sample of cities (Babenko et al., 2017; Piaggesi et al., 2019). In related work on $1km$ grid cells in the US, Rolf et al. (2021) develop a "task-agnostic" learning approach to predict a broad set of localized outcomes.

This paper makes two advances over the existing literature. First, we implement a CNN to predict changes in local economic activity from changes in high-resolution daytime satellite imagery. We achieve high predictive accuracy in the *cross section*, as others have done, and in predicting localized outcomes in the *time series*, which has not been the focus of previous work. Second, we demonstrate that our approach far outperforms nighttime lights at predicting changes at fine spatial scales.[1]

For inputs in model training, we use multi-spectral imagery from Landsat; for

[1]Given their wide use in spatial analysis, nightlights are a natural benchmark for comparison. See, e.g., Chen and Nordhaus (2011), Henderson et al. (2012), Gennaioli et al. (2013), Michalopoulos and Papaioannou (2014), Storeygard (2016), Bruederle and Hodler (2018), Henderson et al. (2018), Hjort and Poulsen (2019), and Jedwab and Storeygard (2021). In the policy domain, the World Bank has produced a quarterly data set, Light Every Night, which records localized nighttime light intensity from 1992 to 2020.

labels, we use household income and population for Census Blocks in the US Census and American Communities Surveys (ACS). Working in the data-rich US setting, we are able to train a CNN from scratch using hundreds of thousands of images and training labels. Matching Census data with Landsat to construct square images with side lengths of $1.2km$ or $2.4km$, we predict levels and changes in income and population.[2] In the test set, model predictions achieve R^2 values of greater than 0.85 in levels and 0.32 in time differences, which compare to R^2 values for predictions in levels of 0.42 for income and 0.75 for population in Rolf et al. (2021). There are no estimates in the literature to benchmark our predictions of changes in local income and population.

Methodologically, we advance the scale and specificity at which machine learning is used to predict local changes in economic activity. Rather than beginning with image features generated by existing models for prediction–which is the standard practice of transfer learning–we train and tune CNN models for all urbanized pixels in the contiguous US from the ground up. This computationally demanding approach allows us to detect the low-level image features (i.e., shapes, shades, edges, clusters) that are informative for predicting income and population, beyond those that have proven useful in other image tasks (Rosenstein et al., 2005).

Our approach complements Rolf et al. (2021), who aim for generality rather than specificity in predicting outcomes from satellite imagery. They use a layer of randomly initialized filters–based on sampling a small patch from the imagery–to extract features from the raw images. These features are then used to predict outcomes of interest. Their process requires little training, is undemanding computationally, and is suitable to predicting many outcomes, but may not be well tuned to specific prediction tasks. Our approach, while highly intensive in training and computation, is bespoke for predicting local changes in income and population.

Our model and code can be used to impute high-frequency outcomes in between the periodic data drawn from large-scale surveys, to train models with imagery where Census data exist but are sparse, and to predict levels and changes in income and population for spatially disaggregated units where Census data are unavailable entirely.[3] We conclude with a discussion of potential applications.

[2]For comparison, in 2010 US Census Blocks had an average size of $0.9km$ x $0.9km$.

[3]Our code, model, and output are available at https://github.com/thomas9t/spatial-econ-cnn.git. This repository includes scripts and computed weights which can be used to augment or extend our modeling approach. It also includes data and instructions for direct applications using our generated income and population measures.

2 Data & Methods

2.1 Imagery and Label Data

For satellite imagery, we use daytime surface reflectance detected by the USGS Landsat 7 satellite, which has 7 spectral bands (3 visible, 2 near-infrared, 1 thermal, 1 mid-infrared), covers the Earth's surface biweekly, and has a spatial resolution of $30m$. Using Google Earth Engine (Gorelick et al., 2017), we construct annual composites of surface reflectance for the May-August median of cloud-free images each year.[4]

To avoid populating the data with a large number of images covering uninhabited areas, we limit the sample to Landsat pixels corresponding to urbanized US Census Block Groups.[5] We first rank Block Groups according to population density in 2000 and identify those in descending rank order that collectively comprised 85% of the continental US population in that year. We then draw a 1-mile buffer around these Block Groups and include all images within the buffer in our sample. Following this procedure, our data cover 93% of the continental US population in 2000. We construct individual images from Landsat imagery as squares. We test two image sizes, one with $2.4km$ sides and one with $1.2km$ sides (see Figure 1).[6] Smaller images, which increase the spatial resolution of the ultimate predictions, may be more useful in some applications, but may also be more challenging to model as they have fewer pixels, and therefore less information available, per image.

Labels for the analysis are constructed from the US Census for 2000, 2010, and 2020, and the ACS five-year samples for 2005-2009, 2008-2012, and 2015-2019, all extracted from Manson et al. (2020). From each sample, we use population by Census Block and total personal income, for residents ages 15 years and older, by Census Block Group.[7] Because income data are only published at the Block Group level, we interpolate income from Block Groups to Blocks according to the population distribution across Blocks within Groups.[8] We further interpolate income and population from Census Blocks to images based on the geographic overlap between the two.

[4]Using summer months averts irregularities due to persistent clouds or snow.

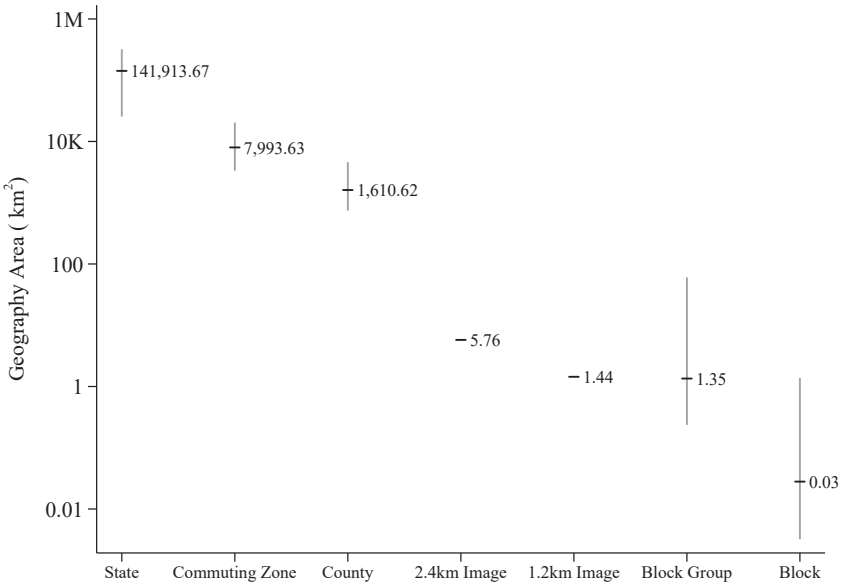
[5]Census Blocks (600 to 3,000 residents) are the smallest geographic unit in the Census; Block Groups are the next smallest unit. In 2000, there were 211,267 Block Groups, with a mean of 39 Blocks per Group. We exclude Census Blocks in which more than 10% of the population was living in group quarters in 2000.

[6]The $2.4km$ and $1.2km$ images have pixel dimensions of 80x80 (6,400 pixels) and 40x40 (1,600 pixels).

[7]Personal income includes wages and salaries, tips and bonuses, proprietor's income, government cash transfers, interest and rental income, and retirement benefits. In-kind government transfers, capital gains, and revenue from property sales are not included (Manson et al., 2020). All values are in 2012 dollars.

[8]Because Block population is unavailable in the ACS data, we use the 2010 population to interpolate 2007 income from Block Groups to Blocks, and similarly use 2020 population to interpolate 2017 income.

Figure 1: Geographic Area of Census and Image Units



Note: This figure shows the geographic area covered by various Census geographic units alongside our constructed images. Horizontal black dashes display the median area for each geographic unit; grey vertical lines show the range from the 10th percentile of area to the 90th percentile of area for each geography. Note that the y-axis is a log-scale of area.

2.2 Convolutional Neural Networks for Spatial Economic Analysis

Although images are an information-rich medium, their unstructured and high-dimensional nature make them difficult to use with conventional learning algorithms, such as LASSO regression. The ability of CNNs to learn structure from data has revolutionized image processing (LeCun et al., 2015). A CNN consists of a sequence of layers, each of which implements a parameterized nonlinear transformation of its inputs. The inputs to the first layer are raw images, in our case 7-dimensional images from Landsat. The output of the first layer is used as input by the second layer and so on. The transformation implemented by each layer is typically either a convolution or pooling operation (Goodfellow et al., 2016), which can be visualized by sliding a rectangular window (e.g., $3 \times 3 \times 7$) over the input image. At each position, an inner product is performed, which aggregates the pixel values in the window into a single number. The output of either a convolution or a pooling operation is another image in which the pixels are these aggregated values.[9] After a sequence of convolutional and

[9]In a convolutional layer, the window contains coefficients used to compute a weighted sum of the pixel values within each window via convolutional filtering. The CNN learns these weights to identify a feature of the image. By applying a sequence of transformations that learn features at increasingly coarse spatial scale,

pooling layers, the transformed image passes through a fully-connected layer, which is a nonlinear regression that maps the image features extracted by the convolutional and pooling layers to a predicted outcome. The parameters of the model are fit using a gradient-based optimization algorithm known as stochastic gradient descent, which minimizes the MSE over labeled training examples.

