Learning about Learning with Deep Learning: Satellite Estimates of School Test Scores

Heather M. Baier

Follow this and additional works at: https://scholarworks.wm.edu/honorstheses

Part of the Educational Assessment, Evaluation, and Research Commons, Geographic Information Sciences Commons, and the Remote Sensing Commons

Recommended Citation

This Honors Thesis is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Learning about Learning with Deep Learning: Satellite Estimates of School Test Scores

Author: Heather BAIER
Advisor: Dan RUNFOLA

A thesis submitted in fulfillment of the requirements for Interdisciplinary Honors in the degree of Bachelors of Science in the Data Science Program

Accepted for Honors
Chair: Dr. Dan Runfola
Matthias Leu

Dr. Matthias Leu
Dr. Anthony Stefanidis

Williamsburg, Virginia
May 4, 2020
Abstract

Dr. Dan Runfola
Data Science Program
Bachelors of Science

Learning about Learning with Deep Learning: Satellite Estimates of School Test Scores
by Heather BAIER

Convolutional neural networks are deep-learning models commonly applied when analyzing imagery. Convolutional neural networks and satellite imagery have shown potential for the global estimation of key factors driving socioeconomic ability to adapt to global change. Unlike more traditional approaches to data collection, such as surveys, approaches based on satellite data are low cost, timely, and allow replication by a wide range of parties. We illustrate the potential of this approach with a case study estimating school test scores based solely on publicly available imagery in both the Philippines (2010, 2014) and Brazil (2016), with predictive accuracy across years and regions ranging from 76% to 80%. Finally, we discuss the numerous obstacles remaining to the operational use of CNN-based approaches for understanding multiple dimensions of socioeconomic vulnerability, and provide open source computer code for community use.
List of Figures

1.1 Example of Landsat imagery transformations for a single school location. Source image courtesy of the U.S. Geological Survey. . . 3
1.2 Relationship and density of each school score metric. . . . . . . . 4
1.3 SG-CNN implementation for the estimation of school test scores based on multiple sources of imagery. . . . . . . . . . . . . 5
1.4 Relationship between error and the percentage of students receiving conditional cash transfers for the Science subject test. . . . . . . 12
### List of Tables

1.1 Results from predictive models for each subject matter test. MAE is derived from the continuous estimates of score; binary accuracy is derived from the classification of schools as above or below average. ........................................... 9

1.2 Results from the application of the same modeling strategy presented in the main body of this work to an independent data set from Brazil. ......................................................... 10

1.3 Contrast of the transfer learning CNN-based approach in (Jean et al., 2016) and the SG-CNN detailed in this piece. The multi-imagery source ensemble showed improvements over high resolution imagery alone in all subjects. .......................... 11

1.4 Test results in which the data withheld for calibration and validation is randomized. ................................................................. 11

1.5 Results from the application of the same modeling strategy presented in the main body of this work to an independent dataset from 2009-2010. .................................................... 12
List of Abbreviations

CCT          Conditional Cash Transfer
CNN          Convolutional Neural Network
ENEM         Exame Nacional do Ensino Médio
FOI          Freedom of Information
PPPP         Pantawid Pamilyang Pilipino Program
SG-CNN       Stacked generalization convolutional neural network
Chapter 1

Thesis

1 Introduction

Our ability to understand socioeconomic vulnerability to climate change is impeded by a lack of accurate and timely socioeconomic data (Andersson and Archila, 2019). This lack of data is most stark in the developing world (Mossoux and Canters, 2018), frequently in areas with a high degree of exposure to climate change (Hedlund and Benzie, 2018). Such a data gap has numerous real-world implications, as policymakers struggle to develop interventions designed to improve human living conditions, as well as to forecast the strength and sustainability of regional economies in the face of climate change (Schwab, 2019).

