On monitoring development using high resolution satellite images

Potnuru Kishen Suraj*  Ankesh Gupta†‡  Makkunda Sharma†‡  Sourabh Bikash Paul§

Subhashis Banerjee†

May 4, 2018

Abstract

We develop a machine learning based tool for accurate prediction of development and socio-economic indicators from high resolution day-time satellite imagery. The indicators that we use are derived from the Census 2011 [The Ministry of Home Affairs, Government of India, 2011] and the NFHS-4 [The Ministry of Health and Family Welfare, Government of India, 2016] survey data. We use a deep convolutional neural network to build a model for regression of asset indicators from satellite images. We show that the direct regression of asset indicators gives superior $R^2$ scores compared to that of transfer learning through night light data, which is a popular proxy for economic development used world wide. We also use the asset prediction model for accurate transfer learning of other socio-economic and health indicators which are not intuitively related to observable features in satellite images, or are not always well correlated with each other. The tool can be extended to monitor the progress of development of a region over time, and to flag potential anomalies because of dissimilar outcomes due to different policy interventions in a geographic region by detecting sharp spatial discontinuities in the regression output. Finally, we demonstrate the utility of our tool by carrying out a case study of regression analysis using predicted variables to understand the determinants of high rate of stunting in India.

Index terms: Satellite images, socio-economic and development indicators, regression, prediction, machine learning, deep learning, CNN.

1 Introduction

In a country like India, where there is a paucity of reliable and high frequency data, evidence based design of policy interventions that are grounded on accurate estimates of economic and development indicators are difficult. Census data collection [The Ministry of Home Affairs, Government of India, 2011] for the 1.2 billion population is cumbersome and expensive, and is carried out infrequently only about once in a decade. Census is also error prone and noisy due to the large variability in the data collection processes across the geography, and there is often no validation [Brown, 1971; Vemuri, 1994; Bose, 2008]. Smaller sample surveys [The Ministry of Statistics and Programme Implementation, Government of India, 2017; The Ministry of Health and Family Welfare, Government of India, 2016] tend to be more accurate, but they too are infrequent, and, in general, they do not comprehensively address all aspects of the economy.
Availability of reliable estimates of economic and development indicators at high frequency may enable policy planning in several ways - for example in block level development planning, addressing local issues related to agriculture, health, sanitation, education, employment, connectivity, resource management etc. Traditionally, extrapolated data estimates or inferential exercises are sometimes used for such purposes [Barro and Lee, 2013]. In this paper we explore the prospect of using predicted data as a viable alternative to using real data [Weide and Fujii, 2016].

Advances in computational facilities and Geographic Information Systems make it possible to predict some important geographic characteristics at a sub-district level [The Ministry of Science and Technology, Government of India]. We investigate the possibility of accurate regression of economic indicators, at the village level, from high resolution daytime satellite imagery, which can be acquired at low cost, accurately and frequently. We carry out our analysis on input images that cover a ground area of 7 Km² on the average, which is at a resolution higher by at least two orders of magnitude compared to what has been attempted in the past for machine learning based econometric analysis [Jean et al., 2016].

We carry out village level analysis in six Indian states - Punjab, Haryana, Uttar Pradesh, Bihar, Jharkhand and West Bengal. The problem is challenging not only because of large variability in language, culture and living styles across these states, but even the geographical diversity is significant. The topography varies from the agricultural plains in Punjab and Haryana to the Indo-Gangetic plains in the middle to the dense forests in parts of Jharkhand to the Chota Nagpur plateau to the mountains in the north and the Sunder Bans delta in West Bengal. The differences, both in the socio-economic patterns and the visual characteristics, make the regression problem interesting. Recent advances in deep convolutional neural networks (CNNs) [Krizhevsky et al., 2012; Goodfellow et al., 2016] and GPU computing make the nonlinear regression problem addressable. We investigate both the accuracy of regression and what makes it possible.

Specifically, the main contributions of our paper are as follows:

1. We train a eight layer deep convolutional neural network based model (VGG CNN-S) [Chatfield et al., 2014] for direct regression of census asset indicators from high resolution daytime satellite images of a village. We obtain superior regression scores when compared to transfer learning [Pan and Yang, 2010] from regression of night light data, a popular proxy economic indicator suggested in the literature [Jean et al., 2016].

2. The regression model learnt from the data of 218000 villages across six states smooths out the error in the census data. This not only gives us accurate prediction of the census indicators, but we also obtain a census validation tool as a by-product.

3. We show that the asset prediction model can be effectively used for transfer learning [Pan and Yang, 2010] of other socio-economic indicators for literacy, education, health, and under-privileged social categories.

4. We show that though the prediction model is trained with cross-sectional data, it can still facilitate monitoring development progress of a region over time.

5. We show that spatial discontinuities (or sharp gradients) in the regression output over a geographical region can potentially indicate dissimilar outcomes of different policy interventions, and can serve as effective alerts for further investigation.

6. Finally, we demonstrate the utility of our tool by carrying out a case study of regression analysis using predicted variables to understand the determinants of the high rate of stunting in Indian children. We show that the predicted variables are helpful to avoid omitted variable biases.
The rest of the paper is organized as follows. In Section 2 we briefly discuss the considerations for using predictive machine learning models for econometrics. In Section 3 we discuss some prior work that use satellite images for econometric analysis. In Section 4 we briefly describe our data sources. In Section 5 we discuss our CNN based regression framework for regression using transfer learning from might light data and direct regression of the asset model. In Section 6 we discuss transfer learning of other socio-economic parameters. In Sections 7 and 8 we discuss monitoring development over time and spatial discontinuities in regression output respectively. In Section 9 we present our case study of regression analysis to understand the determinants of stunting.