In our context, a CNN extracts economic information that is latent in spectral data. Asphalt, cement, gravel, soil, water, vegetation, and other materials vary in their reflectance intensity across the light spectrum (e.g., De Fries et al., 1998). The presence of these materials varies enormously within an urban area: more vegetation and loose soil in green spaces; more asphalt and cement around motorways; more steel and wood, together with concrete, in houses and buildings (Zha et al., 2003). The shapes of these materials exhibit similarly wide variation: irregular edges in green spaces, intermittent grids of grass and roofing material in suburbs, larger rectangular clusters in apartment complexes and shopping malls, and compact, interconnected grids in urban centers (Ural et al., 2011, Pesaresi et al., 2016). It is this complexity that makes a neural network powerful—the network learns the mapping of materials and shapes to the level of economic activity and changes in materials and shapes to changes in economic activity. As an empirical regularity, the features learned by the network are often organized into a hierarchy of complexity (Zeiler and Fergus, 2014), in which early layers learn to identify simple features, such as edges or basic shapes, and subsequent layers learn to compose these simple features into complex objects, such as office buildings, industrial parks, suburban developments.

The predicted values that our analysis generates will be subject to error. In regression analysis, measurement error in the outcome variable does not generate bias in estimating treatment effects if this error is uncorrelated with the treatment being studied.[10] Because treatments may be correlated with initial levels of economic development, we wish to eliminate any correlation between prediction errors and initial conditions. To do so, we include controls for local economic characteristics in the initial time period (as measured in Census data) in our CNN models.[11] An added virtue of this approach is that it may improve model accuracy, thereby reducing the scope for

CNNs are able to represent complex spatial relationships between pixels in an image. In a pooling layer, we condense all pixel values within the window to a single number—typically the maximum pixel value within the window. Pooling differs from convolution primarily in that it does not require any learned weights. Pooling serves to reduce the size of the image, which lowers the computational burden of subsequent layers, and helps make the features detected by convolutions robust to small spatial transformations.

[10]For example, if the assigned treatment (a new highway) had a strong positive correlation with the measurement error in the outcome (larger positive deviations between actual and predicted population or income near the highway), this would lead to an overestimate of the true treatment effect.

[11]A full list of variables included can be found in Appendix Table 1.

prediction errors to contaminate analysis that uses our predictions as outcome variables in the first place. Implementing our approach, we find minimal correlations between prediction errors and initial conditions in our data.[12]

2.3 Training, Tuning, and Testing Procedure

CNNs contain a large number of tunable parameters—known as hyperparameters—which control the model architecture and optimization process (e.g., the dimension of convolution filters, number of channels produced by each convolution layer, strength of regularization on weights, and step size used by the optimization algorithm). CNNs are prone to overfitting, in which a model generates accurate predictions on the data used to fit parameters, but fails to generalize on out-of-sample data. To obtain accurate estimates of the model's out-of-sample performance and to determine the best values for hyperparameters, we follow standard practice in empirical machine learning by partitioning our data into three disjoint subsets for training, validation, and testing (Friedman et al., 2001). The training set is used to fit model parameters, and the validation set is used to estimate the out-of-sample error for a given set of hyperparameters. The final model is obtained by selecting the hyperparameters that yield the lowest prediction error in the validation set. The test set is used to obtain an estimate of out-of-sample error for the final model. Ideally, we would repeat this partitioning many times to obtain an estimate of the distribution of out-of-sample error. However, this is infeasible at our data scale.

Models are trained to minimize the MSE of the prediction using the Adam optimizer (Kingma and Ba, 2014). When training models in levels, we pool training data for the years 2000 and 2010, and train a single model to predict outcomes in this combined sample. An alternative approach would be to specialize models in levels to a particular year. However, this method led to greater over-fitting, where training on pooled data resulted in only modest losses in accuracy. We tune hyperparameters for the learning rate (step size and decay rate) and strength of L2-regularization on weights. The training images are randomly augmented to prevent overfitting (cropping, flipping and zooming). We stop the optimization process after 200 epochs or if the R^2 on the validation set fails to increase for 50 epochs. In the latter case we retain the weights which maximize the validation R^2. Further details are in the online appendix.

To obtain reliable estimates of out-of-sample performance, the training, validation, and test sets must be disjoint. To construct these subsets, we partition the full set

[12]The largest correlation coefficient for the income differences model in the test set is 0.057 (for employment in hospitality services), and the median correlation is 0.002. See Appendix Table 1 for details.

of images meeting our inclusion criteria into contiguous urban areas. We randomize selection into training, validation, and test sets at the level of the urban area, rather than the level of the image. Maintaining a disjoint split of the images removes the possibility of data leakage between the training and testing sets (which may result if we allowed images from the two sets to be adjoining). This procedure leads to a total of 4,710 urban regions, which are each randomly assigned to either the train (roughly 50%), validation (roughly 20%), or test (roughly 30%) sets. An image receives the subset designation of the urban region it is contained by, where we discard images located on borders between urban areas (e.g., images on the border between Minneapolis and St. Paul, which are separate urban areas). Appendix Figure 3 shows the distribution of images into each of these sub-groups.

3 Results

3.1 CNN Model Performance

3.1.1 Baseline Results

Here, we present our main results on the predictive power of CNNs. Table 1 Panel A reports R^2 values for model accuracy, again in levels (2000 and 2010) and time differences (2000 to 2010) for $2.4km$ images; Table 1 Panel B repeats the results for $1.2km$ images. Our smaller images are close in dimension to the $1km$ images that Piaggesi et al. (2019) and Rolf et al. (2021) use in their machine-learning approaches to model, respectively, poverty levels and levels of average income and population density in US data. We report performance in the training, validation, and test sets, with and without incorporating initial conditions in model training.[13] For models in levels, we report results for a single model trained to predict both years; performance in each year separately is very similar (see Appendix Table 5).

Beginning with larger images in Table 1 Panel A, we first consider model performance for outcomes in levels. For income and population, and with initial conditions, the R^2 in the test set are 0.90 and 0.91, respectively. Without initial conditions, performance deteriorates moderately, with the R^2 falling by 0.05 to 0.07. Comparing these results to those for smaller image sizes in Table 1 Panel B, the R^2 for income and pop-

[13]The complete set of initial conditions, all measured for the year 2000, are at the county level, log population, log personal income, and the shares of employment in business services, non-business services, and industrial production; and at the Census Block level, population shares for individuals who are female, ages 25 to 54, Black, non-Hispanic white, Hispanic, and living in group quarters, and employment shares for two-digit manufacturing industries, business services, and non-business services.

ulation are 0.85 and 0.86, with initial conditions, and 0.09 to 0.11 lower, without them. The weaker performance of smaller relative to larger images is expected. For smaller images, the network must form predictions based on a smaller number of underlying pixels, which tends to undermine accuracy.

Table 1: R^2 Values for Baseline Models of Large and Small Images

	2000 and 2010 Levels			2000 to 2010 Difference		
	Train	Valid	Test	Train	Valid	Test
Panel A: National 2.4*km* Imagery						
Income						
With Initial Conditions	0.9254	0.8934	0.9018	0.4863	0.4126	0.3962
Without Initial Conditions	0.8625	0.8289	0.8374	0.4951	0.3960	0.3702
Population						
With Initial Conditions	0.9611	0.9029	0.9132	0.5410	0.4839	0.4573
Without Initial Conditions	0.9187	0.8636	0.8684	0.7004	0.4496	0.4202
Panel B: National 1.2*km* Imagery						
Income						
With Initial Conditions	0.8957	0.8620	0.8543	0.3819	0.3061	0.3216
Without Initial Conditions	0.7969	0.7597	0.7494	0.2959	0.2609	0.2690
Population						
With Initial Conditions	0.9101	0.8716	0.8600	0.4217	0.3401	0.3559
Without Initial Conditions	0.7841	0.7612	0.7492	0.3924	0.3051	0.3036

Note: The table shows R^2 values computed on each subset of the images with $2.4km$ and $1.2km$ sides. The total sample size of spatially unique images in training, validation and test subsets is 112,932 for larger images and 320,880 for smaller images. Income measures the log of total personal income, while population is the log of total population. 2000 and 2010 levels represent a model predicting levels for images in the two years together, while the differences columns show the result predicting the change from 2000 to 2010. Initial conditions included in the model are gender and racial composition, employment shares and county level population and income, all measured in 2000. The results show high accuracy in predicting both levels and differences in income and population; there is not strong evidence of over-fitting in the training set. Model fit is consistently lower on the sample of smaller images; hence, we prioritize the sample of $2.4km$ imagery as our baseline analysis sample.

Turning to our predictions for changes over 2000-2012, for $2.4km$ images the R^2 for income and population growth rates in the test set are 0.40 and 0.46, respectively, with initial conditions, and 0.37 to 0.42 without them. For $1.2km$ images, model performance is again somewhat weaker. The R^2 is 0.32 to 0.36, with initial conditions, and 0.27 and 0.30, without them.

Comparing our results for $1.2km$ images to those for $1km$ grid cells in Rolf et al. (2021), we achieve higher performance for both population density (our R^2 of 0.86 versus theirs of 0.72) and income (our R^2 of 0.85 versus theirs of 0.42). We note that whereas our model is trained from scratch for the express purpose of predicting

income and population, their model is constructed for the general purpose of predicting many possible outcomes and therefore may sacrifice accuracy for any specific quantity. Because we are unaware of any prior work that uses CNNs to predict changes in income or population at spatial resolutions similar to our image sizes, we have no benchmark for comparison in the literature for these results.[14]

To evaluate overfitting, we compare predictive accuracy across training, validation, and test sets. Focusing on the time-difference models and on results in validation versus training sets, the R^2 for income growth in $2.4km$ images falls minimally by 0.02 from the validation to the test set, with initial conditions, and by 0.03, without initial conditions; the change in R^2 is slightly larger for population growth. For $1.2km$ images, the R^2 either rises or changes minimally from the validation to the test set, both for income and population and with or without initial conditions. With cross-validation, overfitting in our model training does not appear to be manifest.