In consideration of the costs and limitations of survey-based collection strategies, recent literature has begun to explore alternative approaches to collecting information on human socioeconomic vulnerability. The most popular of these have been the use of phone-based survey instruments (Blumenstock, Cadamuro, and On, 2015) and modeling based on proxy information such as access to electricity measured by nighttime lights (Bruederle and Hodler, 2018). These methods have shown considerable potential, but are each hampered by important limitations - most notably a decrease in accuracy within impoverished areas (Jean et al., 2016).

In this piece we discuss a deep learning model ensemble approach to estimating specific socio-economic factors, using educational outcomes as an illustrative case study. We specifically model school test scores based on a combination of satellite imagery and street view imagery. By leveraging imagery
from the local area a school is located in, we highlight the large amounts of relevant information geographic contextual features - such as road network conditions, patterns of the built environment, or the success or failure of crops - contain for the approximation of socioeconomic factors (Coleman, 1968; Gamoran and Long, 2007; Suárez-Álvarez, J and R Fernández-Alonso, 2014; Tomul and Savasci, 2012).

We choose education both due to its importance to climate adaptation and the well known limitations of education data. While many developing countries self-report data on educational outcomes, reports can be infrequent and generally only provide summaries for entire nations - rather than individual schools or districts - inhibiting use for many practical applications (World Bank, 2019). This limitation is largely attributed to prohibitive survey costs and a lack of government enthusiasm to collect information that may be reflective of poor performance (World Bank, 2018).

The approach we highlight in this piece allows for the estimation of individual school test outcomes using only publicly available imagery, mitigating (though not removing) the need for costly collection to provision estimates of school quality. We validate this approach using information acquired through Freedom of Information (FOI) requests submitted to the government of the Philippines, providing a novel dataset of school test score outcomes for 5,875 public elementary schools (academic year 2013-2014). Further, we provide additional evidence illustrating the extensible nature of this approach by validating against (a) a second time period, and (b) an additional country (Brazil). Finally, we discuss the tremendous potential of a research agenda that fully explores the limits and opportunities of this approach to the estimation of socioeconomic factors.

## 2 Data and Methods

Recent literature has illustrated improvements in deep-learning models for the automated detection of features relevant to socioeconomic outcomes (Jean et al., 2016), techniques that are proven effective when the outcome of interest can be estimated based on a single high resolution satellite source of imagery. However, some socioeconomic outcomes - such as school quality - may not be able to be
effectively estimated using high resolution satellite information alone (see sections 2.1 and 3.1). In these cases, integrating information from multiple sources of imagery - including car mounted cameras, coarse resolution satellite sensors, and aerial imagery - provides a promising pathway forward.

We explore this approach through a multi-source, stacked generalization convolutional neural network (SG-CNN) model in which we train three different neural network architectures, each tailored to a specific type of imagery (street view imagery, high resolution and coarse resolution satellite imagery). Each model is independently calibrated to estimate the probability a school’s test scores are above or below average, with the goal of identifying features in the imagery that are important for discriminating between these two classes. The outcomes of these three models are then integrated in a meta-model approach, which estimates (1) the absolute numeric grade an average student is expected to achieve for each school (on a continuous 40-point scale), and (2) the probability a school is above or below average.
While multiple measures of school quality are available, they are highly correlated (see figure 1.2). As such, we focus on the use of a SG-CNN to estimate the “Science” subject matter test scores on the Philippines’ 2013-2014 National Achievement Test (NAT) for each of the 5,875 schools in our database; 4,406 (75%) are used for calibration and 1,469 withheld for validation. For the purposes of estimating the science subject test score for each school, our machine learning pipeline follows a four-step procedure.

