AER: Insights 2022, 4(4): 491–506
https://doi.org/10.1257/aeri.20210422

Using Neural Networks to Predict Microspatial Economic Growth[†]

By Arman Khachiyan, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, and Amit K. Khandelwal*

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 1.2 km and 2.4 km (where the average US county has dimension of 51.9 km), our model predictions achieve R^2 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. (*JEL* C45, R11, R23)

Spatial economic analysis evaluates how localized shocks—for example, infrastructure projects (Redding and Turner 2015), factory openings (Greenstone, Hornbeck, and Moretti 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 offers 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 night lights 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

*Khachiyan: Department of Economics, University of San Francisco (email: akhachiyan@usfca.edu); Thomas: Department of Computer Science and Engineering, UC San Diego (email: ahthomas@eng.ucsd.edu); Zhou: Department of Mathematics, UC San Diego (email: h1zhou@ucsd.edu); Hanson: Harvard Kennedy School, Harvard University, and National Bureau of Economic Research (email: gordon_hanson@hks.harvard.edu); Cloninger: Department of Mathematics and Halıcıoğlu Data Science Institute, UC San Diego (email: acloninger@ucsd.edu); Rosing: Department of Computer Science and Engineering, UC San Diego (email: tajana@ucsd.edu); Khandelwal: Economics Division, Columbia Business School, and National Bureau of Economic Research (email: ak2796@columbia.edu). Amy Finkelstein was coeditor for this article. This project was funded through the support of the Russell Sage Foundation's initiative on Computational Social Science (grant number G-2196).

[†]Go to https://doi.org/10.1257/aeri.20210422 to visit the article page for additional materials and author disclosure statement(s).

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 multispectral 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 1 km grid cells in the United States, 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 multispectral imagery from Landsat; for labels, we use household income and population for census blocks in the US Census and American Community Survey (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 1.2 km or 2.4 km, 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 United States 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

[1] Given their wide use in spatial analysis, night lights are a natural benchmark for comparison. See, for example, Chen and Nordhaus (2011); Henderson, Storeygard, and Weil (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 (2022). In the policy domain, the World Bank has produced a quarterly dataset, Light Every Night, which records localized nighttime light intensity from 1992 to 2020.

[2] For comparison, in 2010 US census blocks had an average size of 0.9 km $\times$ 0.9 km.

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.

I. Data and Methods

A. *Imagery and Label Data*

For satellite imagery, we use daytime surface reflectance detected by the US Geological Survey (USGS) Landsat 7 satellite, which has seven spectral bands (three visible, two near-infrared, one thermal, and one mid-infrared), covers the earth's surface biweekly, and has a spatial resolution of 30 m. 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 percent 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 percent of the continental US population in 2000. We construct individual images from Landsat imagery as squares. We test two image sizes, one with 2.4 km sides and one with 1.2 km 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

[3] Our code, model, and output are available at https://github.com/thomas9t/spatial-econ-cnn.git. This repository includes scripts and computed weights that 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.

[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 percent of the population was living in group quarters in 2000.

[6] The 2.4 km and 1.2 km images have pixel dimensions of 80×80 (6,400 pixels) and 40×40 (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 US dollars.

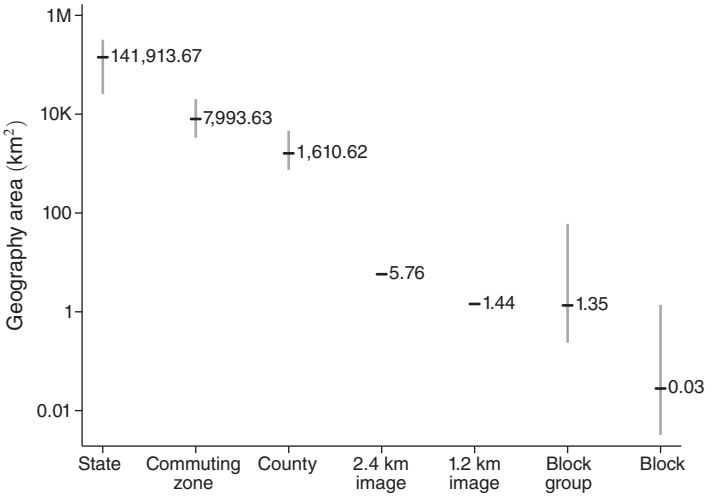

FIGURE 1. GEOGRAPHIC AREA OF CENSUS AND IMAGE UNITS

Notes: 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; gray vertical lines show the range from the tenth percentile of area to the ninetieth percentile of area for each geography. Note that the *y*-axis is a log-scale of area.