2 Machine learning, prediction and causation

Our machine learning based regression model is predictive and it does not directly support any causal inference. However, as has been pointed out in [Kleinberg et al., 2015], deciding on effective policy interventions require both prediction and causation. We reproduce the argument provided in [Kleinberg et al., 2015] below for clarity. Suppose \( Y \) is an outcome variable (such as literacy rate in a village) which depends in some unknown way on variables \( X_0 \) and \( X \) where \( X_0 \) is a policy intervention, and our objective is to maximize a known payoff function \( \pi(X_0, Y) \). Then, the decision \( X_0 \) depends on the derivative

\[
\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial\pi}{\partial X_0}(Y) + \frac{\partial\pi}{\partial Y} \frac{\partial Y}{\partial X_0}
\]

The payoff function is known but evaluation of the first term requires accurate prediction of \( Y \), whereas estimating \( \frac{\partial Y}{\partial X_0} \) in the second term requires causal inference (how much does \( X_0 \) affect \( Y \)).

Thus, for maximizing the payoff, we need to do both - accurately predict the literacy rate in a village, and evaluate how a policy intervention may affect literacy. In some cases, for example when the outcome is an outbreak of a disease like cholera in a village, where the causal relations and the policy implications are already reasonably well understood, the payoff maximization crucially depends on accurate and timely prediction of a possible outbreak. Accurate predictions can not only have large policy impacts but can also provide crucial theoretical and economic insights [Kleinberg et al., 2015], can offer the possibility of using predicted variables when real measurements are not available [Weide and Fujii, 2016], and can be used to guard against omitted variable biases in econometric analysis [Barreto and Howland, 2006]. Machine learning techniques can also be used in econometrics to accurately predict an endogenous variable at the first stage of a linear instrumental regression model [Mullainathan and Spiess, 2017]. [Athey and Imbens, 2017] review some innovative applications of machine learning methods in causal econometric inferences.

The value of accurate prediction has been underemphasized in conventional econometrics. There are two main reasons. First, standard empirical methods are often based on small or infrequent sample surveys which provide inadequate data for accurate prediction estimates, and, consequently, the extrapolations are usually noisy. Building highly accurate predictive models will remain difficult till progress in digitization can facilitate online, real-time gathering of high volume transactional socio-economic and health data. Second, standard linear regression techniques like OLS, which minimize the in-sample error, are not best suited for prediction problems because of their emphasis on reducing bias at a cost of prediction accuracy [Kleinberg et al., 2015]. They typically under-fit the data. In contrast, modern machine learning techniques like deep CNNs [Krizhevsky et al., 2012; Chatfield et al., 2014] explicitly address the bias-variance tradeoff by a) using a nonlinear representational network which can model highly complex functions b) focusing on minimizing out of sample prediction error using \( n \)-fold cross validation during the training phase c) incorporating regularization functions in the training optimization model that minimizes variance and d) selectively using
techniques like dropout to avoid over-fitting [Hastie et al., 2009; Goodfellow et al., 2016]. Hence, they are tuned for accurate prediction though the interpretability of the trained network model is low.

Despite these advantages machine learning models have found limited use in econometrics mainly because of their large data requirement. However, CNN based prediction using satellite images provide more opportunities. For example, the deep CNN model VGG CNN-S [Chatfield et al., 2014] that we use in this work has approximately 138 million parameters and has been trained to recognize 1000 image categories using a training set of 1.3 million labelled images, a validation set of 50,000 images and a test set of 100,000 images. Such a convolutional network, pre-trained with large volumes of data, has already learned complex image representations in terms of highly discriminatory image features, which enables fine tuning the model for regression of economic indicators using a relatively small number of images.

Another significant point of departure of conventional econometric regression and deep CNN based regression manifests in the method of choosing the explanatory variables to predict the outcome. In both conventional econometrics and traditional machine learning the explanatory variables, or the features used for regression, are usually hand-crafted, and are measured from the data using well specified procedures. This makes the explanatory variables clearly identifiable, and the importance of each explanatory variable in predicting the outcome can be evaluated. In contrast, a deep CNN automatically learns the best discriminatory features for the task at hand. The feature representations are distributed over the weights of the CNN making the features unidentifiable, and, consequently, any subsequent causal reasoning is difficult. However, there have been some recent progress in explanatory understanding of knowledge representation in CNNs [Simonyan et al., 2013; Mahendran and Vedaldi, 2016; Zhang et al., 2017; Wei et al., 2015; Zeiler and Fergus, 2013] which may eventually lead to better interpretability of CNN models and facilitate causal analysis.

3 Prior work

Most of the prior work using satellite images have used night light as a proxy for development. See [Ghosh et al., 2013] for a review. Night lights have been used to create a global grid of economic activity [Ghosh et al., 2010b]; to show that during 1992-2009 the centre of gravity of economic activity has shifted towards south and east as economies in India, China and Southeast Asia have ‘lit’ up [Cauwels et al., 2014]; to create a world poverty map using World Development Indicators (WDI) 2006 national level estimates for calibration [Elvidge et al., 2009]; as a measure of poverty to enable health interventions [Noor et al., 2008]; to estimate the informal economy using night light images [Ghosh et al., 2009]; to develop an alternative measure of distribution of wealth [Elvidge et al., 2012]; and even for estimating an information and technology development index [Ghosh et al., 2010a]. Most of the above methods use linear regression models using night lights [Ghosh et al., 2013]. In some recent India related studies, [Bhandari and Roychowdhury, 2011] show that night light is a valid proxy for economic activity in India, [Chakravarty and Dehejia, 2017] use night light images to show that inequality in India has been growing both within and across states, and [Asher and Novosad, 2017] use night light images and regression discontinuities to show that ruling political dispensations tend to favour regions that represent their members.

Daytime satellite images, of resolutions varying from $30m \times 30m$ to $0.5m \times 0.5m$ have also been used for econometric analysis. [Burgess et al., 2012] used satellite images to measure deforestation in Indonesia; [Jayachandran, 2009] estimated the impact of air pollution (particulate matter) resulting from Indonesia’s devastating late-1997 forest fire on infant and fetal mortality; [Marx et al., 2017] used high resolution $0.5m \times 0.5m$ satellite images to measure shiny roof as a proxy for dwelling investments in a Nairobi slum and [Costinot et al., 2016] estimated economic impacts of climate change in agriculture. All
the above methods compute hand-crafted features from satellite images and most use linear regression. See [Donaldson and Storeygard, 2016] for an excellent survey and a primer on remote sensing for economists. See [Gibbons et al., 2015; Varian, 2014] for overviews of econometrics issues associated with spatial data and large datasets.