First, Landsat (Woodcock et al., 2008) satellite imagery (see figure 1.1) is retrieved for an approximately 7km region around each school, with each pixel representing a 30 meter area on the ground. We start with a pre-trained convolutional neural network (CNN), ResNeXt-101-32x8d (Xie et al., 2019; He et al., 2015) - which has already been parameterized to optimally distinguish between classes of imagery found in ImageNet (i.e., “dog”, “cat”, “bridge”) (Russakovsky et al., 2015). We then fine-tune the network parameterization to explicitly distinguish between schools which achieve above-average and below-average scores (using a break point of 25.74 for test scores ranging from 7.83 to 39.89).
In stage 2, a similar process is repeated for imagery acquired from the Google Static Maps API. Because this imagery is of much higher spatial resolution than the Landsat imagery acquired in stage 1, the nature of features that can be detected are fundamentally different than the imagery retrieved from Landsat. While Landsat imagery can be used to observe broad trends - i.e., the number of roads in an image, or if the environment has urban features - the imagery from Google provides information on the school itself - i.e., the presence or absence of a playground or temporary shelters.

In stage 3, we retrieve imagery from 4 headings (north, south, east and west) for each of 3,147 schools for which street view imagery was available. Schools that did not have street view data were represented by a special value, as the lack of street view data was anticipated to be correlated with school accessibility, and thus outcomes. A network of the same architecture as employed in stages 1 and 2 is implemented following the same approach.
Fourth, the resultant probabilities for each school being above- or below-average as estimated by the three CNN models (Landsat, Static Maps, Street View) are taken as inputs into two meta-models. These models are used to estimate the (a) probability a school is above- or below-average, and (b) absolute numeric grade for each school.

Deep learning approaches have proven valuable for imagery-based predictive analyses covering a wide range of different topics. Recent innovations have applied transfer learning - where a network is trained on a large dataset such as ImageNet (Russakovsky et al., 2015), and then refined based on a much smaller dataset - to overcome training data limitations (Jean et al., 2016). Acknowledging that our data set is on its own inadequate to train the millions of parameters in the chosen neural networks, we extend the transfer learning approach to multi-source imagery ensembles (Ju, Bibaut, and Laan, 2018). By leveraging imagery from multiple sources and integrating these into a single ensemble of convolutional neural networks, we are able to tailor the parameters in individual networks to a given source of imagery. Contrasted to a simpler alternative in which we use a CNN directly trained on ImageNet and then tailored based on high resolution imagery (Jean et al., 2016) to extract image features, our approach allows for networks tailored to detect features present in satellite images to operate on one subset of our data, and a network tailored to detect features present in street view images on another subset of our data.

In the first step of the SG-CNN, we independently fine-tune three ResNeXt 101-32x8d base classifiers (He et al., 2015), one for each source of imagery (Landsat, Static Maps, and Street View). Parameters previously derived from classifying images found in the ImageNet dataset are applied to each network, and across multiple epochs of back propagation these parameters are fine tuned to identify features that are correlated with school success or failure. We explicitly treat this stage as a classification problem, in which we seek to correctly classify each school as either above or below the geometric mean of test scores.

Each of the three models receives input imagery from only one source: Landsat, Google Static Maps (zoom level 16), or Google Street View. The three models each produce two classification estimates - the probability that a school was above or below average, as well as the hard-classified estimate of a binary 1 or 0 (above or below average). Each of these six values are passed forward into one
of two grid searches, which sought to identify the best combination of model and parameters for the estimation of (a) the absolute average numeric grade for each school, and (b) the probability a school is above- or below-average. Hyper-parameters searched across included tree depth, number of estimators, neighborhood definition, weights, leaf size, and algorithm. A 10 fold cross-validation was applied to test these parameters, with 80% of data being used for calibration and 20% testing in each permutation. In the case of continuous estimation, the best performing model was identified as a random forest with a maximum depth of 7 and 20 constituent trees. In the case of categorical prediction, the best performing model was identified as a nearest neighbors with 7 neighbors and uniform weights ($p = 1$).

2.1 Exploring the Value of an Ensemble Approach

Integrating imagery across multiple sources comes with necessary time and resource costs: not only do additional models have to be fit, but the source imagery must be collected from each sensor independently. To illustrate the value of a multi-image approach, we contrast the SG-CNN to a single image source transfer learning Convolutional Neural Network proposed in previous work (Jean et al., 2016). To conduct this comparison, we fit a ResNeXt 101-32x8d neural network using initial weights provided by ImageNet, and use the Google Maps Static API as input. We then then refine the model over 50 epochs of training.