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.

B. *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, Bengio, and Hinton 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, seven-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, Bengio, and Courville 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 pooling

[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 the 2020 population to interpolate 2017 income.

[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, CNNs are able to represent complex spatial relationships between pixels in an image. In a pooling layer, we condense all pixel values

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, Gao, and Ni 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, Hussain, and Shan 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, and 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 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]

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 online Appendix Table 1.

[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 online Appendix Table 1 for details.

C. *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 (Hastie, Tibshirani, and Friedman 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 2017). 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 overfitting, 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 that 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 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 percent), validation (roughly 20 percent), or test (roughly 30 percent) 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 Saint Paul, which are separate urban areas). Online Appendix Figure 3 shows the distribution of images into each of these subgroups.

II. Results

A. *CNN Model Performance*

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 2.4 km images; Table 1, panel B repeats the results for 1.2 km images. Our smaller images are close in dimension to the 1 km 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 online 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 population 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.

Turning to our predictions for changes over 2000–2012, for 2.4 km 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 1.2 km 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 1.2 km images to those for 1 km 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]

[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, nonbusiness services, and industrial production; and at the census block level, population shares for individuals who are female, ages 25–54, Black, non-Hispanic white, Hispanic, living in group quarters, and employment shares for two-digit manufacturing industries, business services, and nonbusiness services (US Census Bureau 2020).

[14] In online 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 2.4 km imagery and 0.61 for 1.2 km imagery; in changes for 2000–2010 and with initial conditions, we achieve R^2 in the test set of 0.07 for both 2.4 km and 1.2 km 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.

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

Notes: The table shows R^2 values computed on each subset of the images with 2.4 km and 1.2 km 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 difference 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 overfitting in the training set. Model fit is consistently lower on the sample of smaller images; hence, we prioritize the sample of 2.4 km imagery as our baseline analysis sample.

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 2.4 km 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 1.2 km 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.

Model Prediction Errors.—To evaluate prediction errors in our model, Figure 2 shows scatterplots 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 overpredict 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