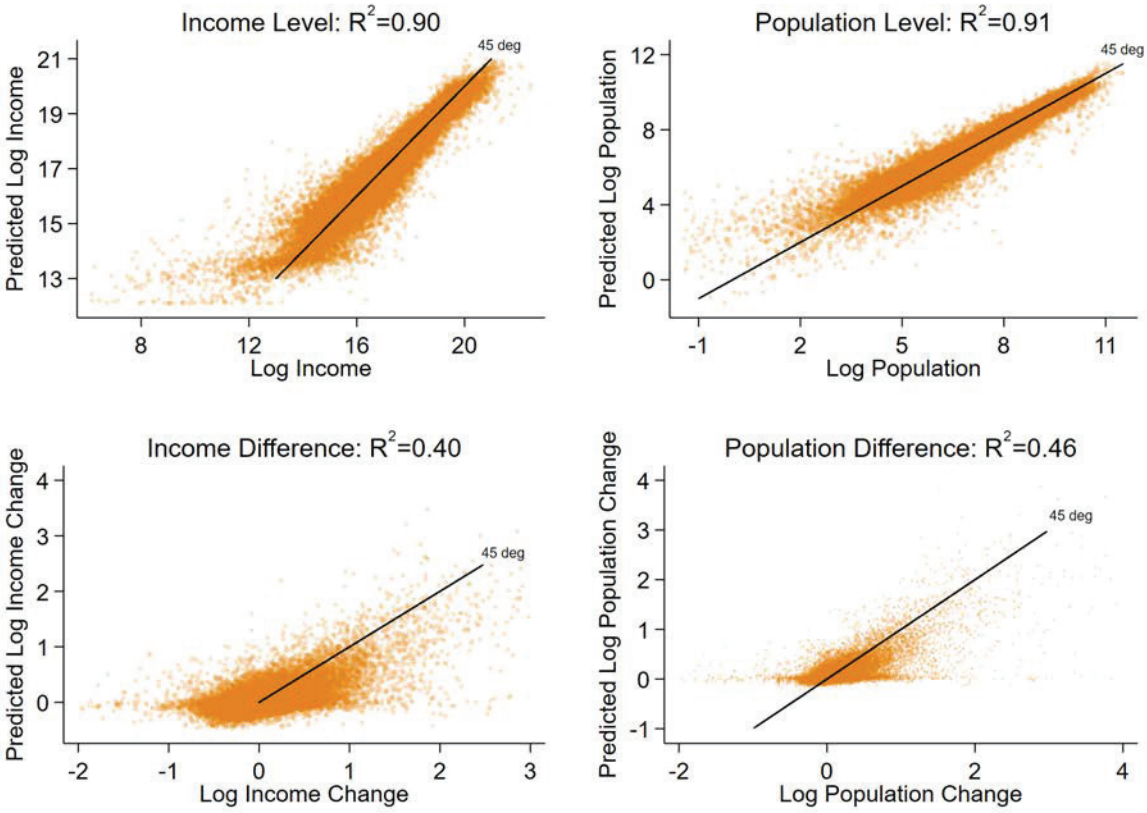
3.1.2 Model Prediction Errors

To evaluate prediction errors in our model, Figure 2 shows scatter plots of model-predicted values and actual values for log income and population, in levels and time differences. In the models for levels, the data are tightly packed around the 45-degree line, indicating that the model accurately captures log income and population across the entire distributions of each. The results for growth rates in the second row show that the prediction of differences is more challenging. The model captures much of the variation for images in which values are growing, but tends to over-predict growth in images for which values are flat or declining, especially for income. The asymmetry in errors for positive and negative growth rates—for income, in particular—may be a result of the slow depreciation of physical capital. Whereas in expanding regions income growth may lead directly to new construction, in declining regions income loss may result in the change or removal of structures over longer time horizons.

To see whether our prediction errors are associated with initial economic conditions, we compute the correlation of our prediction errors with initial industry employment shares and demographic characteristics. These correlations are all below 0.1 and mostly well below 0.02, as seen in Appendix Table 1. Estimating a regression of prediction errors on fixed effects for each urban area in the sample, the fixed effects absorb 11% or

[14]In Appendix Table 2, we report results for log income per capita. In levels for 2000 and 2010 and with initial conditions, we achieve R^2 in the test set of 0.65 for $2.4km$ imagery and 0.61 for $1.2km$ imagery; in changes for 2000-2010 and with initial conditions, we achieve R^2 in the test set of 0.07 for both $2.4km$ and $1.2km$ imagery. Differencing population from income, which removes much of the systematic variation in economic activity from the data, appears to complicate extracting information from satellite imagery.

Figure 2: Model Predictions against Actual Values



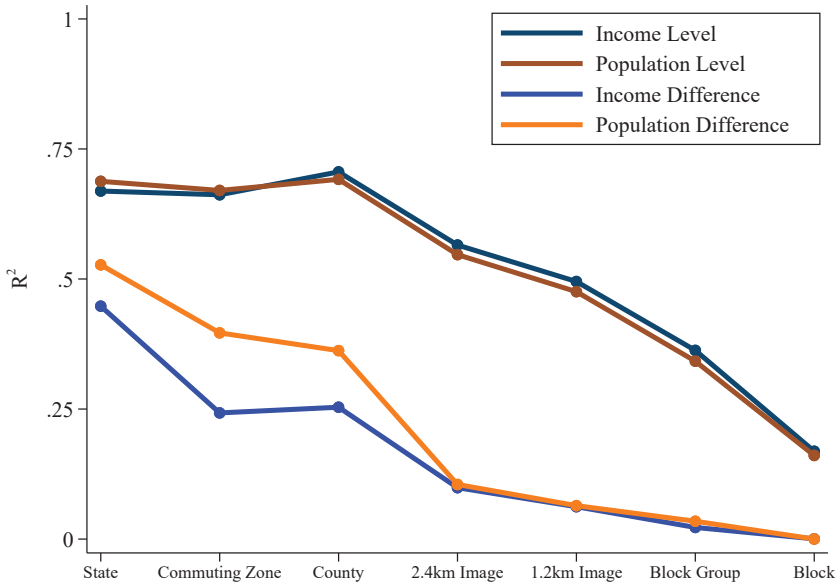
Note: Levels models include data from both 2000 and 2010. Extreme outliers are omitted from this figure to allow visualization of the central tendency in the data.

less of the variation in the errors, as seen in the last row of Appendix Table 1. Appendix Figures 4A and 4B further show no systematic variation in prediction accuracy across geographic regions. In all, there appears to be little covariation between prediction errors and initial economic conditions in our sample.[15]

3.2 Comparison with Nightlight Intensity

Given the growing use of nightlights to detect GDP, as discussed above, we next compare our CNN performance to how well nightlights predict levels and changes in economic activity. In Figure 1, we regress log income or log population on log nightlight intensity, first in levels for the years 2000 and 2010 pooled in a single regression, and then in changes over the 2000 to 2010 time period. The geographies studied range from US states to Census Blocks and include our $1.2km$ and $2.4km$ images. To normalize the size of spatial units, we express all values per km^2.

Figure 3: Nightlight Predictive Accuracy by Geography



Note: This figure shows the linear fit of log income and log population on log night lights for given geographic units, where measures are in values per km^2. Night light intensity is a spatial sum of DMSP-OLS average visible light in both 2000 and 2010. The regression for each geography is conducted with population weights. Results show that nightlights are a powerful predictor of population and income in large geographies, but their effectiveness in smaller geographies is limited.

Figure 3 summarizes the results by presenting the R^2 values for each OLS regression. In the regressions in levels for larger geographies, nightlights are a strong predictor of

[15]In the online appendix, we follow recent literature on interpreting neural network predictions by evaluating saliency maps, which indicate which pixels in an image most influence network prediction.

economic activity, consistent with previous research (Gennaioli et al., 2013; Donaldson and Storeygard, 2016). For income levels in 2000 and 2010, where results for population are very similar, R^2 levels are stable across larger spatial units, at 0.67 for states, 0.66 for commuting zones, and 0.71 for counties. Jumping from counties to our $2.4km$ images, the R^2 drops to 0.57 and drops further to 0.50 for our $1.2km$ images. Even at roughly the neighborhood level—the $1.2km$ images—nightlights are strongly positively correlated with the level of economic activity.

Yet, our CNN trained on daylight imagery substantially outperforms nightlights in cross-section data. Referring to our baseline CNN results in Table 1, the CNN trained on daylight satellite imagery with initial conditions yields an R^2 for log income that is 0.33 higher for $2.4km$ images (0.90 versus 0.57) and 0.35 higher for $1.2km$ images (0.85 versus 0.50); improved accuracy for log population is similar.

The contrast between nightlights and our CNN model is even greater when predicting changes in income or population. For 2000-2010 income changes—where results for population are again similar—R^2 values are 0.10 for nightlights using $2.4km$ images, compared to 0.40 in our CNN with initial conditions (or 0.37 without them), and 0.06 for nightlights using $1.2km$ images, compared to 0.32 in our CNN with initial conditions (or 0.27 without them).[16] At the neighborhood dimension of our $1.2km$ images, changes in nightlights have weak predictive power for changes in economic activity.[17]

3.3 Robustness Exercises

We examine the robustness of our results to changes in the satellite imagery and machine-learning methods used in the analysis.

3.3.1 Performance with RGB Only

We consider the effect of limiting the Landsat imagery used for training to the visible spectrum (i.e., the red, green, and blue (RGB) channels). The non-RGB bands in our imagery more than double the size of the data and therefore significantly increase training complexity. It is therefore useful to examine whether the added modelling complexity of using non-RGB data is justified.

Appendix Table 3 compares test accuracy on models trained with RGB bands alone and those trained with all 7 Landsat bands. For levels models with initial conditions,

[16] Consistent with previous literature, we find that nightlights have sizable predictive power for long-run income changes in larger geographies, achieving R^2 values of 0.45 for states and 0.25 for counties.

