RESEARCH ARTICLE

A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance

D. Runfola[a,c]

A. Stefanidis[b,c]

Z. Lv[a,d]

Joseph O'Brien[c]

H. Baier[a,c]

[a]Department of Applied Science, William & Mary, Williamsburg, Virginia;
[b]Department of Computer Science, William & Mary, Williamsburg, Virginia, USA
[c]Data Science Program, William & Mary, Williamsburg, Virginia, USA
[d]Virginia Institute for Marine Science, William & Mary, Williamsburg, Virginia, USA

ARTICLE HISTORY
Compiled December 6, 2023

ABSTRACT
Convolutional Neural Networks (CNNs) are leveraged for a wide range of satellite imagery information extraction tasks. However, for tasks which seek to estimate aggregated information across highly variable geographic extents, existing techniques are subject to critical limitations. We engage with a specific case study exploring this challenge: estimating census variables across 2,358 Mexican municipalities, which range in scope from 2.21 km^2 (˜74,000 30 meter pixels) to 72,417.9 km^2 (˜2.4 billion pixels). Building on recent literature which has illustrated the capability of deep learning to extract socioeconomic information from satellite imagery, we specifically seek to establish baseline metrics of error that might be expected when estimating a range of census variables based on coarse-resolution (Landsat) satellite imagery alone. For each of 52 variables, we implement a multi-glimpse recurrent attention model, in which we parametrically determine subsets of each municipality to sample across iterative steps. Results of a 5-fold validation indicate that nearly half of the tested variables (22) can be estimated with r^2 values greater than 0.75. Results suggest considerable promise for the use of satellite imagery to estimate socioeconomic factors in both historic time periods for which surveys were not conducted, as well as contemporary inaccessible regions.

Key Policy Highlights

(1) Findings suggest that satellite based estimates of social factors can aid in "filling in gaps" in census collections for more variables than previously tested.
(2) Metrics of infrastructure (i.e., % of households using wood to cook) were well correlated with features visible in satellite imagery, alongside a selection of socio-demographic measurements (notably metrics associated with children).

CONTACT D. Runfola. Email: danr@wm.edu

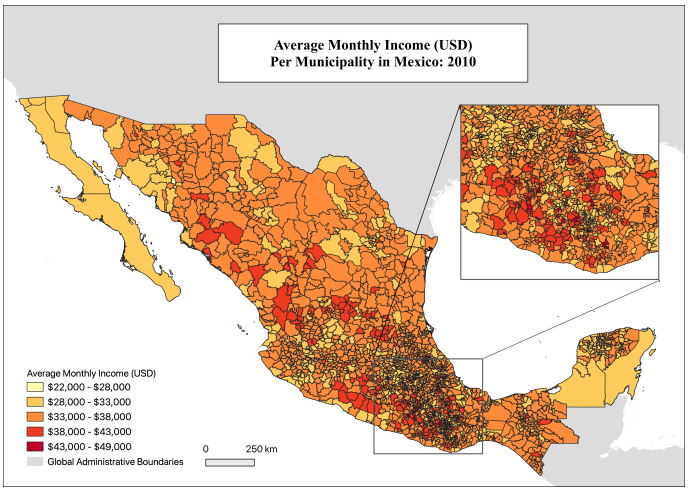

Figure 1. Example of variable size of municipalities within Mexico. Municipalities require between approximately 74,000 and 2.41 billion 30 meter pixels to represent. Municipalities in this figure are colored according to one of the census variables being estimated, average monthly income (Runfola *et al.* 2020a).

(3) The findings in this article help to establish baseline, or floor, accuracy that might be feasible for the presented techniques, limiting analysis to openly available - but relatively coarse resolution - Landsat satellite data.