Use of machine learning techniques on satellite images for econometric analysis is relatively new. [Albert et al., 2017] use deep CNN and satellite imagery to identify land use patterns. The economic survey of 2016-2017 of the finance ministry of government of India used satellite images to calculate built-up area and estimate potential property tax collection [The Ministry of Finance, Government of India, 2017a]. They use Principal Component Analysis (PCA) in conjunction with an ensemble model based on Gradient Boosting Model (GBM) algorithm [Natekin and Knoll, 2013] and multinomial classification regression [Tutorial, b]. They extend this work in the second volume of the economic survey [The Ministry of Finance, Government of India, 2017b] to show that India may be more urbanized than previously thought.

Our work is motivated by [Jean et al., 2016] who use machine learning on daytime satellite images to predict poverty. They use a transfer learning approach [Pan and Yang, 2010] to first fine tune VGG CNN-F [Chatfield et al., 2014], a deep CNN model pre-trained to recognize 1000 image categories, to predict the average night light corresponding to a region from high resolution daytime satellite images. The night light is used here as a noisy but easily obtainable proxy for poverty. This fine tuning builds a deep CNN model that learns to predict economic activity from daytime images. The output vectors of the last fully connected layer (fc7) are then used as input features for regression of asset and consumption estimates using data obtained through sample surveys. The low volume sample survey data turns out to be adequate for the last regression step.

We first reproduce the results of [Jean et al., 2016] to fine tune VGG CNN-S [Chatfield et al., 2014] to predict night light corresponding to 218000 villages in six Indian states. We then follow the same approach of fine tuning VGG CNN-S to directly predict an asset model obtained out of the Census 2011 [The Ministry of Home Affairs, Government of India, 2011] data for these villages. We find that the latter approach works better not only for prediction of assets but also for transfer learning [Pan and Yang, 2010] of other economic indicators.

## 4 Data description

In what follows we briefly describe the data that we have used for our computational experiments.

### 4.1 Asset model

We create our asset model from the Houselisting and Housing data (Table HH-14) of Census 2011 [The Ministry of Home Affairs, Government of India, 2011] corresponding to the six north Indian states of Punjab, Haryana, Uttar Pradesh, Bihar, Jharkhand and West Bengal. This village level data is indexed by the village ids and provides aggregated information of about 140 amenities and assets in the village households. Each column indicates the percentage of households with a facility. We do a dimensionality reduction by manually aggregating, using weighted sums, from the 140 dimensional vector to create a 16 dimensional vector (Table 1). [Jean et al., 2016] used the first principal component of the data as the target of the regression. However, we find that our manual aggregation gives better regression accuracy than using the linear principal component analysis (PCA) for dimensionality reduction. Though the leading principal components capture high variance in the data, other components may be equally important for economic analysis. In Figure 1 we show the correlation matrix of the 140 column census asset data with the 16 aggregated
<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description (% of houses in the village with)</th>
<th>Aggregated from columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>electronics</td>
<td>radio/transistor/tv/laptop</td>
<td>(\frac{([128]+[129]+[130]+[131])}{3})</td>
</tr>
<tr>
<td>water-treated</td>
<td>water from treated source/covered well/ tube-well</td>
<td>([72]+[74]+[77])</td>
</tr>
<tr>
<td>water-untreated</td>
<td>water from untreated source/ uncovered well</td>
<td>([73]+[75])</td>
</tr>
<tr>
<td>water-natural</td>
<td>drinking water from ponds/rivers/lakes</td>
<td>([76]+[78]+[79]+[80]+[81])</td>
</tr>
<tr>
<td>light from electricity</td>
<td>from grid/solar</td>
<td>([85]+[87])</td>
</tr>
<tr>
<td>light from oil</td>
<td>kerosene/other oil</td>
<td>([86]+[88]+[89])</td>
</tr>
<tr>
<td>has-phone</td>
<td>land-line/mobile/both</td>
<td>([132]+[133]+[134])</td>
</tr>
<tr>
<td>transport-cycle</td>
<td>cycle</td>
<td>([135])</td>
</tr>
<tr>
<td>transport-motorized</td>
<td>motorcycle/scooter/car/jeep</td>
<td>([136]+[137])</td>
</tr>
<tr>
<td>no-assets</td>
<td>no assets (cycle/phone etc.)</td>
<td>([139])</td>
</tr>
<tr>
<td>banking-services</td>
<td>availing banking services</td>
<td>([127])</td>
</tr>
<tr>
<td>cook-fuel-processed</td>
<td>LPG/electric stove etc.</td>
<td>([113]+[114]+[115])</td>
</tr>
<tr>
<td>bathroom-within</td>
<td>bathroom within premises</td>
<td>([103]+[104])</td>
</tr>
<tr>
<td>rooms-under-3</td>
<td>less than 3 rooms</td>
<td>([49]+[50]+[51])</td>
</tr>
<tr>
<td>household-size-under-5</td>
<td>less than 5 family members</td>
<td>([56]+[57]+[58]+[59])</td>
</tr>
<tr>
<td>permanent-house</td>
<td>permanent house</td>
<td>([140])</td>
</tr>
</tbody>
</table>

Table 1: Asset indicators aggregated from village level Census 2011 data

indicators. Majority of the census parameters are strongly correlated to one or more of the aggregated indicators. A few have small numerical values and are hence poorly correlated. This demonstrates that the 16 dimensional asset vector is representative of the original 140 dimensional asset data.

The Census 2011 data is noisy and there are large errors for some villages; see Figure 5 for example. For rejecting outliers we compute the distribution of Mahalonobis distances of all villages from their mean as

\[
D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}
\]

where \(\mu\) is the mean of the distribution and \(\Sigma\) is the covariance matrix. We find a steep rise above a distance threshold of 30 and we reject these villages as outliers. Approximately 5% of the villages get rejected. In Section 5 we present regression results with and without outlier rejection.

4.2 Night lights

We use the night light data provided by the Defense Meteorological Satellite Program’s Operational Linescan System (DMSP-OLS) [NOAA/NGDC Earth Observation Group, 2013]. The night light data is available in 30 arc second grids, spanning \(-180^\circ\) to \(180^\circ\) longitude and \(-65^\circ\) to \(75^\circ\) latitude. Each 30 arc second grid cell is mapped to discrete values from \{0, 1, \ldots, 63\}, where 63 corresponds to the highest night light intensity.

