Online Appendix For: Using Neural Networks to Predict Micro-Spatial Economic Growth

By Arman Khachiyan, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, Amit 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 1.2km and 2.4km (where the average US county has dimension of 51.9km), 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.

* This project was funded through the support of the Russell Sage Foundation program on Computational Social Science. Arman Khachiyan (arman.khachiyan@gmail.com), Economics Department, UC San Diego; Anthony Thomas, Computer Science and Engineering Department, UC San Diego; Huye Zhou, Statistics Department, UC San Diego; Gordon Hanson, Harvard University; Alex Cloninger, Mathematics Department and the Halıcıoğlu Data Science Institute, UC San Diego; Tajana Rosing, Computer Science and Engineering Department, UC San Diego; Amit Khandelwal, Columbia Business School.

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 models and training pipelines are implemented in TensorFlow **?**. Our model architecture is a 7-band version of the VGG16 network model, which is widely used in the computer vision community (**?**) 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 $f_\phi : \mathbb{R}^{r \times c \times 7} \to \mathbb{R}^d$. 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 (**?**). 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 (**???**). 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 (**?**). 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

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

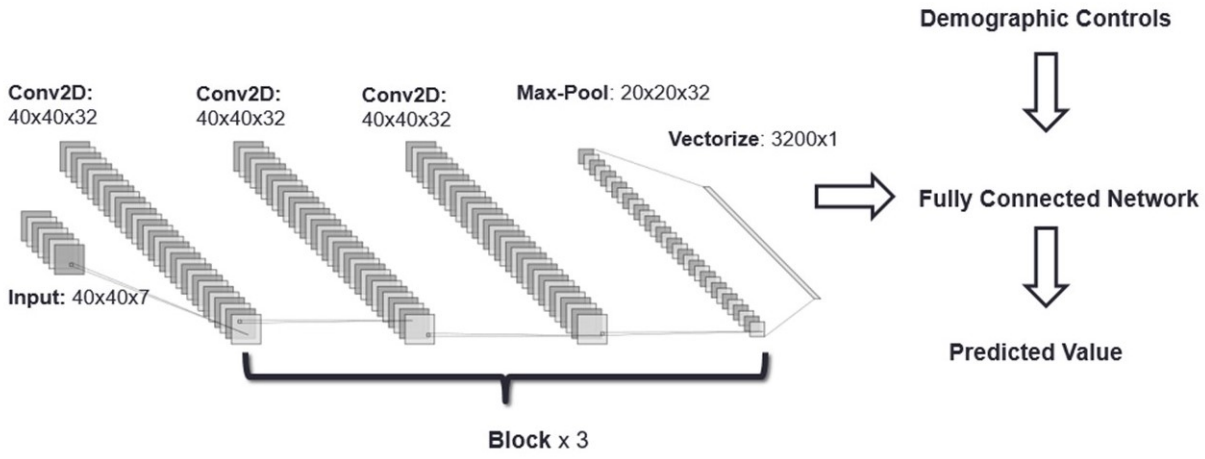
$$\omega = \left. \frac{\partial f}{\partial I} \right|_{I_0},$$

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.