KEYWORDS
Attention Mechanisms, Remote Sensing, Deep Learning, Convolutional Neural Network, Landsat Imagery

1. Introduction

There is a stark lack of systematically collected census data in developing nations today, inhibiting our understanding of human well-being and concomitant vulnerabilities (Andersson Magnus and Archila 2019, Mossoux Sophie and Canters 2018). This lack of information limits our capability to understand or observe the evolution of social processes, effectively allocate resources to improve human conditions, and to measure the effectiveness of such interventions (c.f. Burke *et al.* (2021), Goodman *et al.* (2019), Runfola *et al.* (2020b), Marty *et al.* (2019), Runfola *et al.* (2017), Leyk *et al.* (2017), Miller Runfola and Napier (2016), Miller Runfola *et al.* (2017), BenYishay *et al.* (2017), Nawrotzki *et al.* (2016)).

In response to this gap, a number of practitioners and scholars have begun to turn to more-regularly collected information from satellite sources, specifically focusing on the use of deep learning to estimate socioeconomic information (Burke *et al.* 2021). These techniques have shown considerable promise as a technique for 'gap filling' socioeconomic data across a growing set of domains (Burke *et al.* 2021, Brewer *et al.* 2021, Runfola *et al.* 2021). Recent work in this emergent sub-field have included satellite-derived estimates of assets (income) and household consumption (Jean *et al.* 2016a, Babenko *et al.* 2017a, Perez *et al.* 2017a, Brewer *et al.* 2023), population and related density metrics (Tiecke *et al.* 2017, Hu *et al.* 2019, Fibæk *et al.* 2021, 2022), educational outcomes (Runfola *et al.* 2021), conflict intensity (Goodman *et al.* 2020), and metrics of infrastructure such as road quality (Cadamuro *et al.* 2018, Brewer *et al.* 2021, Lv *et al.* 2023).

Despite the successes of deep learning models in image classification challenges (Russakovsky *et al.* 2015, Xu *et al.* 2021, Zhou *et al.* 2018) and satellite imagery analyses more specifically (Zhong *et al.* 2017, Xia *et al.* 2017, Cheng *et al.* 2017, Yao *et al.* 2016, Dian *et al.* 2018, He *et al.* 2015), this class of models still faces limitations when applied to satellite information for the purposes of estimating socioeconomic outcomes. Of particular note for the work presented here is the challenge of estimating variables collected across large geographic areas ('large area estimation'), and concomitant concerns about extreme scope variance. It is rare for disaggregate location information to be provided alongside socioeconomic survey data, both to protect the anonymity of respondents, and due to the high resource costs associated with such collections (c.f., Kugler and Fitch (2018)). Thus, spatially-explicit survey metrics are generally provided as an aggregate to some geographic unit - i.e., a census block or municipality, rather than a latitude-and-longitude of a household (Kugler and Fitch 2018, Ruggles 1995). However, the geographic regions to which socioeconomic data is most commonly aggregated are not uniform in nature. In Mexico - the focus of this study - the size of these regions can range from 2.21 km^2 (˜74,000 30 meter pixels) to 72,417.9 km^2 (˜2.4 billion pixels; see figure 1).

This contrasts to traditional image analysis, in which input images frequently have dimensions that are - if not identical - broadly similar (i.e., many well-known benchmarks contain images of standardized dimensions (Krizhevsky *et al.* 2017)). In practical application, images that do not conform to selected input dimensions are warped, reshaped, or zero-padded to a standard set of dimensions using a range of interpolation algorithms (Lecun *et al.* 2015). This process is necessary in most convolutional approaches, as the number of convolutions and parameters necessary to fit a given network are frequently dependent on having a standardized input (Lecun *et al.* 2015). A small subset of past literature has engaged with this problem, predominantly exploring solutions to variable-dimension input challenges in the context of scene classification (e.g., determining whether an airplane, boat, or other feature is present in an image; see He *et al.* (2015), Simonyan and Zisserman (2015), Wu *et al.* (2017), Xie *et al.* (2019)). To the authors knowledge, no papers have explored this challenge in the context of large area variable estimation; more broadly, existing models cannot be employed in this context without extreme levels of data loss (i.e., center-cropping).