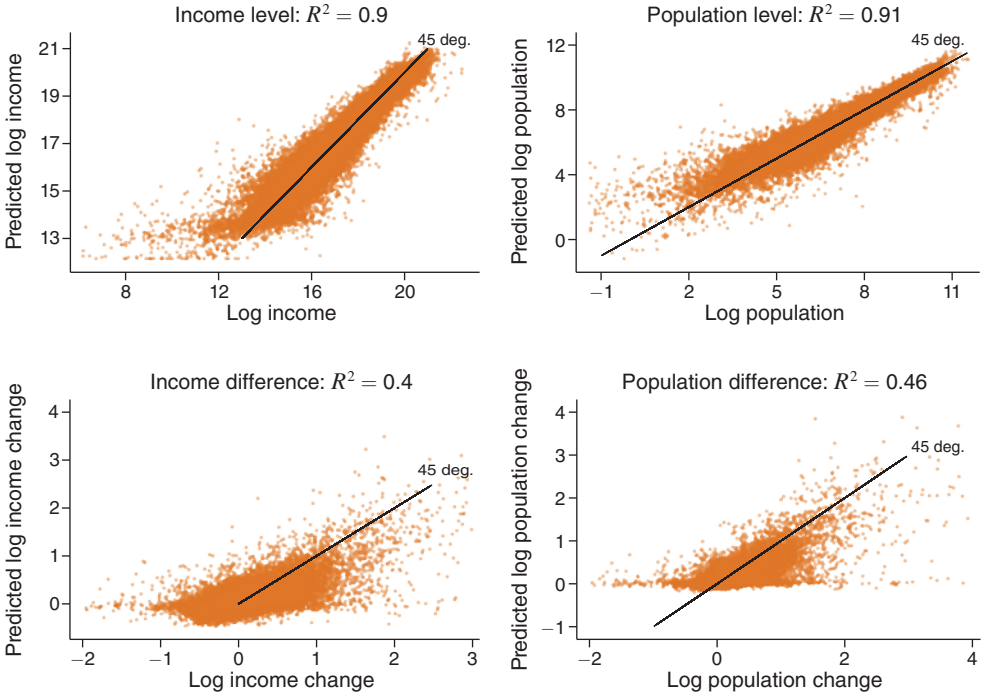

FIGURE 2. MODEL PREDICTIONS AGAINST ACTUAL VALUES

Notes: 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.

employment shares and demographic characteristics. These correlations are all below 0.1 and mostly well below 0.02, as seen in online Appendix Table 1. Estimating a regression of prediction errors on fixed effects for each urban area in the sample, the fixed effects absorb 11 percent or less of the variation in the errors, as seen in the last row of online Appendix Table 1. Online 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]

B. *Comparison with Night Light Intensity*

Given the growing use of night lights to detect GDP, as discussed above, we next compare our CNN performance to how well night lights predict levels and changes in economic activity. In Figure 1, we regress log income or log population on log night light intensity, first in levels for the years 2000 and 2010 pooled in a single regression, and then in changes over the 2000–2010 time period. The geographies

[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.

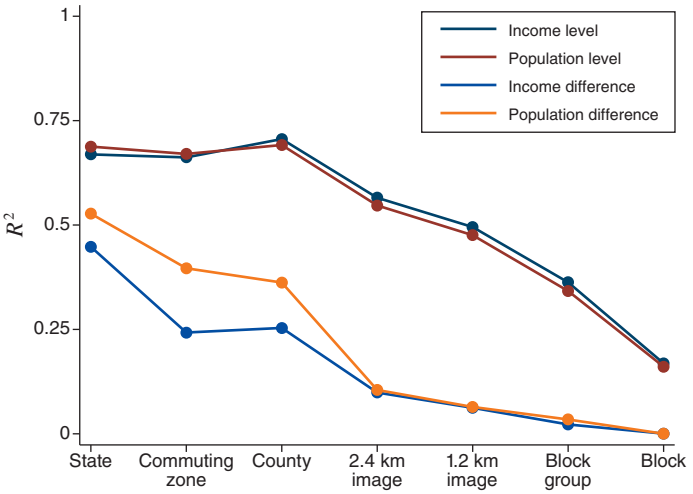

FIGURE 3. NIGHT LIGHT PREDICTIVE ACCURACY BY GEOGRAPHY

Notes: 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 night lights are a powerful predictor of population and income in large geographies, but their effectiveness in smaller geographies is limited.

studied range from US states to census blocks and include our 1.2 km and 2.4 km images. To normalize the size of spatial units, we express all values per km^2.

Figure 3 summarizes the results by presenting the R^2 values for each OLS regression. In the regressions in levels for larger geographies, night lights are a strong predictor of 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 2.4 km images, the R^2 drops to 0.57 and drops further to 0.50 for our 1.2 km images. Even at roughly the neighborhood level—the 1.2 km images—night lights are strongly positively correlated with the level of economic activity.

Yet, our CNN trained on daylight imagery substantially outperforms night lights in cross-sectional 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 2.4 km images (0.90 versus 0.57) and 0.35 higher for 1.2 km images (0.85 versus 0.50); improved accuracy for log population is similar.

The contrast between night lights 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 night lights using 2.4 km images, compared to 0.40 in our CNN with initial conditions (or 0.37 without them), and 0.06 for night lights using 1.2 km images, compared to 0.32 in our CNN with initial conditions (or 0.27 without them).[16] At the neighborhood dimension of

[16] Consistent with previous literature, we find that night lights 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.

our 1.2 km images, changes in night lights have weak predictive power for changes in economic activity.[17]

C. *Robustness Exercises*

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

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 modeling complexity of using non-RGB data is justified.