[17] This lack of predictive power for nightlights may be due to the fact that the resolution of $1.2km$ images is close to that of the $1km$ pixels for which raw nightlight imagery are available. At the pixel level, perhaps unsurprisingly, changes in nightlights have little information about income or population growth.

we find a modest benefit of adding the four non-RGB bands: the R^2 rises by 0.04 for both log income and log population. The gain is larger for difference models: including the additional non-visible Landsat bands raises the R^2 by 0.06 for log income and by 0.11 for log population. For predicting log growth in income and population, having more complete spectral imagery is of substantial value in predictive accuracy.

3.3.2 Performance of 30m (low) vs 15m (high) Resolution Imagery

The resolution of satellite imagery is a key determinant of the information observable in a fixed image region. The USGS Landsat 7 imagery we use has a native $30m$ resolution. Governments and private companies are working to produce more resolute images. DigitalGlobe, for instance, collects and sells satellite imagery with $30cm$ resolution, where a single $30m$ pixel contains 10,000 $30cm$ pixels. Although such high-resolution data promise massive advances in information content, these gains are counter-balanced by similarly massive increases in computational complexity.

To provide a partial evaluation of the gains to prediction from having higher resolution imagery, we compare model performance when doubling the resolution of daytime satellite imagery from $30m$ to $15m$. To perform this comparison, we construct $15m$ Landsat imagery using panchromatic sharpening, as described and used in Jean et al. (2016). This process restricts the Landsat spectral bands to the RGB wavelengths. The results, which appear in Appendix Table 4, contrast the accuracy of CNN models trained on $1.2km$ images for $30m$ versus $15m$ pan-sharpened RGB bands. To reduce computational complexity, we limit the images used in model training to those in the Mid-Atlantic and Southeast US, as shown in Appendix Figure 3. Results on test samples indicate that using the higher resolution imagery leads to no meaningful improvement in fit across model specifications. For all models, increases in R^2 are less than 0.005. This finding suggests that modestly higher resolution imagery is unlikely to offer large improvements in a network's ability to learn relevant features for out-of-sample prediction at a fixed geographic scale. However, we cannot speak to the possible model accuracy if substantially higher resolution imagery were coupled with the computational resources to conduct a similar exercise.

3.4 Out-of-Sample Predictions

A primary application of our model is to use income and population predictions as outcomes for analyses occurring over periods in which Census data are coarse or unavailable. We offer examples of such analyses in Section 4 and guidance on implementing them in the online appendix. To evaluate the accuracy of our predictions in

out-of-sample time periods, we train and tune a modified model in which we allocate 70% of our images to training and 30% to validation. In this case, we evaluate model performance in periods outside of 2000 and 2010, rather than in a dedicated set of test images as in our baseline models. To estimate accuracy in periods as far from our sample period as possible, we use 2020 for population and 2017 for income.[18]

Table 2: Model R^2 for National 2.4km Imagery in Out-of-sample Periods

	In-Sample Period		Out-of-Sample Period		
Population	**2000, 2010**	**2000-2010**	**2020**	**2010-2020**	**2000-2020**
With Initial Conditions	0.9356	0.5132	0.9193	0.1963	0.4967
Without Initial Conditions	0.8806	0.5030	0.8737	0.1702	0.5106
Income	**2000, 2010**	**2000-2010**	**2017**	**2007-2017**	**2000-2017**
With Initial Conditions	0.9043	0.4910	0.8928	-0.0432	0.4193
Without Initial Conditions	0.8463	0.4331	0.8302	-0.0999	0.3731

Note: The table shows R^2 values computed on all images with 2.4km sides. The sample size of spatially unique images in training and validation subsets is 112,932. Income measures the log of total personal income, while population is the log of total population. The columns delineate fit in the training period and in the out of sample periods, both in terms of levels and differences. Because our imagery panel concludes in 2019, predictions on 2019 imagery are evaluated against actual 2020 population, and 2009 to 2019 change predictions against 2010 to 2020 population change. Initial conditions included in the model are gender and racial composition, residential employment shares and county level population and income, all measured in the initial period (2000 for demographics, 2004 for employment).

Table 2 shows the accuracy of these models when used to predict log population and log income in each period for our larger $2.4km$ images. We find in-period accuracy similar to our baseline model, at 0.90 to 0.94 for levels predictions and 0.49 to 0.51 for time differences (when including initial conditions). This approach also performs well in predicting out-of-sample levels: the R^2 for the levels models including initial conditions is 0.92 for 2020 population and 0.89 for 2017 income. There is little loss in accuracy for predictions in levels when we extend beyond our sample period.

For the more challenging task of predicting out-of-sample changes, we achieve an R^2 of 0.20 for the change in log population over 2010 to 2020, approximately half of the accuracy seen in our baseline results in the in-sample-period holdout test set. However, the income model is unable to outperform the true mean (i.e., $R^2 = 0$) when forecasting income changes over 2007 to 2017. Performance improves markedly when we instead set our base period to be the in-sample year of 2000 and let the end period extend 7 to 10 years beyond the sample. R^2 values are 0.50 for the 2000-2020 population change and 0.42 for the 2000-2017 income change (with initial conditions), which are similar to results for the 2000-2010 sample period.

[18]Block population for 2020 is from the Census Redistricting File; income for 2017 is from the 2015-2019 ACS and imputed to Blocks using the 2020 population.

Lower performance in predicting changes, particularly for income over 2007-2017, may be related to the sluggish recovery to the Great Recession, which may have dampened changes in the visible properties of economic growth. During this period, falling unemployment drove economic growth, a type of cyclical adjustment for which our CNN may be poorly suited. A second explanation is lower quality label data in the out-of-sample periods, particularly for income. Because block-level population is only available in decennial census years, we use the 2010 and 2020 population distributions to disaggregate 2007 and 2017 income, respectively, from block groups to blocks. The resulting noise may be more problematic over a 10-year period than over the longer periods tested, explaining the difference in accuracy. Because this label quality issue coincides with recessionary years, we are unable to disentangle the two explanations.

We conclude from the results in Table 2 that, when evaluated against high quality label data, our approach shows strong potential for producing accurate predictions in out-of-sample periods. The results also indicate that this approach is likely to be most effective when predicting changes over long time horizons, and in periods which do not include large business cycle fluctuations.

4 Discussion

Remotely sensed data have the potential to transform spatial economic analysis. Because much of these data are in the public domain, the cost of working at fine geographic scales is now low. We show that applying convolutional neural networks to daytime satellite imagery predicts microspatial changes in income and population at a decadal frequency. An immediate application is to use predictions of income or population at these spatial scales as outcomes in analysis. Our method can also be used to impute income and population between Census years for the US, to extend to other high-income countries where the relationship between multi-spectral imagery and economic activity is likely to be similar, and to initialize layers for training CNNs in other contexts, thereby reducing computational costs. Khachiyan (2021), for example, uses our output to examine the within-county impacts of the US fracking boom.

A related area that would benefit from such data is the study of place-based policies, such as subsidies to firms that invest in designated areas. Justifying these policies hinges on whether new investments have positive spatial spillovers (Kline and Moretti, 2014; Gaubert et al., 2021). Using our model, researchers could evaluate spillovers at much finer spatial scales than is feasible with public data. Estimating the welfare consequences of place-based policies relies further on addressing their non-random location

and timing. With our model, researchers could examine pre-existing trends and control for spatial-temporal shocks at much finer resolutions (e.g., county-year levels) than is possible in conventional data (in which the county-year may be the unit of analysis).

Another application is the evaluation of transport infrastructure, which has seen major recent advances (Redding, 2020). Satellite-based measures of income and population would allow researchers to evaluate specific projects, such intra-city bus lanes or subway lines, at the neighborhood level across many cities. Such granularity would permit refined tests of economic theory, such as whether transport links lead to more agglomeration in larger nodes (via home market effects) or less agglomeration in intermediate nodes (due to agglomeration shadows). Although researchers have obtained granular information from smartphone data (e.g., Akbar et al., 2018, Kreindler and Miyauchi, 2021) and private transport platforms (e.g., Hall et al., 2018), there may be non-random selection of users who supply these data (e.g., taxi riders in New York City may differ from taxi riders in Phoenix). Satellite imagery offers the equivalent of administrative-level data that is consistent across space and time.

A further application is the analysis of natural disasters. Floods, earthquakes, wildfires, and tornadoes tend to have highly localized impacts (Dell et al., 2014). Our model allows analysts to trace the consequences from point of impact to neighboring communities and to broader metro areas. Such disaggregation is important not just for the academic task of evaluating shock transmission across space but for policy makers who, after disasters occur, require tools to assess where need is likely to be acute.

Finally, our results suggest paths for future work developing predictive models from satellite imagery. First, the model does not perform as well in the shorter frequency out-of-sample prediction exercise, although this could be due to business cycles. Addressing this issue could leverage further the ability to use higher-frequency changes in images to predict economic growth. Second, our model is trained on US data, and future work could explore how well model parameters perform in other countries.

References

Akbar, P. A., V. Couture, G. Duranton, and A. Storeygard (2018, November). Mobility and congestion in urban india. Working Paper 25218, National Bureau of Economic Research.

Babenko, B., J. Hersh, D. Newhouse, A. Ramakrishnan, and T. Swartz (2017). Poverty mapping using convolutional neural networks trained on high and medium resolu-

tion satellite images, with an application in mexico. *IPUMS National Historical Geographic Information System: Version 14.0 [Database]*.

Baum-Snow, N., L. Brandt, J. V. Henderson, M. A. Turner, and Q. Zhang (2017). Roads, railroads, and decentralization of chinese cities. *Review of Economics and Statistics 99*(3), 435–448.