4.3 Daytime satellite images

For geo-registration of the villages we use the polygon boundaries generated by Survey of India [The Ministry of Science and Technology, Government of India] \(^1\). The village shape files are linked to the Census data.

\(^1\) Actually, we use proprietary geo-registered boundary data obtained from M/s Pitney Bowes (https://www.pitneybowes.com/in), which is a post-processed and cleaned up version of the original Survey of India [The Ministry of Science and Technology, Government of India].
Figure 1: Correlation between each of the 140 census asset parameters and 16 aggregated indicators. Few census parameters have small numerical values and are hence poorly correlated with any of the 16 handcrafted indicators.

We obtain the daytime satellite images corresponding to the villages from Google static maps using the API provided by Google [Google Static Maps API, 2017]. The Google static maps are free of cloud cover and other noise, because they are carefully constructed mosaics over time. The Google static maps only has recent imagery corresponding to 2017, which we use for regression of the census asset model corresponding to 2011. We assume that the visual characteristics have not changed significantly in six years to make the disparate cross-sectional mapping invalid. We comment more on this in Section 7. We have experimented with both tiling the village shapes using 1 Km$^2$ images and using single images corresponding to the village centroid covering larger areas of 4 and 7 Km$^2$. Tiling the villages with 1Km$^2$ tiles is computationally expensive and gave similar regression scores compared to single images at the village centroid. Consequently, centred images were used in the final analysis.
5 Regression of the asset model using deep convolutional neural network (CNN)

We use the pre-trained VGG CNN-S convolutional neural network [Chatfield et al., 2014], and modify it for our regression tasks described below. We convert the standard input size of $224 \times 224$ of the model to $400 \times 400$. Consequently, only the weights of convolutional layers could be used as a part of pre-trained model, and the three fully connected layers needed to be trained afresh for which weights were initialized with a zero mean Gaussian distribution. Changing the underlying task of the deep CNN from classification to regression necessitated change of the hyper-parameters of the model. The weight decay (regularization parameter) was changed from $0.0005$ to $0.005$. Caffe [Jia et al., 2014] does not support regression by default, so custom layers for data and $R^2$ had to be created. The weights were learnt using Stochastic Gradient Descent (SGD) [Tutorial, a] with a batch size of 32 on a K40 NVidia GPU. The SGD was performed using a step learning policy with a learning rate of $10^{-6}$, $\gamma$ of 0.2 and momentum of 0.8.

5.1 Regression of night light data

We first train the deep CNN model to predict night light values from daytime satellite imagery. We identify the set of relevant night light cells and collect the daytime satellite images corresponding to the centre of these cells. We limit the search space by using ESRI shape files of India available from GADM [GADM database of Global Administrative Areas, 2017]. Out of the approximately 4 million night cells corresponding to India, nearly 50% have zero value resulting in a skewed night light histogram. We reduce the skew (third order moment) from 3.62 to 0.4 by under-sampling to obtain a final dataset of 219000 night cell data points. We perform the regression with a train-test split of 8:2.

We experiment with two different sizes of daytime images of $400 \times 400$ and $640 \times 640$, both at a zoom level of 15. At this zoom level they cover approximately 1 $Km^2$ and 2.5 $Km^2$ respectively, and they both subsume a night light cell which covers less than 1 $Km^2$. We map the daytime images to the input size of $400 \times 400$ required by the modified VGG CNN-S, and build a custom data layer at the input for on-the-fly data augmentations [Goodfellow et al., 2016] like flip, vertical flip and image rotations etc.

We obtain regression $R^2$ scores of 0.69 and 0.79 for 1 $Km^2$ and 2.5 $Km^2$ respectively on the test set. The superior regression accuracy on increasing the ground area can perhaps be explained by the fact that light has a spread; and night light at a particular cell is influenced by the habitation in the surroundings cells as well, and that the neighbourhood context is important for predicting the value at the current cell.

5.2 Transfer learning of the asset model from the night light model

The output of the night light model at the last hidden layer serves as a feature vector which provides a 4096 dimensional representation of the village’s economic prosperity, captured by the night light as a proxy. We use this for transfer learning [Pan and Yang, 2010] of the asset model of Table 1. We use the $640 \times 640$ daytime images at a zoom level of 15 as the input, and design a neural network with a single fully connected layer with rectified linear activation [Goodfellow et al., 2016] to do a regression of the asset model from the 4096 dimensional representation. Despite the high $R^2$ score for the night light data, the $R^2$ scores that we obtain for transfer learning of the asset model, shown in Figure 2, are significantly lower than what has been reported in [Jean et al., 2016]. Since night light as a proxy for economic development has been reported to work well in India [Chen and Nordhaus, 2011; Bhandari and Roychowdhury, 2011; Chakravarty and Dehejia, 2017; Ghosh et al., 2013], the low regression scores can perhaps be ascribed to the noise in the census data.
5.3 Direct regression of the asset model

Availability of a large dataset with census data corresponding 218000 villages from six states allows us to train a deep CNN for direct regression of the asset model with daytime images as input. We modify the last fully connected layer to have a 16 dimensional output corresponding to the asset model described in Section 4.1.

We define the objective function for training the regression model as

\[ C(\theta) = \frac{1}{2M} \sum_{i=1}^{M} \sum_{j=1}^{16} (f_{ij}^i - y_{ij}^i)^2 + \frac{d}{2} \sum_l d_l \sum w_l^2 \]

where \( \theta \) is the vector of all parameters (weights) in the model; the first term is the Euclidean loss where \( f_{ij}^i \) is the prediction value of the \( j^{th} \) indicator for the \( i^{th} \) village and \( y_{ij}^i \) is the value computed from the census data; the second is a \( L2 \) regularization term where \( d \) is a weight decay, \( d_l \) is the decay multiplier for the layer \( l \) and the last sum is over all weights \( w_l \) in the layer. We use \( d = 0.005 \) and \( d_l = 1 \). We initialize the weights in the last three fully connected layers using a Gaussian distribution with zero mean and standard deviation of 0.005. The remaining layers are initialized with the pre-learnt values of VGG CNN-S. During training all weights of all layers are updated during the optimization process. We use Caffe [Jia et al., 2014] for specification and training of the network.