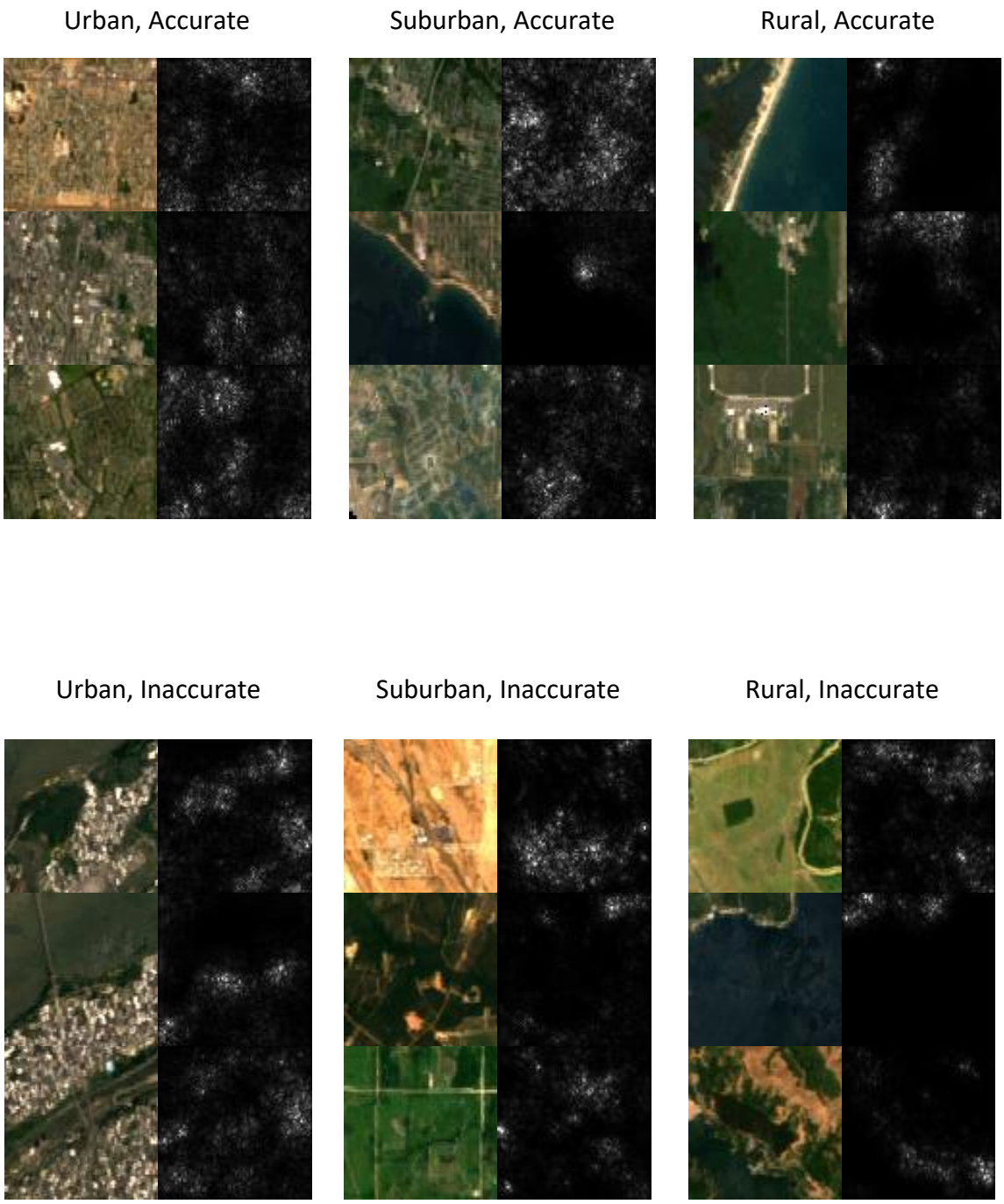
Appendix Figure 2 shows several saliency maps for images in urban, suburban, and semi-rural environs and for which model predictions of income in levels are accurate and inaccurate. Examining cases in which the model performs well and poorly at each population density level gives context on the types 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.