Boustan, L. P., M. E. Kahn, P. W. Rhode, and M. L. Yanguas (2020). The effect of natural disasters on economic activity in us counties: A century of data. *Journal of Urban Economics*, 103257.

Bruederle, A. and R. Hodler (2018). Nighttime lights as a proxy for human development at the local level. *PloS one 13*(9), e0202231.

Chen, X. and W. D. Nordhaus (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences 108*(21), 8589–8594.

De Fries, R., M. Hansen, J. Townshend, and R. Sohlberg (1998). Global land cover classifications at 8 km spatial resolution: the use of training data derived from landsat imagery in decision tree classifiers. *International Journal of Remote Sensing 19*(16), 3141–3168.

Dell, M., B. F. Jones, and B. A. Olken (2014). What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature 52*(3), 740–98.

Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives 30*(4), 171–98.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from china's national trunk highway system. *Review of Economic Studies 81*(3), 1046–1070.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Gaubert, C., P. M. Kline, and D. Yagan (2021). Place-based redistribution. Working Paper 28337, National Bureau of Economic Research.

Gennaioli, N., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2013). Human capital and regional development. *The Quarterly Journal of Economics 128*(1), 105–164.

Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Greenstone, M., R. Hornbeck, and E. Moretti (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy 118*(3), 536–598.

Hall, J. D., C. Palsson, and J. Price (2018). Is uber a substitute or complement for public transit? *Journal of Urban Economics 108*, 36–50.

Henderson, J. V., T. Squires, A. Storeygard, and D. Weil (2018). The global distribution of economic activity: nature, history, and the role of trade. *The Quarterly Journal of Economics 133*(1), 357–406.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring economic growth from outer space. *American economic review 102*(2), 994–1028.

Hjort, J. and J. Poulsen (2019, March). The arrival of fast internet and employment in africa. *American Economic Review 109*(3), 1032–79.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Jedwab, R. and A. Storeygard (2021, 06). The average and heterogeneous effects of transportation investments: Evidence from Sub-Saharan Africa 1960–2010. *Journal of the European Economic Association*.

Khachiyan, A. (2021). The impacts of fracking on microspatial residential investment. Working paper.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kline, P. and E. Moretti (2014). People, places, and public policy: Some simple welfare economics of local economic development programs. *Annual Review of Economics 6*(1), 629–662.

Kreindler, G. E. and Y. Miyauchi (2021, February). Measuring commuting and economic activity inside cities with cell phone records. Working Paper 28516, National Bureau of Economic Research.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature 521*(7553), 436–444.

Manson, S., J. Schroeder, D. Van Riper, T. Kugler, and S. Ruggles (2020). Ipums national historical geographic information system: Version 15.0 [dataset]. *Minneapolis, MN: IPUMS*. http://doi.org/10.18128/D050.V15.0.

Michalopoulos, S. and E. Papaioannou (2014). National institutions and subnational development in africa. *The Quarterly Journal of Economics 129*(1), 151–213.

Pesaresi, M., D. Ehrlich, S. Ferri, A. Florczyk, S. Freire, M. Halkia, A. Julea, T. Kemper, P. Soille, V. Syrris, et al. (2016). Operating procedure for the production of the global human settlement layer from landsat data of the epochs 1975, 1990, 2000, and 2014. *Publications Office of the European Union*, 1–62.

Piaggesi, S., L. Gauvin, M. Tizzoni, C. Cattuto, N. Adler, S. Verhulst, A. Young, R. Price, L. Ferres, and A. Panisson (2019). Predicting city poverty using satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 90–96.

Redding, S. J. (2020). Trade and geography. Working Paper 27821, National Bureau of Economic Research.

Redding, S. J. and M. A. Turner (2015). Transportation costs and the spatial organization of economic activity. In *Handbook of Regional and Urban Economics*, Volume 5, pp. 1339–1398. Elsevier.

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*.

Rosenstein, M. T., Z. Marx, L. P. Kaelbling, and T. G. Dietterich (2005). To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, Volume 898, pp. 1–4.

Samek, W., A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems 28*(11), 2660–2673.

Simonyan, K., A. Vedaldi, and A. Zisserman (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Storeygard, A. (2016). Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. *The Review of Economic Studies 83*(3), 1263–1295.

Ural, S., E. Hussain, and J. Shan (2011). Building population mapping with aerial imagery and gis data. *International Journal of Applied Earth Observation and Geoinformation 13*(6), 841–852.

Zeiler, M. D. and R. Fergus (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer.

Zha, Y., J. Gao, and S. Ni (2003). Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International journal of remote sensing 24*(3), 583–594.

Technical Appendix

Modelling Appendix

Our modelling approach has two stages. We first train a multi-layer convolutional neural network model, which we use to predict outcomes in levels (income and population in a given year). Our model architecture is a 7-band version of the VGG16 network model, which is widely used in the computer vision community (Simonyan and Zisserman, 2014) and consists of three convolutional blocks followed by a fully connected block. Each convolutional block consists of three two-dimensional convolution layers followed by a max-pooling layer. The output of the final convolutional block is flattened into a vector, which is used as input to the fully connected block. The fully connected block consists of three hidden layers, each separated by a dropout layer. The weights of each layer in the fully connected block are regularized using an L2 norm penalty. To incorporate initial conditions in the models, we standardize all features to be of zero mean and unit variance and concatenate the resulting feature vector to the flattened representation obtained by the CNN. The resulting augmented image representation is then processed by the fully connected block to form predictions. A detailed description of model architecture, including filter sizes and strides, is in the Appendix and in our code on `GitHub`. Appendix Figure 1 shows our model architecture.

We use the model trained in levels to construct a model for predicting time differences in the outcome variables over a given time period (e.g., 2000 to 2010). For each year, we first extract the image representation using the convolutional filters learned by training the levels model, as described above. We then concatenate the vectorized representations for each year and use this as input to a new fully connected block, which is used to predict the difference in outcomes between the two years.

More formally, let $I_a, I_b \in \mathbb{R}^{r \times c \times 7}$ be the input images in years a and b, respectively. We first instantiate a copy of the convolutional layers of the levels model described above, which we denote as a function $f_\phi : \mathbb{R}^{r \times c \times 7} \rightarrow \mathbb{R}^d$. The parameters ϕ are initialized to the weights learned by training the levels model. The predicted outcome of interest is then modeled as $\hat{y} = f_\psi(f_\phi(I_a), f_\phi(I_b))$, where f_ψ can be described by concatenating its two arguments and then applying a dense block as described above. The set of parameters ϕ and ψ is then optimized to minimize the mean-squared-error of the prediction. In this process, we use the levels model to "warm-start" the training of the differences model, based on the intuition that features salient for predicting differences are likely related to, but not coincident with, those for predicting levels.

Levels Models. The levels models consist of three convolution blocks, a "flatten"

layer which vectorizes the output of the convolution layers, and a dense block, which is used to predict the outcome of interest from the features extracted by the convolution blocks. Weights in all layers are initialized using the Glorot Normal random initialization (Glorot and Bengio, 2010). Each convolution layer block consists of three 2D convolution layers followed by a max pooling layer. The convolution layers use a stride of 1 and a kernel size of 3 with ReLu activations. The convolution kernels are regularized using an L2 norm penalty where the strength of the penalty is chosen using cross-validation as described in the body of the paper. The number of filters is constant within each block and increases by a factor of 2 between each block. In other words, if the first block has n filters, the second block outputs $2n$ filters and the third outputs $4n$. The max-pooling layer pools over a 2×2 window. For models that incorporate nightlight intensity, these are included as another channel in the input image.

The output of the convolution blocks is flattened into a vector which is then passed to the dense block. For models that incorporate baseline features (e.g., county level income or population), these features are concatenated to the vectorized output of the convolution blocks. The dense block consists of three fully connected layers each separated by a dropout layer. The fully connected layers use ReLu activations and are regularized by an L2 norm penalty where the strength of the penalty is again chosen using cross-validation and grid-search. The specific set of parameters considered can be found in our code on GitHub. The dropout probability in dropout layers is fixed at 0.5. The number of hidden units in each fully-connected layer in the dense block is set based on the number of filters used in the convolution layers. If the first convolution layer outputs n filters, then each fully connected layer uses $l_i \cdot n$ hidden units, where $l_1 = 16, l_2 = 16$ and $l_3 = 8$. The output of the dense block is passed through a final linear layer which produces a scalar value that is the predicted output. This layer is also regularized by an L2 penalty.

Differences Models. The differences model takes a pair of images, of the same spatial region, in different years as input and produces an estimate of the change in the outcome of interest as output. The images for both years are passed through the levels model as described above and the output of the flatten layer is extracted for each year. For models that incorporate auxiliary features, these features are again concatenated to the output of the flatten layer. The image representations extracted for each year are then concatenated and passed to a dense block as described above. The output of the dense block is again passed to a final linear layer which generates the predicted difference in the outcome of interest. The entire architecture, including the convolution filters in the levels models, is then trained end-to-end.

Computing R^2 Values from CNN Predictions

To compute R^2 in our case of highly non-linear CNN models, we use the general formula of $1 - \frac{SSR}{TSS}$. Here SSR is the sum of squared residuals, where each residual is the difference between the predicted and true value for an image. TSS is conversely the Total Sum of Squares, which is the sum across images of squared differences between each true value and the mean true value of the given outcome.

Code and Data Appendix

While highly effective, developing and training CNN models requires significant computational resources and technical expertise. To assist researchers interested in using our predicted outcomes for their own applications, or in adapting our approach to predict other outcomes or generate predictions in periods or countries outside of our analysis, we have made publicly available our entire code pipeline, image labels and predicted values used to generate results in this paper, and trained CNN models. These resources, along with documentation can be found in our project GitHub at `https://github.com/thomas9t/spatial-econ-cnn.git`.

