Using Publicly Available Satellite Imagery to Address Data Scarcity Problems in Developing Economies

Cooper Nederhood

University of Chicago

Abstract

Past studies have used nighttime luminosity as a proxy for economic activity while other studies have relied upon proprietary high-resolution satellite imagery to estimate economic conditions. We exploit the rich Landsat 7 imagery, publicly available since the 1970s to estimate ground truth wealth index in sub-Saharan Africa, as measured by the DHS asset index. We employ a convolutional neural network (CNN) to analyze the high-dimensional data and train models both via transfer learning and from random initialization. In all cases, we find reasonable explanatory power despite limitations in model training capacity and the medium resolution satellite imagery. These results suggest numerous avenues for future research.

Keywords: computer vision, CNN, satellite imagery

1. Introduction

New technologies and methodologies have led to a recent data revolution for governments, researchers, and private companies in developed nations. However, many developing economies are unable to take advantage of these advancements, resulting in a growing data divide. Many developing countries still lack reliable information on socio-economic and health outcomes. For example, during the years 2000 to 2010, 39 of 59 African countries conducted fewer than two surveys of national poverty measures. Further, 14 countries conducted no such survey. This data scarcity is often most acute in the poorest and most remote regions, precisely the populations of greatest concern. Even when official measures are conducted, the quality or coverage can be dubious. Finally, informal economies, often of greater significance in developing economies, can remain unseen by official statistics. Making

informed policy decisions relies on the presence of high quality data and ambitious initiatives like the UNs Millennium Development Goals highlight the need to develop accurate and cost-effective statistics for even the most under-represented regions. [1]

While satellites have been circling the Earth and acquiring data since the 1970s, only recently have developments in computer vision, engineering, and geography as well as partnerships between governments and private companies led to this rich data being publicly available and accessible to researchers. The richness of the data, in terms of the information encoded in the variety of bands, the spatial and temporal granularity, and the nearly worldwide geographic coverage offers many possibilities for research, including addressing data scarcity issues in developing economies. However, because of the high-dimensionality and unstructured nature of the data, many of these advantages can only be fully attained by exploiting state-of-the-art approaches in computing and algorithmic design. Just this December, the United Nations Task Team on Satellite Imagery and Geospatial Data published a guide for all National Statistical Offices which provides a brief introduction to the use of earth observations (EO) data for official statistics. Satellite imagery can be used to fill in gaps where no data exists and to improve the spatial and temporal resolution of current measures. Because of the different types of bands, satellite imagery can be used to improve measures across a variety of areas including statistics on agriculture, environment, business activity, and transport. Finally, after initial development costs, satellite imagery techniques scale far more efficiently than tradition designed data sources like censuses and surveys, and at a fraction of the cost. By using modern techniques in computer vision we show that these rich, high-dimensional satellite images can be used to estimate ground-truth economic conditions in regions with otherwise infrequent economic statistics. [2] [3]

2. Theory and Literature

The most common satellite data used by economists is nighttime luminosity (i.e. nightlights). Since the original study using nighttime luminosity as a proxy for economic activity by Henderson et al nightlights have been used in a host of research, far too much to recall here. [4] The widespread use

of nightlights data highlights the potential for fully exploring much richer datasets like Landsat 7. The nightlights data has a 1km resolution, while Landsat has 30m resolution. The nightlights maps each pixel to a 1-63 value while Landsat has 8 continuous bands (not to mention all the other available satellites with different bands). While there have been many novel uses of satellite data, the potential of satellite data to measure economic conditions is incipient. The vastness of the research using nightlights as a proxy for economic activity despite the coarseness of the luminosity data illustrates the potential of using a more fine-grained dataset and modern methods. Most research employing satellite imagery simply uses the pixel band values as a data input, without regard to any surrounding pixels. Thus, approximations of vegetation or urbanization can be calculated over a region which is correlated with economic outcomes, but the actual features of the image are lost. For example, a normalized vegetation index (NVI) cannot distinguish between a gray pixel that is a part of a rock outcrop or part of an interstate highway system. A convolutional neural network (CNN) is a modified deep learning framework which has had recent success in computer vision object detection tasks. Because of the high dimensionality of images, having many fully connected layers would lead to an intractable number of parameters to estimate (even by machine learning standards). Most CNNs begin with a partially connected (convolutional) layer which reduces dimensionality and allows for the location of the target object within the image to vary, critical for analyzing satellite imagery. The first layers of a CNN typically learn to recognize basic features, like edges and color blobs. The features increase in complexity with each layer, until the last layer, a fully-connected layer, maps the resulting feature-vector to a given classification category. Critically, training CNNs require massive amounts of labeled training data. However, because early layers of CNNs identify low-level features like edges models pre-trained on vast datasets like ImageNet can be modified on new tasks such that only later layers need retraining. This transfer learning approach is common practice in computer vision tasks and is central to estimating economic outcomes where ground-truth data is usually in short supply. [5]