2.2 Crossfold Validation

In addition to the single split validation presented, we conducted a 5-fold cross validation in which the data used for calibration (75% of the dataset) and validation are randomized for the “Science” test score estimations. While computational limitations precluded testing additional folds on additional subsets of our data, this preliminary analysis is intended to illustrate the robustness of results to changing sample makeup.
2.3 Model performance in different socioeconomic contexts

Previous research has highlighted a significant bias towards disadvantaged communities in the measurement errors of socioeconomic variables. This can be true for both traditional survey approaches (Johnson et al., 2006) and satellite imagery approaches to measurement (Jean et al., 2016). To identify if similar bias existed in the approach presented here, we leveraged a secondary variable present in our dataset - the percentage of students receiving conditions cash transfers (CCTs) as a part of the Pantawid Pamilyang Pilipino Program (PPPP). The PPPP was implemented by the Philippines in 2008, and provides cash transfers to poor households to encourage increased participation in the education system (Chaudhury et al., 2012). This relationship is analyzed in our results.

2.4 Testing in other Contexts: Extensibility Across Time and Space

Recognizing the limits inherent to a single cross-sectional analysis, in this section we present the results of two additional tests. The first test examines the accuracy of the SG-CNN approach when applied to another point in time. The second test applied the SG-CNN to a different geographic settings, using the country of Brazil as a case study. These tests seek to provide preliminary positive or negative evidence that the proposed approach is generally extensible to other geographic and temporal settings.

2.4.1 Testing additional time periods: 2009-2010

To support testing additional time periods, we submitted multiple FOI requests for school test score data during the 2009-2010 school year to the Philippines’ government. We chose 2009 - 2010 as it represented the earliest year for which information can be requested. This data is of the same nature as the data described in our materials and methods, with only the academic year changing. Using this information, we replicated the modeling approach described in the main body of the text to estimate the average test scores across all disciplines for each school, using Landsat information from 2009-2010.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Acc. (%)</th>
<th>MAE (σ)</th>
<th>Obs. Mean (σ)</th>
<th>Pred. Mean (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>78.82%</td>
<td>2.29 (2.33)</td>
<td>28.33 (7.36)</td>
<td>28.29 (7.36)</td>
</tr>
<tr>
<td>Filipino</td>
<td>77.05%</td>
<td>1.53 (1.54)</td>
<td>31.11 (4.49)</td>
<td>31.07 (4.49)</td>
</tr>
<tr>
<td>Math</td>
<td>78.44%</td>
<td>2.31 (2.45)</td>
<td>29.79 (7.34)</td>
<td>29.75 (7.34)</td>
</tr>
<tr>
<td>Science</td>
<td>79.53%</td>
<td>2.38 (2.44)</td>
<td>27.21 (7.43)</td>
<td>27.2 (7.43)</td>
</tr>
<tr>
<td>AP</td>
<td>78.68%</td>
<td>2.18 (2.19)</td>
<td>26.92 (6.59)</td>
<td>26.89 (6.59)</td>
</tr>
<tr>
<td>All (Avg)</td>
<td>79.68%</td>
<td>1.67 (1.82)</td>
<td>28.67 (6.25)</td>
<td>28.63 (6.25)</td>
</tr>
</tbody>
</table>

TABLE 1.1: Results from predictive models for each subject matter test. MAE is derived from the continuous estimates of score; binary accuracy is derived from the classification of schools as above or below average.

2.4.2 Testing additional countries: Brazil

To examine the general utility of the SG-CNN approach for school score estimations, we independently repeated the analysis for a country in a different geographic context: Brazil. The Brazil Exame Nacional do Ensino Médio (ENEM) data is free and publicly available through the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. The ENEM is a non-mandatory, standardized exam taken by high school students in Brazil. The exam evaluates students on composition, Portuguese, history, geography, math, physics, chemistry and biology. The exam results we retrieved for this analysis were from the 2016 school year and cover 5,464 schools in the states of Sao Paulo, Rio de Janeiro, Minas Gerais and Parana. Information provided included the percentage of students that passed the ENEM at each school, which was selected for this test. These percentages range from 47.7% to 100% with a standard deviation of 9.14.