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

Performance of 30 m (Low) versus 15 m (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 30 m resolution. Governments and private companies are working to produce more resolute images. DigitalGlobe, for instance, collects and sells satellite imagery with 30 cm resolution, where a single 30 m pixel contains 10,000 30 cm pixels. Although such high-resolution data promise massive advances in information content, these gains are counterbalanced 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 30 m to 15 m. To perform this comparison, we construct 15 m 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 online Appendix Table 4, contrast the accuracy of CNN models trained on 1.2 km images for 30 m versus 15 m pan-sharpened RGB bands. To reduce computational complexity, we limit the images used in model training to those in the mid-Atlantic and southeast United States, as shown in online 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

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

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.

D. *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 IV 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 percent of our images to training and 30 percent 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 shows the accuracy of these models when used to predict log population and log income in each period for our larger 2.4 km images. We find in-period accuracy similar to our baseline model, at 0.90–0.94 for levels predictions and 0.49–0.51 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 2010–2020, 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 2007–2017. Performance improves markedly when we instead set our base period to be the in-sample year of 2000 and let the end period extend seven to ten 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.

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

[18] Block population for 2020 is from the Census Redistricting Data Files; income for 2017 is from the 2015–2019 ACS and imputed to blocks using the 2020 population.

Table 2—Model R^2 for National 2.4 km 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

Notes: The table shows R^2 values computed on all images with 2.4 km 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 the actual 2020 population and 2009–2019 change predictions against 2010–2020 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).

over a ten-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 that do not include large business cycle fluctuations.

III. 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 United States, to extend to other high-income countries where the relationship between multispectral 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, Kline, and Yagan 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 nonrandom location and timing. With our model, researchers could

examine preexisting 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 as intracity 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, Palsson, and Price 2018), there may be nonrandom 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, Jones, and Olken 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 policymakers 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

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." arXiv: 1603.04467.

Akbar, Prottoy A., Victor Couture, Gilles Duranton, and Adam Storeygard. 2018. "Mobility and Congestion in Urban India." NBER Working Paper 25218.

Babenko, Boris, Jonathan Hersh, David Newhouse, Anusha Ramakrishnan, and Tom Swartz. 2017. "Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, with an Application in Mexico." arXiv: 1711.06323.

Baum-Snow, Nathaniel, Loren Brandt, J. Vernon Henderson, Matthew A. Turner, and Qinghua Zhang. 2017. "Roads, Railroads, and Decentralization of Chinese Cities." *Review of Economics and Statistics* 99 (3): 435–48.

Boustan, Leah Platt, Matthew E. Kahn, Paul W. Rhode, and Maria Lucia Yanguas. 2020. "The Effect of Natural Disasters on Economic Activity in US Counties: A Century of Data." *Journal of Urban Economics* 118: Article 103257.

Bruederle, Anna, and Roland Hodler. 2018. "Nighttime Lights as a Proxy for Human Development at the Local Level." *PLOS ONE* 13 (9): Article 0202231.

Chen, Xi, and William D. Nordhaus. 2011. "Using Luminosity Data as a Proxy for Economic Statistics." *Proceedings of the National Academy of Sciences* 108 (21): 8589–94.

De Fries, R.S., M. Hansen, J.R.G. 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): 3141–68.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2014. "What Do We Learn from the Weather? The New Climate-Economy Literature." *Journal of Economic Literature* 52 (3): 740–98.

Donaldson, Dave, and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives* 30 (4): 171–98.

Faber, Benjamin. 2014. "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System." *Review of Economic Studies* 81 (3): 1046–70.

Gaubert, Cecile, Patrick M. Kline, and Danny Yagan. 2021. "Place-Based Redistribution." NBER Working Paper 28337.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer. 2013. "Human Capital and Regional Development." *Quarterly Journal of Economics* 128 (1): 105–64.

Glorot, Xavier, and Yoshua Bengio. 2010. "Understanding the Difficulty of Training Deep Feedforward Neural Networks." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ed. Yee Whye Teh and Mike Titterington, 9: 249–56.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment* 202 (1): 18–27.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *Journal of Political Economy* 118 (3): 536–98.