Jean et al [6] are leaders in using CNNs to extract economic information from satellite imagery and our results rely heavily on their methods. They use two stages of transfer learning to estimate economic expenditure and wealth indices in five countries in sub-Saharan Africa. However, Jean et all use satellite imagery from Google Static Maps which, while high-resolution,

is proprietary and is thus only made public for the current time period. Our current analysis will use only data entirely within the public domain and thus all data sources are available extending for years into the past, providing a path for future studies regarding time-series properties.

Banergee et all [7] , a direct extension of Jean, analyzes development in various sub-regions of India, which provides an interesting test case because of the geographic and economic diversity. They directly challenge Jeans choice to train the CNN on the noisy nightlights data before using the CNN as a feature extractor. Rather, they show that a direct regression of asset indicators gives superior R-squared when compared to a transfer-learning via nightlights approach. They maintain the first transfer learning step, again beginning with the VGG-F model pre-trained on ImageNet. In keeping with this approach, our current analysis will compare both transfer learning approaches and random initialization models.

Engstrom et al [8] , another direct reply to Jean, use CNNs to measure poverty from high-resolution private imagery of Sri Lanka. Because of their high-resolution images (< 5m) they do not need the second transfer learning on nightlights step. They can also extract extremely fine-grained features, like the number of cars in a parking lot, which is not possible with the medium-resolution Landsat images I intend to use. However, they are unable to attain nationwide coverage because of the high cost of purchasing the data. While their research illustrates the bright future as costs decrease and resolution increases, currently they cannot properly scale their approach, limiting its applicability. Their data also does not extend 40 years into the past like the Landsat data, which remains the gold standard in terms of resolution and spatial and temporal coverage.

Finally, these computer vision methods are also increasingly applied to urbanization questions, which often has development applications. Machine learning excels at classification tasks which is especially applicable when classifying land usages or mapping urban form. For example, Albert et al [9] use Google Static Maps and two different CNN architectures to extract features relevant to land-use classification, comparing urban formation across European cities. Given the importance of urbanization to economic outcomes in the both developed and developing economies, these methods will become increasingly related.

3. Data

Our analysis uses two sources of satellite imagery which are cleaned and prepared for analysis within the Google Earth Engine platform, a free platform and IDE written in javascript combining catalogs of publicly available satellite imagery with geospatial analysis capabilities. We rely on the Demographic and Health Surveys (DHS) for our ground-truth economic data.

3.1. Nighttime luminosity

The Defense Meteorological Program (DMSP) Operational Light Scan (OLS) program can detect visible and near-infrared light emissions at night. The stable lights band represents a cleaned average light value emitted by cities, towns, and other sites with persistent lighting. Noise caused by events like fires and background noise are removed. At 1km resolution, the nighttime luminosity data is one of the coarser satellite derived products. Each 1km pixel takes on an integer value from 0-63, with 0 indicating no light and luminosities above 63 capped. Thus, the extremes are subject to measurement distortions. We utilize yearly composites, specifically the 2013 composite to coincide with the year of the DHS survey data. Similar data is available every year going back to 1992. Immediately below, we show the resulting luminosity image from our area of interest.