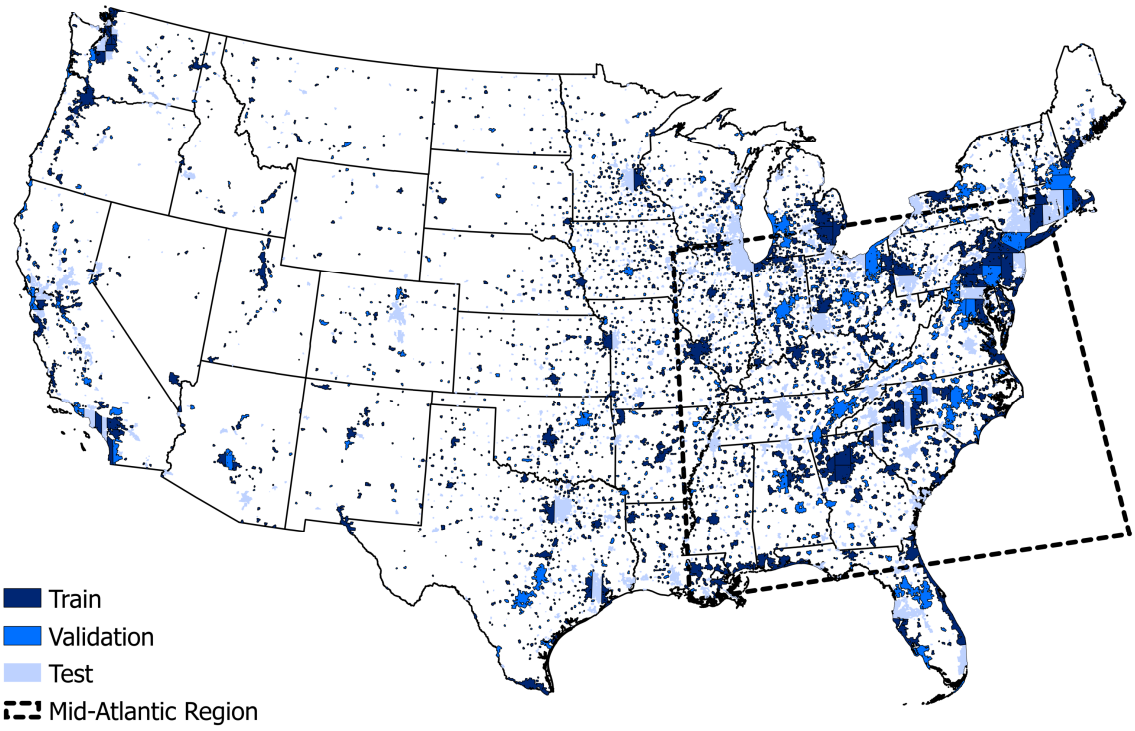
Appendix Figure 1: Convolutional Neural Network, Landsat Imagery Model Architecture