The input daytime satellite imagery was collected via Google static maps [Google Static Maps API, 2017] at a zoom level of 16 and a size of 640 × 640 which corresponds to a ground area of 7 km². 95% of the villages have a ground area less than 7 km². The training and test set split used was 8:2.

We train the network both with and without outlier rejection (described in Section 4.1). As a result of outlier rejection 17000 villages are removed from the dataset resulting in a reduced dataset of 201000
villages. The mean Euclidean loss for the 17000 outlier points was 14800 which is significantly higher than the mean overall Euclidean loss of 5188 corresponding to the final model. The mean overall Euclidean loss without outlier rejection was 5944. In Figure 2 we compare the direct regression $R^2$ scores, with and without outlier rejection, with that of transfer learning from night light regression.

We also experimented with varying the input image sizes to cover 1, 4 and 7 $\text{km}^2$ of the ground area; and with tiling each village with 1 $\text{km}^2$ tiles and training each image tile with the aggregate village census data. 95% of the villages are fully covered by a 5 $\text{km}^2$ square located at the village centre. In Figure 3 we show the per-indicator $R^2$ scores for various choices. The tiling model and the choice of 7 $\text{km}^2$ were comparable and gave better $R^2$ scores for regression. We choose the simpler 7 $\text{km}^2$ input image size for subsequent experiments.

As can be noted from Figures 2 and 3, the prediction accuracy of direct regression is superior to what is obtained from transfer learning from training a night light model.

To ensure that there is no “placebo effect” we also attempted to train the network after randomizing the input-output mappings. As expected, the training did not converge and the $R^2$ scores were close to zero or negative.

In Figure 4 we plot a) the original census values and b) the predicted output of the direct regression model, for some asset indicators on choropleth maps for the six states. In Figure 5 we zoom into West Bengal for the “water treated” indicator. The salt and pepper noise in the census data indicates random errors in the census data. As is evident, the noise is smoothed out in the regression output which has more geo-spatial consistency. Similar effects are observed for all the asset indicators.

It is to be noted that the smoothing of the errors in the census data is due to the non-linear regression model learnt from over 200000 villages spread across six states which averages out the error. Smoothing
Figure 4: Choropleths for census data and model prediction for the ‘bathroom-within’, ‘water-treated’, ‘light from electricity’ and ‘electronics’ indicators. The values are percentages. The predicted output is sometimes greater than 100%. See http://web.iitd.ac.in/~suban/satellite/asset-model/ for the choropleths for the other indicators.

Indeed, as we show in Section 6, transfer learning of the raw census data over local neighbourhood of villages can also reduce the salt and pepper noise, but such local smoothing fails as a proxy for economic activity.
other socio-economic and health indicators fails from such local smoothing, but succeeds from regression of the asset model.

A significant deviation of the original census data from the predicted value for a village would indicate error in the census data. Thus, not only do we obtain more accurate prediction, but we also get a census validation tool as a by-product.

5.4 Salient image features for prediction

As we have mentioned in Section 2, the interpretability of a deep CNN model is low. The features measured from images that are responsible for accurate regression are distributed in the weights of the CNN making it hard to decipher what exactly attributed to the accurate regression. This, in turn, makes any causal analysis difficult. However, there have been some promising recent research [Simonyan et al., 2013; Mahendran and Vedaldi, 2016; Zhang et al., 2017; Wei et al., 2015; Zeiler and Fergus, 2013] towards interpreting deep CNNs which may eventually lead to better interpretation of CNN based models.

We use the method suggested in [Zeiler and Fergus, 2013] in some sample images to understand what makes the regression possible. We slide a $16 \times 16$ occluder object over the images to investigate which parts of an image is responsible for the regression accuracy and plot this as a two dimensional heatmap. We show some sample results in Figure 6. A sharp drop in the heatmap value (indicated in blue) when a region is occluded indicates that the region is significant for the regression outcome. Note that the regression model automatically learns what to control for in the images.
Figure 6: We plot heatmaps to understand what parts of an image may be responsible for regression of an asset indicator. The built-up and inhabited areas clearly show a dip when occluded for the “permanent house” and “bathroom within” indicators. Motorized transport are likely to be observed in developed villages with good roads as clearly illustrated by a heatmap for the “transport motorized” indicator.
6 Transfer learning of other socio-economic parameters

<table>
<thead>
<tr>
<th>Indicator</th>
<th>from original asset data</th>
<th>from the night light model</th>
<th>from the direct regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literacy rate</td>
<td>0.338</td>
<td>0.170</td>
<td>0.340</td>
</tr>
<tr>
<td>Scheduled caste (SC) percentage</td>
<td>0.033</td>
<td>0.032</td>
<td>0.102</td>
</tr>
<tr>
<td>Scheduled Tribe (ST) percentage</td>
<td>0.303</td>
<td>0.323</td>
<td>0.531</td>
</tr>
<tr>
<td>Percentage of working population</td>
<td>0.181</td>
<td>0.125</td>
<td>0.194</td>
</tr>
<tr>
<td>Overall variance weighted R2 score</td>
<td>0.19</td>
<td>0.176</td>
<td>0.314</td>
</tr>
</tbody>
</table>

Table 2: Cross validated $R^2$ scores for the three transfer learning models

Intuitively, the economic indicators in the asset model of Table 1 may be correlated, either directly, or in some kernel space through nonlinear mappings, with the observable features in the daytime satellite images like proportion of built-up area, road area and road types, density and type of housing, water bodies, forest cover and green areas etc. This is confirmed by the regression models presented in Section 5. In this section we investigate whether the direct regression model trained to predict the asset indicators can be used for transfer learning [Pan and Yang, 2010] of other socio-economic indicators for literacy, health and demographics which are not directly related to what can be measured from satellite imagery.