Specifically, the following resources are available:

- **Code**: The code base used in this paper is available on GitHub. This includes code to (1) extract raw publicly-available imagery from Google Earth Engine and to link imagery with census labels, (2) process image files and convert data to input to the CNN, (3) define and train CNN models, (4) generate predictions using the trained CNN models, and (5) evaluate the accuracy of predictions. Our code base can be directly adapted by researchers to develop new CNN models predicting other outcomes of interest.
- **Model Predictions**: We include CSV files with image-level predictions of income and population levels in each year from 2000 to 2019. We also share predictions of 10-year changes in each outcome for every 10-year period from 2000 to 2019. These predictions are generated for our large images using our out-of-period model (Table 2). These include an image id variable (img_id) and predictions based on models both with and without initial conditions. Each of these files is at the image level, with variables predicting outcomes in a given year and over 10 year changes. Shapefiles of our urban image samples are included for researchers wishing to directly study these geographies. For those interested in aggregating to census geographies, we include a cross-walked version of these predictions for 2010 Census Blocks, which can be further aggregated to containing census geographies (i.e. Counties). Those wishing to study other geographic units will need

to construct their own crosswalk between our images and their units of interest. One way to do this would simply be based on the spatial overlap between units of the different geographies.

- **Trained Models**: For researchers interested in generating predictions on geographies or time-periods not described above, we have also made available the trained parameters of our CNN models. Using these pre-trained models, along with the data processing scripts described above, researchers can input their own Landsat data into our models and compute predicted values. Another use-case would be to use lower levels of our trained CNNs in a transfer learning application in order to reduce the computational cost of training on different economic outcomes.

We also provide documentation and a step-by-step example illustrating how to generate out-of-sample predictions on LANDSAT data not used in our analysis.

Saliency Maps

In addition to validating model performance on a held out test set, it is also useful to assess network performance qualitatively by interpreting the features that appear to be learned by the network. There is a large literature on techniques for interpreting neural network predictions; we focus on saliency mapping, which is simple and widely used (Simonyan et al., 2013; Zeiler and Fergus, 2014; Samek et al., 2016). Saliency maps typically take the form of a heat map showing which pixels in a particular image most strongly influenced the network's prediction. They provide qualitative assurance that the network utilizes "reasonable" features of the image.

The saliency map is generated by calculating the derivative of the score of a class of interest S_c with respect to the input $I \in \mathbb{R}^{r \times c \times d}$ at any image I_0 (Simonyan et al., 2013). In our case, the problem is regression rather than classification. To adapt saliency maps to this setting, we generate the saliency map $M \in \mathbb{R}^{r \times c}$ by

$$M_{ij} = \sum_{c=1}^{7} \left| \omega_{h(i,j,c)} \right|,$$

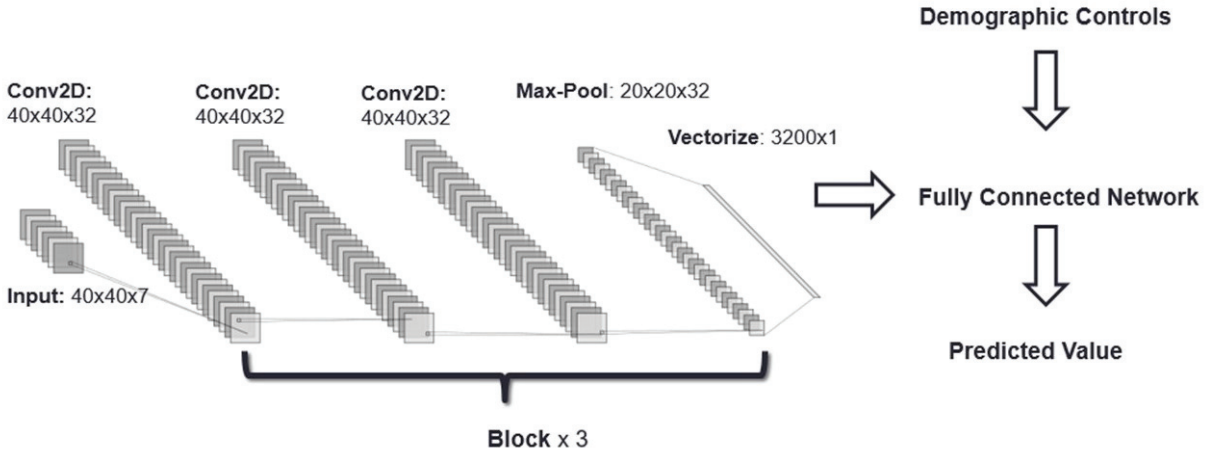
$$\omega = \left. \frac{\partial f}{\partial I} \right|_{I_0},$$

where f is the entire model for prediction, $\omega_{h(i,j,c)}$ is the i-th row, j-th column and c-th channel of ω. In this way, the saliency map will show which features increase the output most across all the channels.

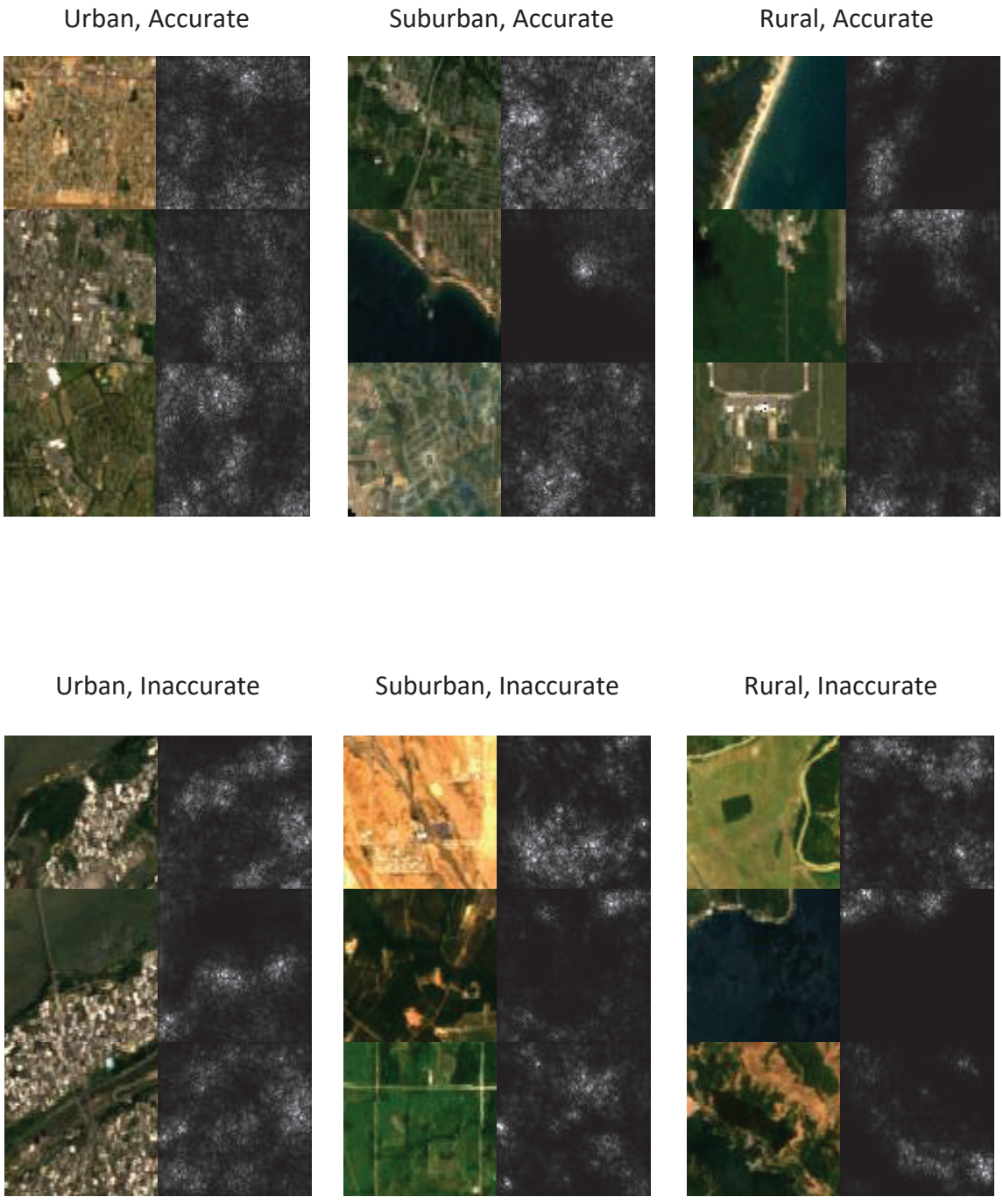
Appendix Figure 2 shows several saliency maps for images in urban, suburban, and semi-rural environs and for which model predictions of income in levels are accurate and inaccurate. Examining cases in which the model performs well and poorly at each population density level gives context on the types of land cover features that are being captured accurately in our models and those that are not. We caution that interpreting saliency is challenging—the motivation for using a CNN is that the relevant image features are unknown and thus one would not expect saliency maps to have in each instance a visually obvious and precise interpretation. Nonetheless, it may be possible to extract some lessons from their examination. Reassuringly, the network ignores water and tends to focus on developed regions in images. For instance, in the second row of the column titled "Rural, Accurate," the network is focusing on the small developed region at the top of the image. Similarly, in the first row of the column "Suburban, Accurate," the model is focusing on the developed region at the lower left. However, in the first and second rows of the column "Urban, Innacurate" the network seems to prioritize undeveloped regions. This is not necessarily a concern, as in some contexts green space is predictive of income. Taken with our quantitative results, which show relatively little evidence of overfitting, the saliency maps suggest that our model is extracting relevant economic information from the images.