Appendix Figure 2: Selected Saliency Maps

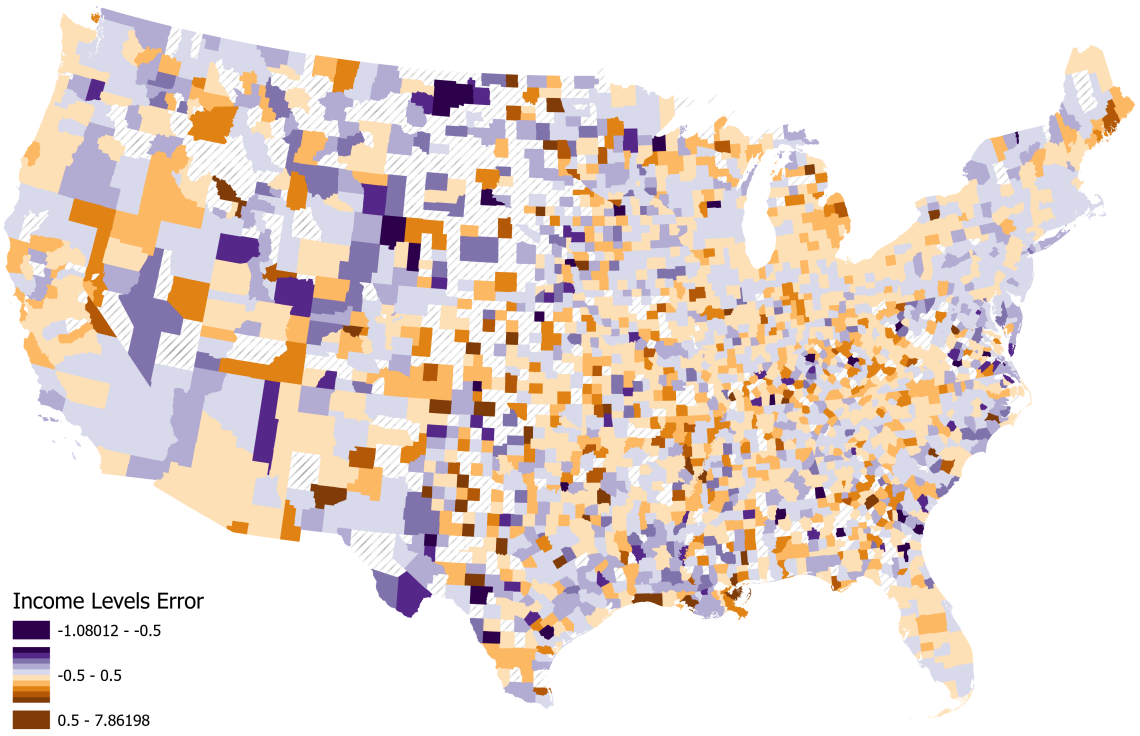


Appendix Figure 3: Spatial Extent of Urban Areas and Model Development Subsets



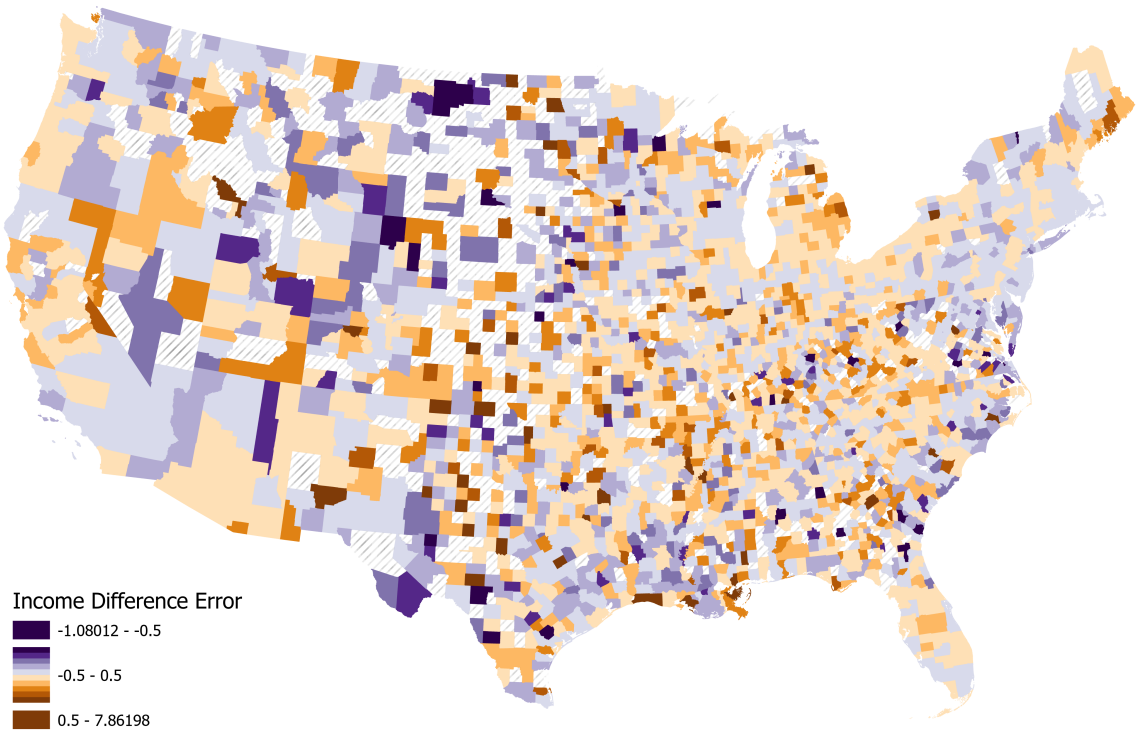
Notes: This map shows the urban areas of the contiguous Unites States used to assign images into training, validation and testing subsets for our CNN models. The black dotted line represents the Mid-Atlantic region used for our high-resolution imagery robustness tests discussed in Section **??**. The 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



Notes: 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



Notes: 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

Notes: 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

Notes: 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

Notes: 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

Notes: 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

Notes: 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.