In Table 2 we present the leave-one-out cross validation $R^2$ scores of three fully connected two layered neural network models with rectified linear activations [Goodfellow et al., 2016] trained to predict a few socio-economic indicators obtained from Census 2011 [The Ministry of Home Affairs, Government of India, 2011] (Population Enumeration Data) in all villages in the six states using the following as input:

1. the asset model computed from the raw census data.
2. the 4096 dimensional feature vector obtained from the last layer of the night light model of Section 5.1.
3. the 4096 dimensional feature vector obtained from the last layer of the direct regression model of Section 5.3.

Clearly, transfer learning from the direct regression model outperforms the other two. This also demonstrates that the predicted output of the direct regression model captures the asset model better than the original census data. In Figure 7 we present the choropleth maps for transfer learning from the direct regression model.

We also try transfer learning of some socio-economic and health indicators from the asset model using the Census 2011 [The Ministry of Home Affairs, Government of India, 2011] and NFHS-4 survey data [The Ministry of Health and Family Welfare, Government of India, 2016] data. The NFHS-4 indicators are described in Table 3. The “education levels” data in Census 2011 (Population Enumeration Data - C-08...
Educational Level By Age And Sex For Population Age 7 And Above (Total, SC/ST) (India & States/UTs-District Level)), and the NFHS-4 data are available at a lower granularity of district level, and there are a total of 192 districts in the six north Indian states. Out of these we consider only the 182 'rural' districts for transfer learning of the NFHS-4 data.
<table>
<thead>
<tr>
<th>Indicator</th>
<th>NFHS-4 section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-rural</td>
<td>Population and Household Profile (rural)</td>
<td>Households using improved sanitation facility (%)</td>
</tr>
<tr>
<td>33-rural</td>
<td>Maternity Care (rural)</td>
<td>Mothers who had full antenatal care (%)</td>
</tr>
<tr>
<td>35-rural</td>
<td>Maternity Care (rural)</td>
<td>Mothers who received post-natal care from a doctor/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (%)</td>
</tr>
<tr>
<td>41-rural</td>
<td>Delivery Care (rural)</td>
<td>Institutional births in public facility (%)</td>
</tr>
<tr>
<td>47-rural</td>
<td>Child Immunizations and Vitamin A Supplementation (rural)</td>
<td>Children age 12-23 months fully immunized (BCG, measles, and 3 doses each of polio and DPT) (%)</td>
</tr>
<tr>
<td>68-rural</td>
<td>Child Feeding Practices and Nutritional Status of Children (rural)</td>
<td>Children under 5 years who are stunted (height-for-age) (%)</td>
</tr>
<tr>
<td>69-rural</td>
<td>Child Feeding Practices and Nutritional Status of Children (rural)</td>
<td>Children under 5 years who are wasted (weight-for-height) (%)</td>
</tr>
<tr>
<td>70-rural</td>
<td>Child Feeding Practices and Nutritional Status of Children (rural)</td>
<td>Children under 5 years who are severely wasted (weight-for-height) (%)</td>
</tr>
</tbody>
</table>

Table 3: Socio-economic and health indicators from NFHS-4.

For transfer learning of these indicators we aggregate the asset model regression output for all villages in a district by averaging, and train a neural network with a single fully connected layer with rectified linear activation [Goodfellow et al., 2016] to do a regression of the NFHS-4 and “education level” indicators from the 16 dimensional input. We do a 5 part split for the 182 ‘rural’ districts and train using leave-one-out cross validation. The $R^2$ scores that we report is on the validation set for the best performing model.

In Figures 8, 9 and 10 we show the state-wise choropleth maps for the ground truth and prediction of the NFHS-4 and “education level” indicators. We also indicate the regression $R^2$ scores for a cross validation set.

In Figure 11 we show the pair-wise scatter plots of the NFHS-4 indicators computed over all 182 districts. As can be noted that all the NFHS-4 indicators apart from “institutional birth’ and “wasting” can be accurately predicted from regression output of the asset model, even though some of them are poorly correlated among each other. The highly nonlinear functional relationships are surprisingly well captured by the transfer learning.

We also try transfer learning of all the 90 ‘rural’ parameters in NFHS-4. In Figure 12, we show the histogram of $R^2$ scores. Even though the NFHS-4 parameters are not intuitively related to what can be observed from satellite images, we obtain reasonable prediction accuracy for over 75% of them. As we demonstrate in Section 9, such predictions can be used to remove possible omitted variable biases and predict endogenous variables at the first stage of a linear regression experiment.
Figure 8: NFHS-4 data and transfer learning prediction for “8-rural”, “33-rural”, “35-rural” and “41-rural” indicators. The $R^2$ scores for regression were 0.73, 0.57, 0.45 and -0.6 respectively. The grayscale indicates proportions.
Figure 9: NFHS-4 data and transfer learning prediction for “47-rural”, “68-rural”, “69-rural” and “70-rural” indicators. The $R^2$ scores for regression were 0.61, 0.61, -0.45, -0.45 respectively. The grayscale indicates proportions.
Figure 10: Census data and transfer learning prediction for “average education level”. The $R^2$ score for regression was 0.54. The grayscale indicates percentages.

Figure 11: Pair-wise scatter plots of the NFHS-4 indicators computed over 182 districts.

Figure 12: Histogram of $R^2$ scores of regression of all 90 ‘rural’ parameters of NFHS-4. See http://web.iitd.ac.in/~suban/satellite/nfhs4_r2_scores.html for the $R^2$ scores and http://web.iitd.ac.in/~suban/satellite/nfhs4/ for the choropleths.
7 Monitoring development over time

Figure 13: Images of Bhadana village, Sonipat tehsil, Haryana state captured at different times between 2011-2017.

Despite the fact that we use disparate cross-sectional data - daytime satellite images captured in 2017 and census data of 2011 - to train the regression model, we obtain reasonably high regression accuracy. This can be attributed to the fact that the large number of villages (over 200000) that we use to train the model are at varying stages of economic development, and their collective diversity is rich enough to represent the characteristics of economic development spanning several years. This provides us with an opportunity to use the static regression model trained with cross-sectional data to monitor the temporal evolution of a village.