3 Results

Our SG-CNN model predicts if a school is likely to receive above or below average test scores correctly for 79.53% of an independent set of 1,469 test schools. In the absolute score estimation approach, we are able to predict the numeric grade with a mean absolute error (MAE) of 2.26 across all schools; this contrasts to the best performing subject (Filipino), for which the MAE was 1.47 (see table 1.1).
Chapter 1. Thesis

<table>
<thead>
<tr>
<th>Model</th>
<th>Binary Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat</td>
<td>61.30%</td>
</tr>
<tr>
<td>Static Maps</td>
<td>68.85%</td>
</tr>
<tr>
<td>Street View</td>
<td>68.48%</td>
</tr>
<tr>
<td>Binary Ensemble</td>
<td>77.67%</td>
</tr>
<tr>
<td>Continuous Ensemble MAE*</td>
<td>3.11 (σ = 3.29)</td>
</tr>
</tbody>
</table>

**Table 1.2:** Results from the application of the same modeling strategy presented in the main body of this work to an independent data set from Brazil.

Of note in the binary classification models is that the ensemble outperformed any of the individual model components, indicating that different scales of imagery are providing unique information (see also results from Brazil in section 3.5 for additional evidence of the value of information across different scales of imagery).

### 3.1 Results: Value of a Multi-source Ensemble Approach

The single-image CNN approach resulted in accuracy ranging from 67.96% (Filipino) to 71.90% (Math), with an overall all subject accuracy of 72.05%. For science test scores we observed a 10% improvement (69.16% to 79.53%) in the ability of the SG-CNN to discern above- and below-average schools for our testing dataset relative to the single-source transfer learning approach. Improvements ranged from 6.54% (Math) to 9.59% (A.P.) across other subjects (see table 1.3).

We observed a similar difference in our analysis of Brazil, with the SG-CNN approach improving on the static maps approach by 8.82% (see table 1.2). Of particular note in Brazil was the inability of any individual model - Static Maps, Landsat, or Street View - to exceed 69% accuracy; however, the SG-CNN ensemble approach which integrated these models was able to accurately classify 77.67% of schools.

### 3.2 Results: Cross Fold Validation

The findings from our 5-gold validation are presented in table 1.4. Our findings appear to be generally insensitive to the makeup of the training and validation
Chapter 1. Thesis

CNN w/ Static Maps | Stacked Generalization CNN

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Accuracy</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>70.25%</td>
<td>78.82%</td>
<td>8.57%</td>
</tr>
<tr>
<td>Filipino</td>
<td>67.96%</td>
<td>77.05%</td>
<td>9.09%</td>
</tr>
<tr>
<td>Math</td>
<td>71.9%</td>
<td>78.44%</td>
<td>6.54%</td>
</tr>
<tr>
<td>Science</td>
<td>69.16%</td>
<td>79.53%</td>
<td>10.37%</td>
</tr>
<tr>
<td>Araling Panlipunan</td>
<td>69.09%</td>
<td>78.68%</td>
<td>9.59%</td>
</tr>
<tr>
<td>All Subject (Avg)</td>
<td>72.09%</td>
<td>79.68%</td>
<td>7.59%</td>
</tr>
</tbody>
</table>

Table 1.3: Contrast of the transfer learning CNN-based approach in (Jean et al., 2016) and the SG-CNN detailed in this piece. The multi-imagery source ensemble showed improvements over high resolution imagery alone in all subjects.