Hall, Jonathan D., Craig Palsson, and Joseph Price. 2018. "Is Uber a Substitute or Complement for Public Transit?" *Journal of Urban Economics* 108: 36–50.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Henderson, J. Vernon, Tim Squires, Adam Storeygard, and David Weil. 2018. "The Global Distribution of Economic Activity: Nature, History, and the Role of Trade." *Quarterly Journal of Economics* 133 (1): 357–406.

Hjort, Jonas, and Jonas Poulsen. 2019. "The Arrival of Fast Internet and Employment in Africa." *American Economic Review* 109 (3): 1032–79.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94.

Jedwab, Rémi, and Adam Storeygard. 2022. "The Average and Heterogeneous Effects of Transportation Investments: Evidence from Sub-Saharan Africa 1960–2010." *Journal of the European Economic Association* 20 (1): 1–38.

Khachiyan, Arman. 2021. "The Impacts of Fracking on Microspatial Residential Investment." Unpublished.

Khachiyan, Arman, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, and Amit K. Khandelwal. 2022. "Replication data for: Using Neural Networks to Predict Microspatial Economic Growth." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E158002V1.

Kingma, Diederik P., and Jimmy Ba. 2017. "Adam: A Method for Stochastic Optimization." arXiv: 1412.6980v9.

Kline, Patrick, and Enrico Moretti. 2014. "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs." *Annual Review of Economics* 6: 629–62.

Kreindler, Gabriel E., and Yuhei Miyauchi. 2021. "Measuring Commuting and Economic Activity Inside Cities with Cell Phone Records." NBER Working Paper 28516.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2020. "IPUMS National Historical Geographic Information System: Version 15.0." IPUMS. http://doi.org/10.18128/D050.V15.0 (accessed April 6, 2020).

Michalopoulos, Stelios, and Elias Papaioannou. 2014. "National Institutions and Subnational Development in Africa." *Quarterly Journal of Economics* 129 (1): 151–213.

Pesaresi, Martino, Daniele Ehrlich, Stefano Ferri, Aneta J. Florczyk, Sergio Freire, Matina Halkia, Andreea Julea, Thomas Kemper, Pierre Soille, and Vasileios Syrris. 2016. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014*. Brussels: Joint Research Centre of the European Commission.

Piaggesi, Simone, Laetitia Gauvin, Michele Tizzoni, Natalia Adler, Stefaan Verhulst, Andrew Young, Rhiannan Price, Leo Ferres, Ciro Cattuto, and André Panisson. 2019. "Predicting City Poverty Using Satellite Imagery." Paper presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, June 16–20.

Redding, Stephen J. 2020. "Trade and Geography." NBER Working Paper 27821.

Redding, Stephen J., and Matthew A. Turner. 2015. "Transportation Costs and the Spatial Organization of Economic Activity." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 1339–98. Amsterdam: North-Holland.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. 2021. "A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery." *Nature Communications* 12: Article 4392.

Rosenstein, Michael T., Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. 2005. "To Transfer or Not to Transfer." Paper presented at the NIPS 2005 Workshop: Inductive Transfer: 10 Years Later, Whistler, British Columbia, December 9.

Samek, Wojciech, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. "Evaluating the Visualization of What a Deep Neural Network Has Learned." *IEEE Transactions on Neural Networks and Learning Systems* 28 (11): 2660–73.

Simonyan, Karen, and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv: 1409.1556v6.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." arXiv: 1312.6034v2.

Storeygard, Adam. 2016. "Farther On Down the Road: Transport Costs, Trade, and Urban Growth in Sub-Saharan Africa." *Review of Economic Studies* 83 (3): 1263–95.

US Census Bureau. 2020. "LEHD Origin-Destination Employment Statistics Data (2004): Version 7." Longitudinal Employer-Household Dynamics Program. https://lehd.ces.census.gov/data/ (accessed January 22, 2020).

Ural, Serkan, Ejaz Hussain, and Jie Shan. 2011. "Building Population Mapping with Aerial Imagery and GIS Data." *International Journal of Applied Earth Observation and Geoinformation* 13 (6): 841–52.

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In *Computer Vision—ECCV 2014*, ed. David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 818–33. Cham, Switzerland: 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): 583–94.