In Figure 13 we show images of a village captured at different times between 2011 and 2017. In Figure 14 we present the regression output of our asset indicators for these images. Despite the fact that the cross-sectional regression of Census 2011 indicators from the 2017 image is not very accurate for this village, indicating that the Census 2011 data for this village is not accurate, the near monotonic development of the village through the years is evident from the predicted values.
We argue that spatial discontinuities or sharp spatial gradients in the regression output of a development indicator require special attention. Such high gradients can be computed by carrying out edge detection [Gonzalez and Woods, 2009] in the regression output choropleth maps - see Figure 15 for an example. Edge detection involves computing the image gradients using finite differences and thresholding the high gradient.

Figure 14: The evolution of asset indicators over time

8 Spatial discontinuities in regression output

We argue that spatial discontinuities or sharp spatial gradients in the regression output of a development indicator require special attention. Such high gradients can be computed by carrying out edge detection [Gonzalez and Woods, 2009] in the regression output choropleth maps - see Figure 15 for an example. Edge detection involves computing the image gradients using finite differences and thresholding the high gradient.
image locations. An edge point in the regression output indicates that geographically neighbouring villages have dissimilar economic development outcomes. Often such edges can occur at village-forest or city-village boundaries and can be easily explained away. Otherwise, if two neighbouring villages are similar in all other aspects, then an edge in the regression output may suggest that the villages have been subjected to different policy interventions or there may be other socio-economic anomalies. All such edge points require special attention and investigation to find causal explanations. Indeed, not surprisingly, the edges in Figure 15 often coincide with state or district boundaries where neighbouring villages belonging to different states or districts have been subjected to different policy interventions.

9 Using predicted variables for regression: a case study to understand stunting

Finally, we do a case study of regression analysis using predicted variables to understand the determinants of the disturbingly high rate of stunting (see second row of Figure 9) in the north Indian states. The determinants of stunting are not yet completely understood [Coffey et al., 2013], and we analyse the robustness of a recent regression study carried out by [Spears et al., 2013].

Identification with observational data in econometrics crucially depends on to what extent one can maintain the assumption of exogeneity of explanatory variables. One of the main reasons for violation of exogeneity arises from inability to control for omitted sources of variation which may be correlated with the included explanatory variables. Finding a good proxy for an omitted variable is a difficult task, either because of missing variables in the data, or because of missing observations, or both. Missing observations usually require extrapolation or interpolation based projections, whereas missing variables require finding a suitable proxy from other data sources. We show that machine learning based predictions can not only account for missing observations but can also be useful for generation of proxy variables and correcting for omitted variable biases. Our predictive tool can also correct for attenuation biases due to measurement errors such as in Census and account for the classical “error in variables” problem.

[Spears et al., 2013] hypothesise that prevalence of open defecation is an important correlate of stunting
rates in India. The unit of their analysis is a non-representative sample of 112 districts out of over 600 districts in India. They use four different data sets, spread over 2005 to 2011, to collect the relevant variables. The outcome variable (stunting rates) is from a survey in 2010-11, the open defecation rates are collected from the 2011 Census report, infant mortality rates are from Annual Health Survey 2010-2011 and the consumption and calorie data are collected from 2005 National Sample Survey. The district level stunting rate is regressed on log of open defecation rate, controlling for several other covariates like urban population (quadratic), log of monthly per capita consumption expenditure (MPCE), calories per capita, cereal calories per capita, household size, overall literacy rate and female literacy rate. Their main conclusion is that districts with more open defecation have more stunted children and that it is robust with respect to multiple controls. With the full set of controls, a 10% increase in open defecation rate is associated with 0.7% increase in the stunting rate. This association fades away a bit when they also control for infant mortality rate (IMR), which is endogeneous in the specification (see Table 4).

Table 4: Comparison of our models with [Spears et al., 2013]. Note that all but the calories are proportions. The calorie values range from 1000 to 2500.

We replicate this regression using our predicted values for all the districts in the six north Indian states. We leave out a few districts where the predicted rate for the outcome variable or that of an important control variable turns out to be negative (this can happen in a predictive model). Hence, the sample districts in our study are different from that of [Spears et al., 2013]. Our constructed sample of district level variables are only for the rural areas, therefore our first model is comparable with the model 2 of [Spears et al., 2013] (See Table 4). We obtain our district level predictions from transfer learning of the NFHS-4 and “education
level” indicators described in Section 6.

<table>
<thead>
<tr>
<th></th>
<th>Predicted data</th>
<th>Actual data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m1</td>
<td>m2</td>
</tr>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>ln(open defecation)</td>
<td>7.404***</td>
<td>3.621***</td>
</tr>
<tr>
<td>(0.66)</td>
<td>(0.68)</td>
<td>(0.60)</td>
</tr>
<tr>
<td>ln(mpce)</td>
<td>(1.55)</td>
<td>(1.44)</td>
</tr>
<tr>
<td>kcal2</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>kcal-cereal</td>
<td>0.020***</td>
<td>0.011***</td>
</tr>
<tr>
<td>household size under 5</td>
<td>-0.006</td>
<td>-0.271***</td>
</tr>
<tr>
<td>literacy rate</td>
<td>-0.016***</td>
<td>-0.043***</td>
</tr>
<tr>
<td>women-lit</td>
<td>-0.151***</td>
<td>-0.022</td>
</tr>
<tr>
<td>(0.05)</td>
<td>(0.12)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>mom-folic</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td>women-sec-edu</td>
<td>-0.174**</td>
<td>0.083</td>
</tr>
<tr>
<td>mom-full-ant-care</td>
<td>-0.021</td>
<td>-0.021</td>
</tr>
<tr>
<td>caesarean-birth</td>
<td>-0.019</td>
<td>0.028</td>
</tr>
<tr>
<td>children-vitA</td>
<td>-0.095***</td>
<td>-0.048***</td>
</tr>
<tr>
<td>women-bmi-below-norm</td>
<td>0.192*</td>
<td>0.023</td>
</tr>
<tr>
<td>clean-fuel</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>Constant</td>
<td>12.827***</td>
<td>80.076***</td>
</tr>
<tr>
<td>(2.62)</td>
<td>(14.17)</td>
<td>(11.09)</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.411</td>
<td>0.733</td>
</tr>
<tr>
<td>N</td>
<td>178</td>
<td>178</td>
</tr>
</tbody>
</table>