<table>
<thead>
<tr>
<th>Test</th>
<th>Landsat</th>
<th>Static Maps</th>
<th>Street View</th>
<th>Ensemble</th>
<th>Ensemble MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.04</td>
<td>70.57</td>
<td>69.58</td>
<td>78.99</td>
<td>1.53</td>
</tr>
<tr>
<td>2</td>
<td>76.83</td>
<td>70.13</td>
<td>69.97</td>
<td>77.61</td>
<td>1.51</td>
</tr>
<tr>
<td>3</td>
<td>76.73</td>
<td>70.57</td>
<td>69.93</td>
<td>77.83</td>
<td>1.51</td>
</tr>
<tr>
<td>4</td>
<td>77.14</td>
<td>69.11</td>
<td>70.36</td>
<td>79.7</td>
<td>1.52</td>
</tr>
<tr>
<td>5</td>
<td>76.75</td>
<td>70.35</td>
<td>68.5</td>
<td>77.73</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Table 1.4: Test results in which the data withheld for calibration and validation is randomized.

sets of data, with minor variations in predictive accuracy across runs (for both binary classifications and mean absolute error in score estimates).

3.3 Results: Socioeconomic bias in errors

Figure 1.4 shows the relationship between our model error (the percentage error in our classification model) and the percentage of students receiving CCTs for each school. As this figure illustrates, the error of the SG-CNN approach is apparently largely insensitive to socioeconomic status in the presented case, with a slight bias towards more extreme errors in more socioeconomically affluent areas (i.e., those with lower numbers of CCTs). This suggests the SG-CNN ensemble approach could be generally useful across a wide range of socioeconomic settings.
Figure 1.4: Relationship between error and the percentage of students receiving conditional cash transfers for the Science subject test.

<table>
<thead>
<tr>
<th>Model</th>
<th>Binary Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat</td>
<td>75.66%</td>
</tr>
<tr>
<td>Static Maps</td>
<td>72.06%</td>
</tr>
<tr>
<td>Street View</td>
<td>69.50%</td>
</tr>
<tr>
<td>Binary Ensemble</td>
<td>76.48%</td>
</tr>
</tbody>
</table>

Table 1.5: Results from the application of the same modeling strategy presented in the main body of this work to an independent dataset from 2009-2010.
3.4 Results: Additional Time Periods

Results for our tests of the 2009-2010 academic year indicate the quality of the estimates remained similar to the 2014-2015 case (see table 1.5). In terms of absolute score estimation, the mean absolute error of the estimate was 0.59 (\(\sigma = 0.66\)), relative to an overall mean of 16.03 (\(\sigma = 2.43\)). The estimate of the overall mean (16.06, \(\sigma = 2.17\)) remained very similar to the observed data. In terms of classification accuracy, the SG-CNN ensemble model estimated if each school was above or below the global average with 76.48% accuracy.

3.5 Results for Brazil

Following an identical methodology to the Philippines cases, the binary ensemble model using Brazilian school test score data correctly classified 77.67% of schools as either above- or below-average in terms of the percent of students that passed (see table 1.2). Of particular note in the case of Brazil, the constituent members of the SG-CNN ensemble did not perform well individually, with no single model achieving greater than 68.89% accuracy (Static Maps). This provides additional evidence as to the value of the ensemble approach described in this paper, as traditional approaches using a single source of imagery alone would suffer a 10% degradation in accuracy. In terms of absolute accuracy, the Brazil model performed similarly to our observations in the Philippines (\(MAE = 3.11, \sigma = 3.29\)). The overall estimate of the mean score from imagery was 88.62; this contrasts to an observed mean of 88.75.

4 Discussion

The integration of convolutional neural networks with satellite and other sources of imagery represents an enormous opportunity to “fill in the gaps” in our measurements of socioeconomic factors. This has immediate ramifications for policymakers who seek to allocate aid to improve the adaptive capacity of localities exposed to negative impacts from climate change.

There are many benefits to this style of estimation, as contrasted to survey or other approaches. Because it relies strictly on publicly available data, a deep learning approach can be scaled at low costs to nearly any country or time period
for which imagery is available. It is more readily replicable by stakeholders; a fact of particular value in contexts where the accuracy of public data may be questionable or biased. Additionally, it can be updated as frequently as new remotely sensed information is made available.