We bridge this gap, focusing on the most ubiquitous source of socioeconomic survey information - country census instruments. We specifically explore an illustrative example using Mexico's 2010 census, which is available aggregated to each of Mexico's 2,358 municipalities (i.e., a single value is made available for each municipality, for each variable we seek to test). To provide an explicit example of the challenge this paper engages with, a single Landsat satellite image covering the Mexican municipality of Baja California Sur, at a 30 meter resolution, requires image dimensions of approximately 21,670 by 18,950 pixels; this contrasts to Natividad, which requires an image of 80 x 60 pixels. Attempting to scale or pad either of these images to a set dimension would result in untenable introduction of noise and/or data loss.

In the context of this methodological challenge, recent research has highlighted the importance of improving our understanding of what types of socioeconomic variables are reasonable to estimate using space-borne instruments (Burke *et al.* 2021). This paper also responds to this call, explicitly seeking to identify baseline accuracy that might be expected if satellite data is used to estimate a wide range of census variables. We make two specific contributions to the literature. First, we introduce a multi-glimpse technique for analyzing units of observation (census units) that are represented by imagery of widely varying dimensions. Second, we apply this technique to publicly

available satellite imagery (Landsat) to estimate 52 variables from Mexico's 2010 Population and Housing Census, providing the first evidence of the efficacy of leveraging satellite imagery to conduct large area estimation traditionally only conducted with in situ survey instruments.

1.1. *Literature: Estimating Socioeconomic Information with Satellite Imagery and Deep Learning*

Estimating socioeconomic factors from satellite imagery using deep learning techniques is a relatively recent topic in the literature, with some of the first pieces exploring the use of convolutional neural networks (CNNs) to predict poverty using a combination of high resolution imagery and nighttime lights being published circa 2016 (Jean *et al.* 2016a, Xie *et al.* 2016a). One of the earliest pieces in this domain (Jean *et al.* 2016b) showed that CNN approaches could explain up to 75% of the variation in local-level economic outcomes in sub-Saharan Africa. Building on this work, researchers showed that these CNN techniques were identifying and exploiting correlations with man-made structures including roads, buildings, and farmlands (Xie *et al.* 2016b).

In later studies comparing model architectures and imagery resolutions, researchers found that similar techniques employed in Mexico may be of use, identifying that a GoogleNet model trained on Digital Globe imagery performed slightly better than other approaches in explaining poverty variance in 896 Mexico municipalities (Babenko *et al.* 2017b). However, researchers concluded that the model accuracy did not carry over to municipalities outside the 896 included in the 2014 Módulo de Condiciones Socioeconómicas-Encuesta Nacional de Ingresos y Gastos de los Hogares (MCS-ENIGH) survey (a survey of socioeconomic conditions in Mexico). Researchers hypothesized that this lack of generalizability could be due to weighting geographic tiles by land area rather than population, or that MCS-ENIGH municipalities have characteristics different from non-MCS-ENIGH municipalities. The authors conclude that more work is needed to determine the ways in which training processes influence sample validation.

A number of other studies have explored alternative architectures and approaches to the estimation of poverty information. Some of the earlier tested approaches, Tile2Vec (Jean *et al.* 2019) and generative adversarial network (GAN) approaches (Perez *et al.* 2019) illustrated the considerable promise of these techniques. Tile2Vec, an unsupervised representation learning algorithm that uses theories from linguistics and applies them to geospatial information, was able to explain 49.6% of the variance in poverty data in Uganda (Jean *et al.* 2019), as contrasted to transfer learning approaches which explained 41% of the variance (Jean *et al.* 2016b). In the GAN approach, the authors achieved a validation accuracy of up to 68% when using only visible (RGB) bands (Perez *et al.* 2019). Other researchers have contrasted a range of more common model architectures, finding that techniques such as focal loss can provide considerable benefits when paired with DenseNet (Ni *et al.* 2020).

A range of teams have explored the capability of models trained on data from a broader range of sources, including the National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) Day/Night Band (DNB) nighttime light (NTL) data, Google satellite imagery, land cover and road maps, NDVI, and location data, to predict household wealth (Zhao *et al.* 2019, Kondmann and Zhu 2020, Chen 2017). In one of the initial models of these types, extracted features from these datasets were able to successfully predict a household wealth index with an R^2

value of 0.70 in Bangladesh and 0.61 in Nepal (Zhao *et al.* 2019). The researchers also identified that proximity to urban areas was the most important variable in explaining poverty, contributing 37.9% of the explanatory power. Further contributions to the literature have been made by teams who showed that these techniques generalize to areas in Mexico (Babenko *et al.* 2017b), the Philippines and Thailand (Hofer *et al.* 2020), India (Daoud *et al.* 2021), and Brazil (Castro and Álvarez 2023).