* p < 0.10, **p < 0.05, ***p < 0.01

Table 5: Comparison of models with predicted data and actual data

The estimates are remarkably similar. However, when we control for log of MPCE, total calorie, calorie from cereal and percentage of household size under 5, our estimates are quite different from that of [Spears et al., 2013]. We should keep in mind that our consumption data comes from 2011-12 NSSO survey [The Ministry of Statistics and Programme Implementation, Government of India, 2017], whereas the consumption data for [Spears et al., 2013] are quite disparate (2005). A significant part of the variation in stunting is explained by overall economic condition of the households as measured by district level average MPCE. As expected, higher calorie consumption is negatively associated with stunting rate whereas calorie from cereals is positively associated with stunting rate. This is not surprising because given two households with similar levels of total calorie intake, the one with more calorie from cereals (nutritionally not so rich) will have higher stunting possibilities. We may extend the same logic to district level, where indeed the variance of the total per capita calorie intake is low. Unlike in our model, model 3 in [Spears et al., 2013] does not indicate any of the control factors to be significant, whereas open defecation remains very significant and the estimated value does not drop. In the third specification, we control a few more variables related to literacy rates (m3). The effect of open defecation drops further but remains significant at 5 percent level. However, instead of MPCE, overall literacy rates, women literacy rates and household sizes dominate as significant factors affecting stunting. This specification is comparable with model 4 of [Spears et al., 2013], but the estimates of the partial effect of open defecation remain very high in [Spears et al., 2013]. A 10% decrease in open defecation rate is associated with only 0.13 percentage point decrease in stunting rate whereas [Spears et al., 2013] conclude that for their sample, the reduction is about 0.7 percent. We should keep in mind that the sample districts in [Spears et al., 2013] are non-random and are selected based on high prevalence of stunting.
There are several other important covariates which we believe can have effect on stunting rates. In our final specification (m4) we also include proportion of women having secondary or higher education, proportion of mothers who took folic acid supplement during pregnancy, proportion of mothers with full antenatal care, proportion of mothers who had caesarean birth, proportion of women with below normal BMI, proportion of children who had vitamin A, and proportion of households with clean fuel source for cooking. As infant mortality rate is a problematic explanatory variables, we exclude it from our final specification. With this full set of controls, our estimation shows that literacy rates and education levels, particularly women having secondary or higher level of education, are important factors negatively associated with stunting rate. Open defecation remains a significant variable but its effect is much lower as compared to that in [Spears et al., 2013]. The main contrast between our results and that of [Spears et al., 2013] is that for our sample education plays an important role in improving stunting, particularly through mothers’ education, whereas in the [Spears et al., 2013] sample, education does not play a significant role. Thus our findings are also in agreement with those in [Som et al., 2007; Kumar and Sinha, 2017].

In Table 5 we compare the regression results using predicted and actual variables. They are qualitatively similar, but the predicted variables are more “noise free” and provide more consistent and interpretable estimates.

Note that all the above regression analyses, either by [Spears et al., 2013] or by us, were carried out using district level aggregated data. Most of the survey variables, including stunting rates, are available only at the district level. However, we can use our machine learning tool to predict corresponding village level
Figure 17: Kernel density plots of regression coefficients for some of the significant explanatory variable. We show the density plot for only one state (categorical variable), the others are similar.
estimates from satellite images. Consequently, we can also carry out the same regression analysis at a much lower level of data aggregation at the village level. Our six north Indian states have 218000 villages, and linear regression with a such a large sample size makes the $p$ values go to zero very quickly. In view of this we take random uniform samples of villages multiple times and do repeated regression analysis. We carry out a power and sample size estimation using the procedure outlined in [Sullivan], and determine the optimal sample size for regression to be approximately 3500 villages. We carry out the regression analysis 100 times using 3500 randomly choses villages and compute the histogram of 1%, 5% and 10% significance levels for each variable and the kernel densities of the corresponding regression coefficients. For the village level analysis we introduce the proportion of permanent houses in a village as an additional control which measure a village level asset infrastructure. For the village level regression we also control for the states keeping the largest state, Uttar Pradesh, as the base (note that UP also has the highest rate of stunting). In Figure 16 we plot the histogram of the fraction of times a variable is 10%, 5% an 1% significant. We show the positive and the negative associations separately. In Figure 17 we show the kernel density plots of the regression coefficients for the significant variables. We show it for only one state, West Bengal.

The village level analysis clearly reinforces that although the rate of open defecation is an important factor, women’s education levels are the most significant determinant of stunting. Moreover, there are several other infrastructure and diet related explanatory variables that turn out to be at least as important as open defecation. The limitation of the sample in [Spears et al., 2013] due to unavailability of data for the whole set of districts (or a representative sample of districts), led them to quite a different conclusion. In contrast, using satellite based predictions we generate quite accurate measures of some more controls and construct a representative sample of districts. Using this richer sample, our conclusion is that along with open defecation, women education is also an important factor that affects stunting. We also confirm this by detailed village level analysis using predicted data.

10 Conclusion

We have presented a tool for monitoring development using high resolution day-time satellite images. Satellite images can be obtained cheaply, easily and frequently, and can provide valuable estimates when survey data are infrequent or missing. We demonstrate this using a regression case study to understand the determinants of stunting. We use machine learning to build a deep CNN based regression model for a hand crafted asset vector from input satellite images. Though the model is static and is trained with cross-sectional data, we demonstrate that it can be effectively used to predict the asset model from satellite images acquired at different times, making it extremely useful alternative between surveys. Further, the asset model can be used for transfer learning and prediction of a variety of other socio-economic and health parameters. We also demonstrate an interesting application of our tool to generate alerts by detecting dissimilar regression outcomes in geographically neighbouring regions.

Acknowledgement

We thank Armaan Bhullar for research assistance. Part of this work was initiated during his Bachelor’s project at IIT Delhi.

We thank the IIT Delhi High Performance Computing facility for providing the computational resources for this work.

We thank Dr. Manish Sharma and M/s Pitney Bowes (https://www.pitneybowes.com/in) for providing us with the proprietary shape files for all the villages in the six north Indian states, indexed by the village
census ids.

References