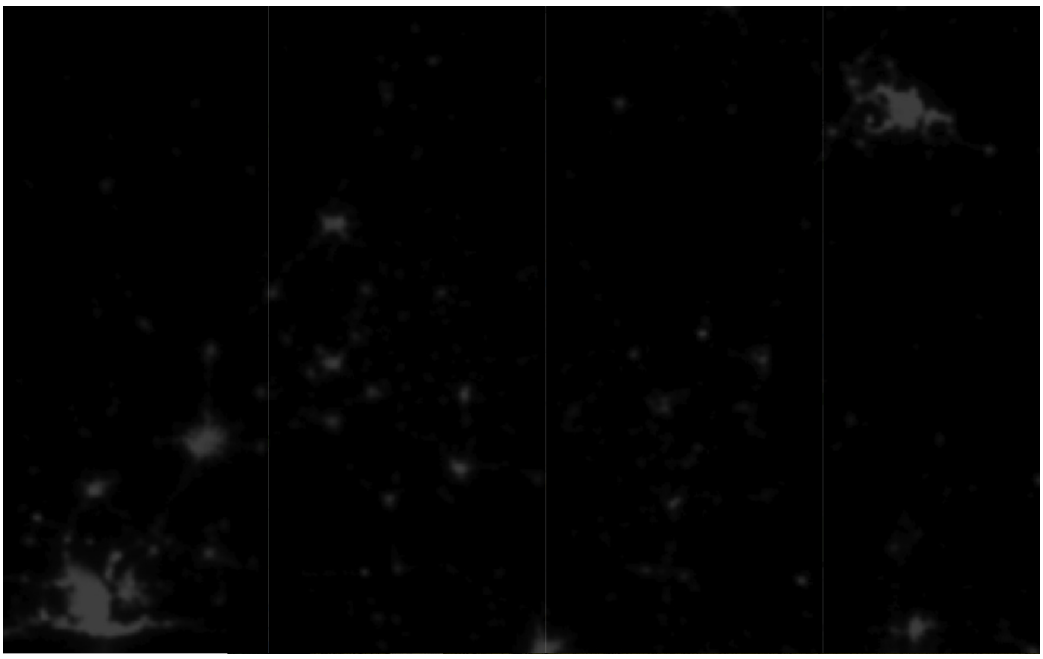


Figure 1: Nighttime luminosity across Nigeria

3.2. Landsat 7

The Landsat series of satellites is the gold standard in publicly available satellite imagery with respect to its resolution, the available bands, and its temporal and spatial coverage. The Landsat 7 series is available from 1999 to 2018 with 30m resolution in all bands. Earlier Landsat series go back to the 1970s. Again, we utilize the 2013 images. There are 8 bands ranging from the RGB (red-green-blue) visible spectrum to infrared and pan-chromatic. For our current analyses, we use the visible RGB spectrum. The Landsat series captures an image of the same region roughly every two weeks. To match the temporal resolution of our luminosity and DHS data we composite the Landsat 7 images into a single cleaned yearly composite. This mitigates any noise caused by cloud cover and any discontinuities within the image collection process. Immediately below, we show the resulting 3 band image from our area of interest. Note, the satellite image is very dark, affecting viewability but a computer does not have any issue in discerning the numerical values given that the visual rendering for humans is arbitrary. While the image is dark it does not affect the underlying information. In the bottom-left corner you can discern the splotch that is Lagos and the Niger river is also visable

cutting through the image north to south. Bright areas in the luminosity image correspond to the slightly lighter areas in the Landsat image, reflecting man made structures rather than vegetation.