This approach has the further benefit of being relatively intuitive, though at costs discussed below. For example, it is reasonably intuitive that a school with a playground and ample greenspace may - on average - perform better than a school without; a school with a temporary shelter may perform worse than one without. Further, because of the CNN architecture employed, heterogeneous patterns in these relationships can be exploited (i.e., playgrounds may not be correlated with improvements in densely populated urban areas).

At the same time, there are many precautions researchers seeking to advance this research agenda should consider; here we note a few key items identified through our own explorations into this topic. First, the literature on extracting and interpreting what ‘features’ are identified by deep learning algorithms is relatively nascent, with no literature available in the context of satellite data to the knowledge of the authors. This means that - while it may be intuitive to suggest a school with a playground could have a better score than one without - there is currently no way to illustrate that a “playground” is identified within the model itself. Research on this topic is ongoing, with ample room for contribution.

Second, CNN algorithms are extremely data hungry, requiring thousands to millions of observations to reliably converge on meaningful results. This is unfortunately at odds with many socioeconomic datasets that could be used for training; i.e., small-scale studies using a few hundred household surveys may not be able to leverage this approach. Ongoing research into transfer learning seeks to mitigate this challenge, but our current understanding of the accuracy of transfer learning is extremely limited.

Third, while we present some limited evidence in this piece on the spatio-temporal extensibility of this style of work, we lack a body of literature examining a range of locales, substantive themes (i.e., education), or times. Further, we have observed relatively little experimentation with different classes of sensors, with the limited number of studies in the public domain currently focused heavily on Landsat and google street view imagery.
Finally, there is a significant gap in our understanding of spatial independence in the context of convolutional networks. This has real ramifications in the work presented here, though we do not engage directly with it. For example, the satellite images for two schools in a single city overlap in many cases, thus biasing our convolutional networks information towards regions with higher densities of schools. An entire research agenda is needed which examines the errors in machine learning that can be attributed to such a lack of independence; researchers are left with very little guidance on such issues today.

5 Conclusion

In this paper, we seek to highlight the potential of convolutional neural networks (“Deep Learning”) - especially multi-source ensembles - for the estimation of socioeconomic factors of relevance to climate vulnerability. We highlight a case study in the Philippines in which we achieve 80% accuracy in our estimation of if a school is above or below-average test score. Further, we provide two additional analyses examining the potential of this approach. First, we show that we are able to predict school outcomes in the Philippines for a second time period with 76.48% accuracy (see 3.4). Second, we test the approach in Brazil using a completely distinct dataset, finding a SG-CNN accuracy of 77.67% and mean absolute errors comparable to the Philippines case study (see 3.5). More broadly, we illustrate that absolute school scores can also be estimated with an apparently reasonable degree of accuracy based on imagery alone, though the utility of such estimated would depend on the specific use case.

The SG-CNN model presented here illustrates how using imagery data from multiple sources can provide a unique window into socioeconomic outcomes of interest to policy and research communities. Relying solely on public information, we provide a solution to estimating school test scores in an open, replicable way in areas where information may otherwise be limited. This illustrates that, by integrating imagery from multiple sources, it is possible to estimate a broader subset of socioeconomic indicators than has been feasible to date. Extending this approach to other socioeconomic factors would have broad implications for our understanding of adaptive capacity - and socioeconomic progress more broadly - in data-scarce environments.
6 Acknowledgements

We would like to thank the committee (Dan Runfola (chair), Anthony Stefani-dis and Matthias Leu. We acknowledge William Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. We would also like to thank the faculty and students of the William Mary geolab (geolab.wm.edu) for their feedback and support.
Bibliography

Andersson Magnus, Ola Hall and Maria Francisca Archila (2019). How Data-Poor Countries Remain Data Poor: Underestimation of Human Settlements in Burkina Faso as Observed from Nighttime Light Data.


Brueederle, Anna and Roland Hodler (2018). Nighttime lights as a proxy for human development at the local level.


Google Scholar