Results Appendix

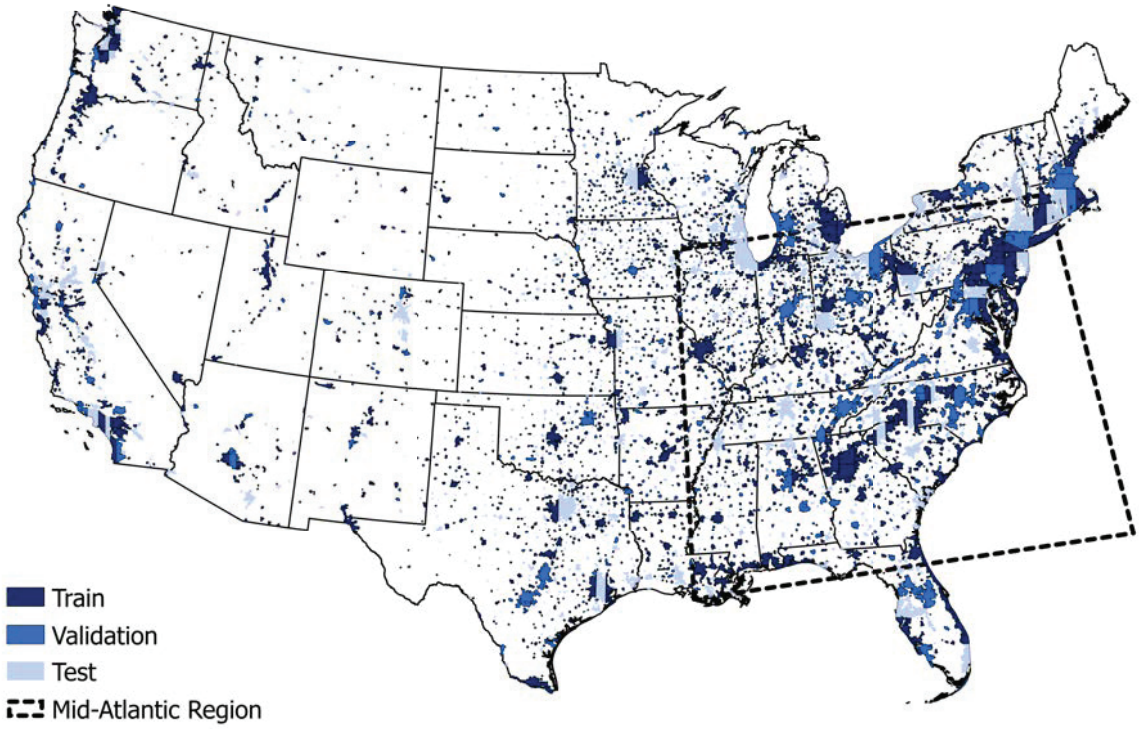
Appendix Figure 1: Convolutional Neural Network, Landsat Imagery Model Architecture



Appendix Figure 2: Selected Saliency Maps

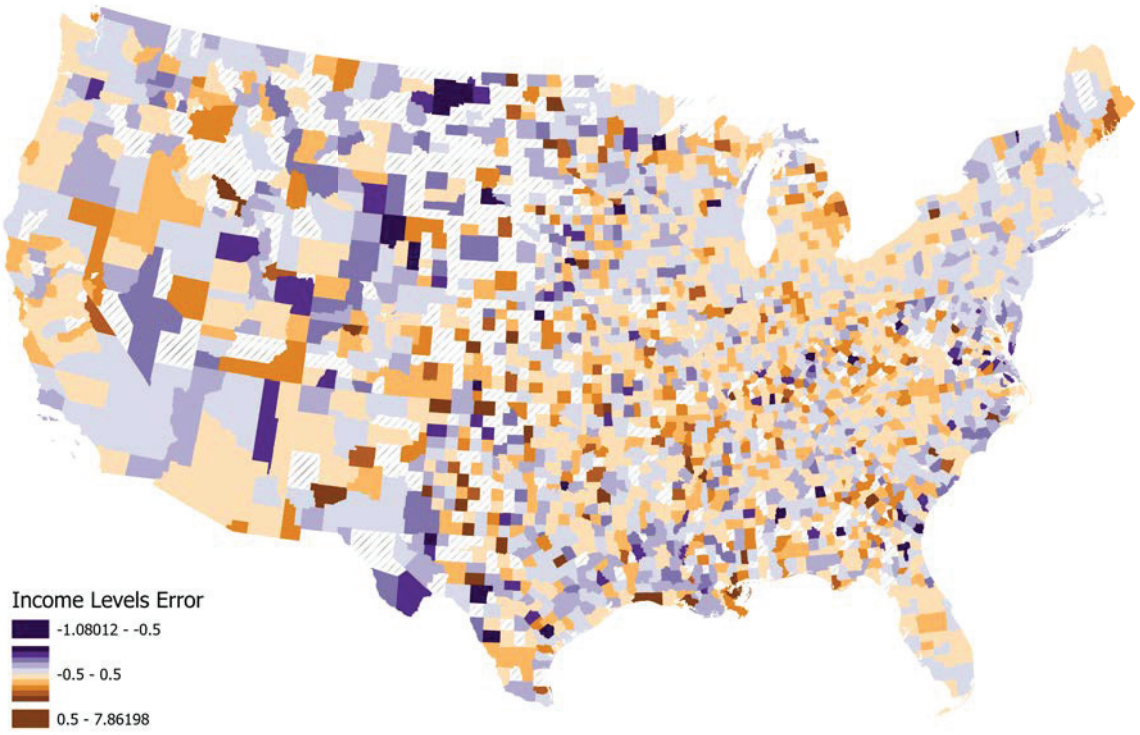


Appendix Figure 3: Spatial Extent of Urban Areas and Model Development Subsets



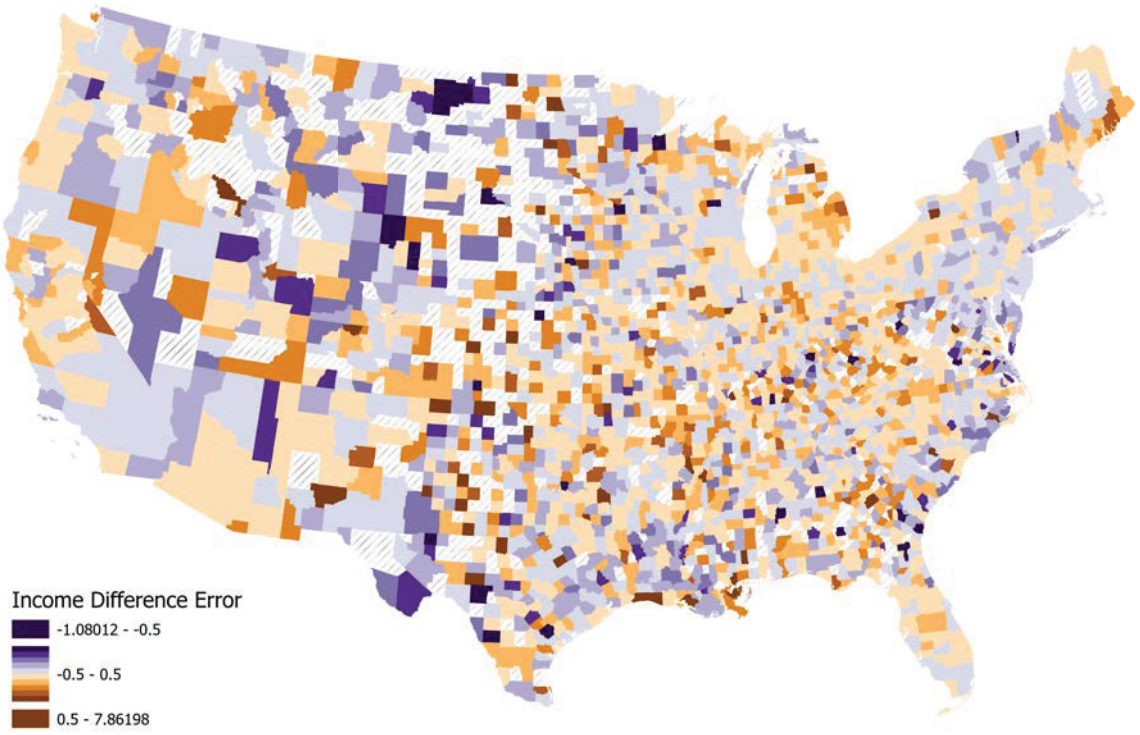
Note: This map shows the urban areas of the contiguous Unites States used to assign images into training, validation and testing subsets for our CNN models. The black dotted line represents the Mid-Atlantic region used for our high-resolution imagery robustness tests discussed in Section 3.3.2. The sample subsets represented in this map are randomly generated and used in training the model for large images; separate randomization is conducted to subset the areas for small images and the high-resolution robustness check. Blank space on this map represents low population density regions (approximately 7% of total population), which are not included in our analysis of urban areas.

Appendix Figure 4A: Levels Model Average Prediction Error across Counties



Note: This map shows the spatial distribution of our modelling error across US counties. We compute image-level prediction error as the average of predicted log income minus actual log income in 2000 and 2010. The color of each county in the map represents the average of these prediction errors across all images in the county.

Appendix Figure 4B: Differences Model Average Prediction Error across Counties



Note: This map shows the spatial distribution of our modelling error across US counties. We compute image-level prediction error as predicted income change from 2000 to 2010 minus actual income change for the same period. The color of each county in the map represents the average of this prediction error across all images in the county.