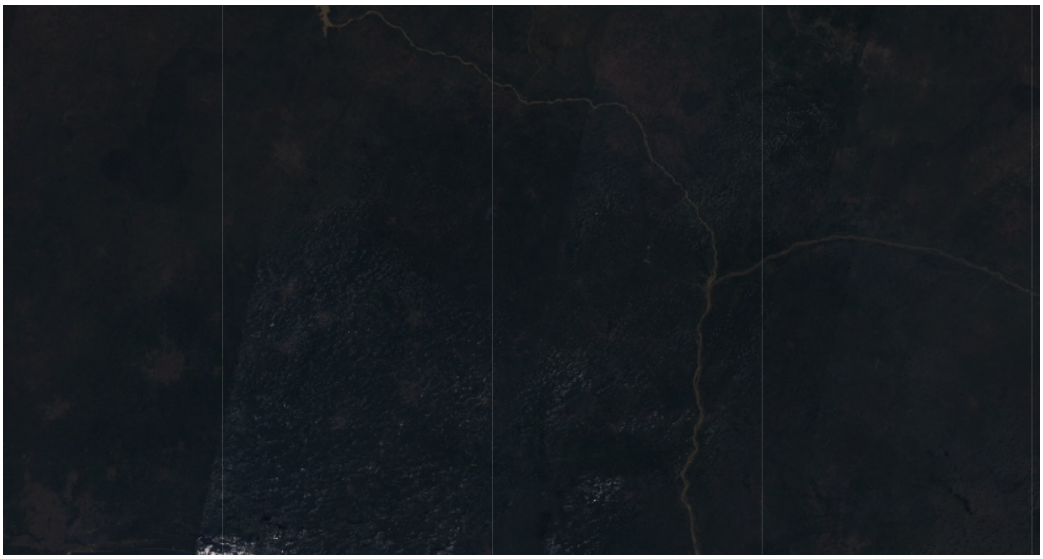


Figure 2: Landsat 7 across Nigeria

3.3. DHS survey data

Any analysis of satellite imagery needs some ground-truth value for training. The Demographic and Health Surveys from US Aid conducts detailed health surveys of households all around the developing world, with wide coverage in sub-Saharan Africa. We select the region surrounding the city of Lagos in Nigeria. We use the 2013 data as it is the most current, although similar data is also available for 1990, 2003, and 2008. Lagos has grown tremendously over recent years and is now the most populous urban area in the African continent. It is a major financial center and one of the busiest ports on the continent. The surrounding area is also home to many smaller cities and towns as well as the Niger river, thus providing a rich set of features for our neural network to learn from. While the DHS data is mostly health information they do record questions relating to assets like housing material, access to plumbing, etc. The DHS aggregates all of the asset information into an asset index, which is the first principal component of the asset questions. Below we provide the map of our target area with black dots representing

the 279 cluster survey locations of the DHS data. This includes the megacity of Lagos in the southwest corner as well as the cities of Ibadan and Abuja.

The DHS survey is conducted at the household level with households grouped to a specific cluster. Each cluster is then geo-coded via GPS coordinates with noise added to preserve anonymity of the respondents. Thus, the cluster level is the smallest unit of analysis recommended by the DHS. Both the survey data and the GPS data are available from DHS via request, with the GPS data requiring an additional permission. To connect the satellite imagery with the GPS data we parse the GPS shapefiles and extract the coordinates, then bring the coordinates into Google Earth Engine, constructing a new satellite image with flags denoting the locations of the DHS cluster. Failure to correctly tie the DHS data with the appropriate satellite imagery would invalidate the entire exercise.

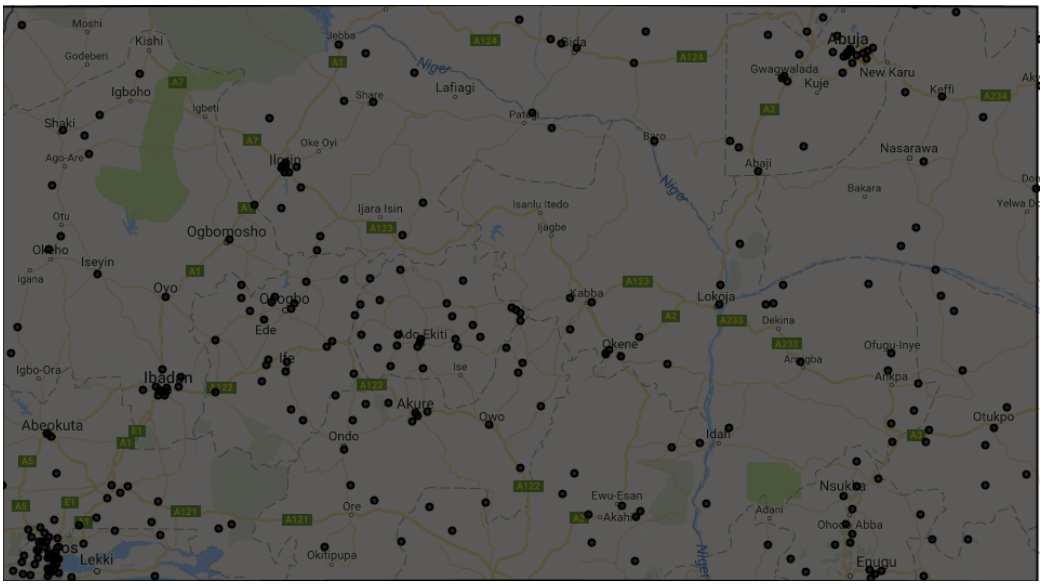


Figure 3: Map of 279 DHS cluster locations

4. Model training

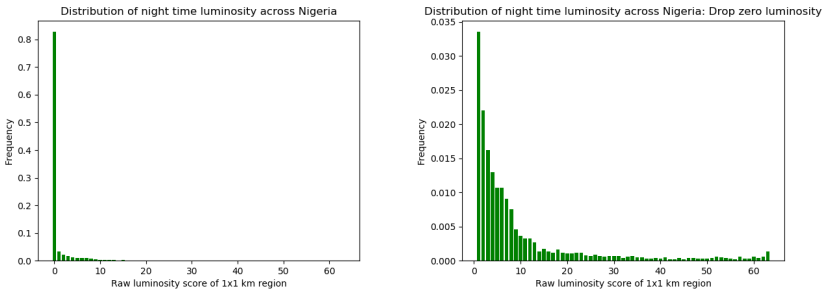
The ultimate goal is to build a model that can predict the ground-truth economic level over a region given the corresponding Landsat 7 image. However, as we will show the CNN required by the high-dimensional image data has a huge numbers of parameters, with some designs having over one million

parameters to estimate. Even with the rich DHS survey coverage, there are only 279 sampled clusters within our area of interest. Thus, to circumvent the data scarcity we construct an intermediate task that is similar to the final task but that is data-rich rather than data-poor. As discussed, nighttime luminosity is a common proxy for economic activity. Thus, we would expect that a model that learns to predict nightime luminosity given daytime satellite imagery has also learned to identify features that would be correlated with the estimation of ground-truth asset levels. We can then retrain the model that identifies nighttime luminosity to identify the DHS wealth index. This process whereby a model is pre-trained on a data-rich task then used as a basis for a related task is called Transfer Learning and is commonplace in computer vision tasks.

4.1. Training task 1: Landsat to luminosity

Our first transfer learning step is to train a CNN to predict nighttime luminosity given daytime Landsat imagery. The region of interest surrounding Lagos is a 323 km by 584 km region represented by 210 million pixels (for each of the 3 RGB bands). Because the resolution of the luminosity data is 1km and the Landsat resolution is 30m each luminosity value corresponds to a 34x34 pixel image of Landsat data. To build the training data, we therefore randomly draw 60,000 points from our region, pairing the luminosity level with its corresponding Landsat image. We partition the 60,000 points into 30,000 training points, 15,000 validation points, and 15,000 test points.

While the luminosity data takes on integer values from 0-63, in keeping with Jean et all we bucket the luminosity scores into 3 buckets, reflecting low, medium, and high luminosity. As the full-scale luminosity image shows, most of the region is dark, corresponding to a luminosity score of zero. Thus, our 3 buckets are extremely imbalanced, making the learning task more difficult as our baseline is already quite high. Immediately below we show a histogram of the luminosity of our sampled points both with the 0 luminosity scores included and with them excluded. Once we exclude the 0 luminosity scores it is easier to discern the general 3 bins.



(a) Raw luminosity scores (b) Luminosity scores dropping 0 scores

Figure 4: Luminosity is highly skewed across low, medium, and high

Most CNNs are not trained completely from scratch but rather start with a convolutional base trained on ImageNet, a dataset of millions of labeled images. Nearly any image recognition task relies on early layers in the CNN learning to identify low-level features like edges and basic color patterns. Thus, the early layers of CNNs trained on ImageNet can be used in almost all vision tasks. For example, Jean et all begins with a CNN trained on VGG-8. However, because of the Landsat imagery is medium resolution rather than high resolution exact image detection is not possible. We can identify general regional differences but cannot identify specific objects likes roads. This raises potential questions as to whether a model trained on ImageNet is sufficiently similar to the task at hand. Further, there is the question of whether a 34x34 pixel image even contains meaningful features. Jean et all, with its higher resolution proprietary data, has over 200x200 pixel images, which is similar to the resolution seen in most ImageNet tasks. However, given that the crude luminosity data has proven predictive of economic activity and the Landsat image has 34x34 pixels for each luminosity data point, there is almost certainly a meaningful signal to extract. To address the resolution concerns, we also construct another 60,000 image sample but rather than a 34x34 region we extract a 128x128 region. By zooming out perhaps we will regain some resolution and thus be able to extract more meaningful features. By comparing the performance of various models we can also gain some insight into how the CNN process is working. Thus, we actually train 3 different model structures for the initial CNN task:

1. **Small**: 34x34 images trained from scratch
2. **Large**: 128x128 images trained from scratch

3. **Transfer**: 128x128 images trained with weights from VGG16 (trained on ImageNet)

We cannot train a small model via VGG initialization because the 34x34 image size is simply too small for the VGG convolutional base. For details on the layer structure of each model see the Appendix.