Appendix Table 1: Prediction Error Correlations with Covariates and Geography

	Income Level	Income Difference	Population Level	Population Difference
Female	–0.0925	0.0409	–0.0620	0.0591
Emp in Business Services	–0.0750	–0.0026	–0.0268	–0.0227
Emp in Accommodation & Food Services	0.0562	0.0570	0.0659	0.0446
Emp in Wholesale Trade	–0.0521	–0.0336	0.0013	–0.0450
Log Income, County	–0.0520	–0.0049	–0.0011	–0.0230
Emp in Administrative/Support/Waste/Remediation Services	–0.0499	0.0247	–0.0127	0.0166
Emp in Production, County	–0.0467	0.0211	–0.0294	0.0387
White	0.0428	–0.0022	0.0101	–0.0052
Emp in Non-Business Services	0.0420	0.0358	–0.0076	0.0408
Log Population, County	–0.0417	–0.0038	0.0015	–0.0179
Hispanic	–0.0415	0.0088	–0.0054	–0.0010
Emp in Business Services, County	–0.0369	0.0173	–0.0426	0.0101
Emp in Construction	–0.0369	0.0169	–0.0470	0.0054
Emp in Professional/Scientific/Technical Services	–0.0364	–0.0008	–0.0288	–0.0077
Emp in Real Estate, Rental & Leasing	–0.0358	0.0042	–0.0273	–0.0046
Emp in Production	–0.0337	–0.0000	–0.0112	0.0189
Emp in Finance & Insurance	–0.0336	–0.0123	–0.0035	–0.0397
Emp in Non-Business Services, County	0.0335	0.0211	–0.0403	0.0451
Emp in Public Administration	0.0293	0.0074	–0.0132	0.0045
Black	–0.0266	0.0052	–0.0266	0.0245
Emp in Information	–0.0251	–0.0000	0.0048	–0.0247
Emp in Transportation and Warehousing	–0.0243	–0.0012	0.0125	–0.0145
Group Quarters	–0.0230	0.0010	–0.0090	0.0161
Emp in Mining/Quarrying & Oil/Gas Extraction	0.0226	–0.0431	–0.0060	0.0080
Emp in Manufacturing	–0.0224	0.0030	0.0165	0.0175
Emp in Retail Trade	–0.0178	0.0060	–0.0353	0.0153
Emp in Agriculture, Forestry, Fishing, & Hunting	–0.0153	0.0010	–0.0227	–0.0004
Emp in Arts, Entertainment & Recreation	–0.0142	0.0258	–0.0062	0.0030
Emp in Health Care & Social Assistance	0.0131	0.0081	–0.0170	0.0352
Emp in Management	–0.0128	–0.0017	0.0067	0.0029
Emp in Utilities	0.0122	–0.0060	0.0049	0.0002
Emp in Educational Services	0.0062	–0.0077	–0.0380	–0.0020
Emp in Other Services	–0.0023	0.0146	–0.0112	0.0123
Working Age	0.0023	–0.0555	0.0020	–0.0718
Urban Area Fixed Effects	0.1067	0.0803	0.0890	0.0669

Note: The table reports correlation coefficients between covariates and prediction errors in each of the four prediction exercises: log income in 2000 and 2010, the change in log income from 2000 to 2010, and the corresponding values for population. The final row shows the R^2 coefficient of an OLS regression on fixed effects by contiguous urban areas (as shown in Appendix Figure 1). All covariates measure an initial value (2004 for employment, 2000 for the rest) at the image-level, and all but the initial county income and population columns represent shares of the relevant image population. These values are spatially interpolated to images from Census Block labels, with the exception of rows listed as County. Residential employment shares are broken down by two-digit NAICS manufacturing industries as well as the aggregates Business Services, Non-Business Services, and Production. Rows are sorted from highest to lowest correlation for income levels. Prediction errors are constructed based on models which include initial conditions.

Appendix Table 2: R^2 Values for Income Per Capita in Large and Small Images

	2000 and 2010 Levels			2000 to 2010 Difference		
	Train	Valid	Test	Train	Valid	Test
National 2.4*km* Imagery						
With Initial Conditions	0.7049	0.6795	0.6533	0.1220	0.0624	0.0674
Without Initial Conditions	0.5077	0.4276	0.3884	0.0984	0.0407	0.0461
National 1.2*km* Imagery						
With Initial Conditions	0.7011	0.6166	0.6091	0.0838	0.0621	0.0653
Without Initial Conditions	0.4502	0.3037	0.3317	0.0534	0.0360	0.0306

Note: The table shows R^2 values computed on each subset of the images with $2.4km$ and $1.2km$ sides. The total sample size of spatially unique images in training, validation and test subsets is 112,932 for larger images and 320,880 for smaller images. Income per capita measures the log of total personal income per person. 2000 and 2010 levels represent a model predicting levels for images in the two years together, while the differences columns show the result predicting the change from 2000 to 2010. Initial conditions included in the model are gender and racial composition, employment shares and county level population and income, all measured in 2000. The results show that predictions on income per capita are less accurate than those on income or population separately, particularly when predicting differences and excluding initial conditions.

Appendix Table 3: Model R^2 for National 2.4km Imagery: All Bands vs RGB Only

	2000 and 2010 Levels			2000 to 2010 Difference		
	RGB Only	LS Bands	LS + NL	RGB Only	LS Bands	LS + NL
Income						
With Initial Conditions	0.8580	0.9018	0.8949	0.3330	0.3962	0.3917
Without Initial Conditions	0.7502	0.8374	0.8429	0.2815	0.3702	0.3827
Population						
With Initial Conditions	0.8781	0.9132	0.9025	0.3467	0.4573	0.4408
Without Initial Conditions	0.7952	0.8684	0.8571	0.3197	0.4202	0.4538

Note: The table shows R^2 values computed on the test set of images with 2.4km sides. The total sample size of spatially unique images in training, validation and test subsets is 112,932. Income measures the log of total personal income, while population is the log of total population. 2000 and 2010 levels represent a model predicting levels for images in the two years combined, while the differences columns show the result predicting the change from 2000 to 2010. Initial conditions included in the model are gender and racial composition, employment shares and county level population and income, all measured in 2000. RGB Only refers to the red/green/blue Landsat 7 bands, LS refers to all 7 Landsat Bands, LS+NL Refers to all 7 Landsat bands plus the DMSP-OLS nightlight band. The results show that including the non-visible Landsat bands improves the model performance, particularly in predicting differences. Further including nightlight data does not improve models (with initial conditions).

Appendix Table 4: Model R^2 in Mid-Atlantic Region: 30m vs 15m Resolution RGB Imagery

	2000 and 2010 Levels		2000 to 2010 Difference	
	30m RGB	15m RGB	30m RGB	15m RGB
Income				
With Initial Conditions	0.7997	0.7970	0.2499	0.2320
Without Initial Conditions	0.6644	0.6683	0.2014	0.1773
Population				
With Initial Conditions	0.8167	0.8189	0.2749	0.2492
Without Initial Conditions	0.7159	0.6995	0.2545	0.2265

Note: The table shows R^2 values computed on the test set of images with 1.2km sides. The total sample size of spatially unique images in training, validation and test subsets is 163,250. Income measures the log of total personal income, while population is the log of total population. 2000 and 2010 levels represent a model predicting levels for images in the two years combined, while the differences columns show the result predicting the change from 2000 to 2010. Initial conditions included in the model are gender and racial composition, employment shares and county level population and income, all measured in 2000. 30m RGB refers to the same Landsat 7 RGB bands used in the rest of the analysis. 15m RGB refers to pan-sharpened RGB bands which are refined from 30m to 15m resolution using the panchromatic Landsat band. All results in this table are based on the Mid-Atlantic subset of the national imagery, shown in Appendix Figure 3, to address the additional computation of analyzing imagery with double resolution. Results show that the extra information of 15m resolution images does not meaningfully improve model accuracy relative to equally sized images with 30m pixels.

Appendix Table 5: Model R^2 for National Imagery By Year

	2000			**2010**			**Diff**		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Panel A: National 2.4*km* Imagery									
Income									
With Initial Conditions	0.9287	0.8981	0.9029	0.9221	0.8887	0.9006	0.4863	0.4126	0.3962
Without Initial Conditions	0.8672	0.8330	0.8373	0.8577	0.8248	0.8375	0.4951	0.3960	0.3702
Population									
With Initial Conditions	0.9610	0.9061	0.9146	0.9613	0.8996	0.9119	0.5410	0.4839	0.4573
Without Initial Conditions	0.9186	0.8620	0.8669	0.9189	0.8652	0.8700	0.7004	0.4496	0.4202
Panel B: National 1.2*km* Imagery									
Income									
With Initial Conditions	0.9032	0.8729	0.8615	0.8883	0.8512	0.8470	0.3819	0.3061	0.3216
Without Initial Conditions	0.7988	0.7604	0.7482	0.7949	0.7591	0.7507	0.2959	0.2609	0.2690
Population									
With Initial Conditions	0.9149	0.8788	0.8650	0.9052	0.8645	0.8548	0.4217	0.3401	0.3559
Without Initial Conditions	0.7815	0.7602	0.7452	0.7867	0.7623	0.7532	0.3924	0.3051	0.3036

Note: The table shows R^2 values computed on each subset of the images with $2.4km$ and $1.2km$ sides. The total sample size of spatially unique images in training, validation and test subsets is 112,932 for larger images and 320,880 for smaller images. Income measures the log of total personal income, while population is the log of total population. Results for 2000 and 2010 are shown separately here, while the differences columns show the result predicting the change from 2000 to 2010 as in Table 1. Initial conditions included in the model are gender and racial composition, employment shares and county level population and income, all measured in 2000. The results show high accuracy in predicting both levels and differences in income and population; there is not strong evidence of over-fitting in the training set. Model fit is lower using the smaller images.