Training the hundreds of thousands of parameters in each of the above models requires computation far beyond a personal computer. Thus, I employ Google Cloud Computing, specifically an instance featuring 3 NVIDIA Tesla K80 GPUs, for the training. The Keras framework used to construct the models automatically assigns tasks to any available GPUs without the need for further implementation. When training the models, we use a batch size of 100 images and 300 steps per epoch (100 x 300 = 30,000 training images). The model explicitly chooses weights which minimize the error on the 30,000 training images and to estimate the out-of-sample fit after each epoch we evaluate the predictive power on the 15,000 validation points. Finally, upon completion of model training we evaluate the performance on the 15,000 test points, which are previously completely unseen by the model. To prevent overfitting and to mediate unnecessary training, we implement an early termination condition, whereby the model training terminates if the validation error is not decreasing for two epochs. The validation accuracy and the validation loss for each of the three models is shown below

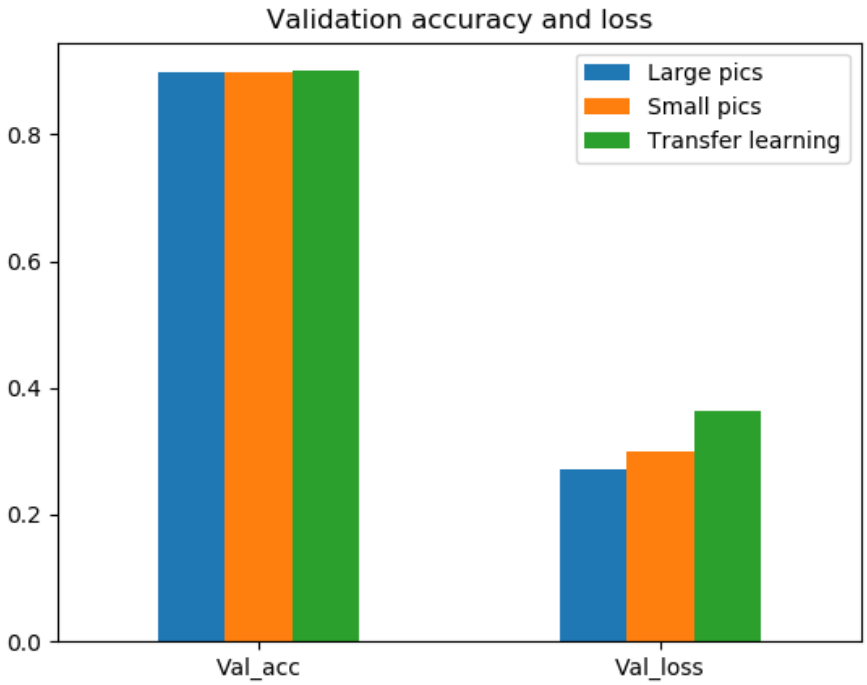


Figure 5: Validation acc and loss of trained CNN

4.2. Training task 2: Landsat to DHS

Once we have trained the CNN to predict nighttime luminosity based on the daytime Landsat image, we can move to the ultimate task of predicting the ground-truth DHS asset index. Nearly every CNN has fully-connected layers in its final layers. After layer upon layer of convolutions, the last layer maps a fixed set of lower-dimensional features to a corresponding class, in this case low, medium, and high luminosity. The final vector of lower-dimensional features essentially summarizes all information contained in the original high-dimensional input image that is deemed relevant to luminosity. Thus, we can discard the final layer of each CNN and use it as a feature extractor. Specifically, if we discard the last layer then pass in a new Landsat image the CNN will return vector of length 512 summarizing the input image. Because nighttime luminosity is correlated with economic activity these 512 features may also be informative when predicting the DHS asset index. In the second training task we use these feature vectors to predict the

ground-truth DHS asset index.

Ideally, we we would simply train another neural net but given the small DHS sample size this is not feasible. Therefore, we use the resulting feature vectors as the explanatory variables in a ridge regression on DHS asset index. Each of the 279 DHS clusters has a corresponding GPS location. However, there is up to 10 km of noise added to this GPS location to preserve anonymity. Therefore, for each of the 279 GPS clusters we extract the corresponding 10x10km Landsat image. We then tile this 10x10km image into 34x34 or 128x128 pixel images and pass each image collection through the appropriate CNN. This yields a set of feature vectors for each cluster which we then average, providing a single final feature vector for each DHS location. We then use the resulting 279 feature vectors in the left-hand-side of a ridge regression on DHS asset index. Because we have more features than observations the linear regression procedure is prone to overfitting. The ridge regression enforces squared penalties on the coefficient estimates, thus penalizing model complexity. We also need to find the optimal value of the alpha parameter. By performing 10-fold cross-validation across various values of alpha we can find the optimal parameter value with minimal fears of overfitting. Below we report the cross-validation R-squared corresponding to each of our three CNN structures. For reference, Jean et all finds cross-validation R-squared of about .70. While our findings of roughly .45 are much lower, we do have more coarse grained satellite imagery and, as discussed in the limitations section, there are various improvements to be made here for better predictive power. The current estimates should be viewed as a lower bound on the predictive power.

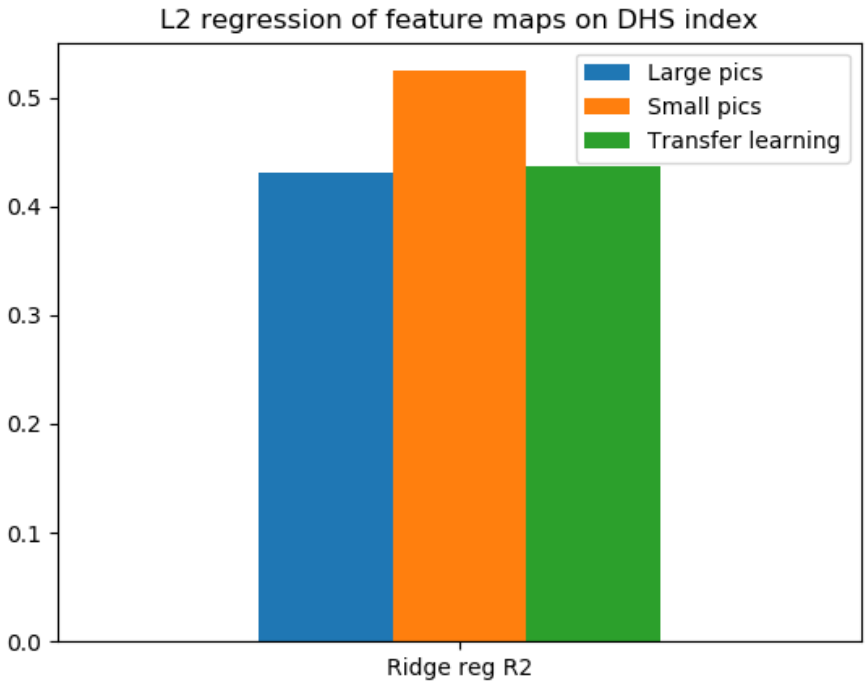


Figure 6: Ridge regression R2 of feature vectors from each trained CNN on DHS asset index

5. Discussion and Limitations

The most immediate limitation is in the incredibly unbalanced luminosity data. We have investigated stratified sampling to force a more balanced distribution, but doing so required categorizing each of the millions of pixels. The initial naive algorithm for this was intractable and a more efficient algorithm needs to be developed. Alternatively, we are investigating an asymmetric loss function. Adjusting for the unbalanced data would make the learning task easier and would allow for more straightforward assessment of performance. Currently, given that over .80 of luminosity is in the low luminosity bin, our baseline accuracy is already quite high. Mitigating this concern is that despite the high baseline, the declining validation loss function is evidence that the model is in fact learning.

The structure and training of the CNNs is also almost certainly sub-optimal. Given the vast number of engineering decisions behind CNN training, accuracy can surely be increased by the continued testing of various model structures and parameter settings. Some of these limitations are a result of the Keras platform, which is the most high-level neural network platform. For example, the VGG16 convolutional base is a 16 layer model which is far too deep for the current task. Jean et all use a VGG8 base, which has only 8 layers. However, the VGG16 model is the only pre-trained model available through Keras compatible with the given 128x128 image size. Similarly, there is no available pre-trained model for the smaller 34x34 images. By moving to a lower level language like PyTorch we can regain flexibility with respect to the neural network construction and have access to the wealth of pretrained models in Cafe. Additionally, despite the NVIDIA-enabled Google Compute Engine, each training took from 1-2 hours of compute time. This severely limited the ability to run different levels of dropout and data-augmentation. Each of these helps protect against overfitting which could then allow for greater model complexity and thus greater predictive power. Greater computing power would allow for enhanced testing of the wealth of tunable parameters in a CNN.

6. Conclusion and Future Research

Given the incipient nature of the task, these modest results suggest numerous courses for future work. As discussed, all of this data is available historically. The data-rich first training task, estimating nighttime luminosity from daytime images, can be repeated for each year going back to 1999. Interestingly, the data will become a georeferenced panel data where each geographic region has a corresponding sequence of yearly image pairs. This structure of a time-indexed sequence of images is how a video is structured. Thus, rather than using image classification techniques a time-series approach could use video classification techniques, with each year viewed as a frame in the video. To my knowledge, the video classification methods have not yet been applied to the economic development literature.

Ultimately a measurement tool must be applied to a specific task. By developing a time-series approach, the fine-grained estimates of economic conditions could be used to evaluate policy changes in the developing world. The more granular and accurate the satellite-derived measures become the more

useful they will be in constructing counter-factuals and in policy evaluation.

Appendix A. CNN Layer Specifications

MODEL: 34x34 pictures, random initialization

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 32, 32, 32)	896
max_pooling2d_1 (MaxPooling2	(None, 16, 16, 32)	0
conv2d_2 (Conv2D)	(None, 14, 14, 64)	18496
max_pooling2d_2 (MaxPooling2	(None, 7, 7, 64)	0
conv2d_3 (Conv2D)	(None, 5, 5, 64)	36928
flatten_1 (Flatten)	(None, 1600)	0
dense_1 (Dense)	(None, 512)	819712
dense_2 (Dense)	(None, 3)	1539

Total params: 877,571
Trainable params: 877,571
Non-trainable params: 0

Figure A.7: CNN structure for small 34x34 image model trained from scratch

MODEL: 128x128 pictures, random initialization

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d_1 (MaxPooling2	(None, 63, 63, 32)	0
conv2d_2 (Conv2D)	(None, 61, 61, 64)	18496
max_pooling2d_2 (MaxPooling2	(None, 30, 30, 64)	0
conv2d_3 (Conv2D)	(None, 28, 28, 128)	73856
max_pooling2d_3 (MaxPooling2	(None, 14, 14, 128)	0
conv2d_4 (Conv2D)	(None, 12, 12, 128)	147584
max_pooling2d_4 (MaxPooling2	(None, 6, 6, 128)	0
flatten_1 (Flatten)	(None, 4608)	0
dense_1 (Dense)	(None, 512)	2359808
dense_2 (Dense)	(None, 3)	1539

Total params: 2,602,179
Trainable params: 2,602,179
Non-trainable params: 0

Figure A.8: CNN structure for large 128x128 image model trained from scratch

MODEL: **128x128 pictures, VGG convolutional base**

Layer (type)	Output Shape	Param #
vgg16 (Model)	(None, 4, 4, 512)	14714688
flatten_5 (Flatten)	(None, 8192)	0
dense_9 (Dense)	(None, 512)	4194816
dense_10 (Dense)	(None, 3)	1539

Total params: 18,911,043
Trainable params*: 11,275,779
Non-trainable params*: 7,635,264

*NOTE: fully-connected layers added on top of VGG16 are trained first while fixing the convolutional base. Then, the last convolutional layer is un-fixed and finetuned with a slower learning rate

Figure A.9: CNN structure for large 128x128 image model trained from VGG16 convolutional base

[1] L. G. Morales, Y.-C. Hsu, J. Poole, and B. R. I. Rutherford, "A world that counts: Mobilising the data revolution for sustainable development," *United Nations Expert Report*, 2014.

[2] "Big data for development: Challenges and opportunities," *UN Global Pulse*, May 2012.

[3] "Earth observations for official statistics," *United Nations Satellite Imagery and Geospatial Data Task Team Report*, Dec 2017.

[4] V. Henderson, A. Storeygard, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[5] N. Jean, M. Xie, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *Association for the Advance of Artificial Intelligence*, 2016.

[6] N. Jean, M. Xie, M. Burke, D. Lobell, S. Ermon, and M. Davis, "Combining satellite imagery and machine learning to predict poverty," *Science*, 2016.

[7] S. Banergee, S. B. Paul, M. Sharma, A. Gupta, and P. K. Suraj, "On monitoring development using high resolution satellite images," *Working Paper*, Dec 2017.

[8] R. Engstrom, J. Hersh, and D. Newhouse, "Poverty from space - using high resolution satellite imagery for estimating economics well being," *World Bank Group Poverty and Equity Global Practice Group*, Dec 2017.

[9] A. Albert, J. Kaur, and M. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale," *arXiv pre-print*, Aug 2017.