Recent work has begun to explicitly look for ways to leverage only open-source data to estimate socioeconomic indicators, recognizing the prohibitive costs associated with downloading imagery and training deep learning models using proprietary (satellite) information. Building on the work done by Jean et al. in 2016, (Jean *et al.* 2016b), researchers showed that models trained using lower-resolution, publicly available Landsat 7 imagery can achieve accuracy that exceeds previous benchmarks (Perez *et al.* 2017b). This increase in performance while using lower-resolution imagery was theorized to be due to the inclusion of hyperspectral imagery bands - information which is unavailable in many high resolution image products. However, this research again illustrated the challenge of external validity, as the techniques employed struggled to generalize across country boundaries. Other work has focused on using publicly available crowd-sourced geospatial information in the Philippines (Tingzon *et al.* 2019). Here, the research team was able to achieve a R^2 value of 0.63 for estimating asset-based wealth, and showed that models trained on a combination of OpenStreetMap (OSM) and nighttime light-derived features are effective for real-time poverty mapping. Other research has proposed a novel reinforcement learning algorithm that aims to decrease the number of high-resolution imagery tiles needed to accurately predict consumption and other economic outcomes, thus mitigating image acquisition costs (Ayush *et al.* 2021).

Despite these successes, a number of open questions remain as to the external validity and robustness of estimates. For example, researchers found that these methods struggled to quantify changes in the economic status of communities in Rwanda between 2005 and 2015 (Kondmann and Zhu 2020). Further, other authors have concluded that spatial perturbation injected in the coordinates of poverty indicators significantly reduces the prediction power of the models (Jarry *et al.* 2021). As many of the underlying survey datasets used to train these models have spatial perturbations integrated into them to promote anonymization, these challenges are important to consider (Jarry *et al.* 2021). Other authors have noted that uptake in the policy community for these techniques is likely to remain low due to the lack of interpretability using existing tools (Yeh *et al.* 2020).

1.1.1. Predicting Other Socioeconomic Indicators

Building on the body of literature seeking to estimate metrics related to income, a small subset of researchers have begun to explore our ability to estimate a broader range of socioeconomic indicators. Studies showed that a Resnet50 model trained on nighttime lights was capable of predicting gross domestic product (GDP) and total retail sales of consumer goods (TRSCG) with a Pearson coefficient of 0.85 (Wu and Tan 2019a,b). In experiments in Chongqing, China, researchers showed that a range of variables, including general public budgetary expenditure (GPBE), per-capita living expenditure of rural residents (PCLERR), and per-capita disposable income of rural residents (PCDIRR), could be estimated (Tan *et al.* 2020).

Initial research into generalization of these techniques to other sociodemographic indicators in Africa did not show promise, including educational attainment, access to

drinking water, and a variety of health-related indicators (Head *et al.* 2017). However, recent work exploring the use of deep learning in predicting education outcomes showed that the estimation of school test scores using could be implemented with an accuracy between 76% to 80% (Runfola *et al.* 2022a). Further, work leveraging the Demographic and Health Survey (DHS) data to predict health outcomes has shown that, while model errors remain, many of the errors were only one class off, and were broadly reflective of errors also common to manual human interpretation (Irvin *et al.* 2017). More recent work has further shown that data fusion which combines census and satellite imagery can be used to predict human migratory flows with an accuracy of R^2=0.72, an improvement over models leveraging only socioeconomic data by 10% (Runfola *et al.* 2022b). Some application of these techniques has started to emerge in the context of evaluation of anti-poverty programs using satellite imagery (Huang *et al.* 2021).

2. Materials & Methods

2.1. *Data & Study Area*

We seek to establish the capability of satellite imagery alone to estimate socioeconomic information collected during the Mexican Census. Mexico was selected for these tests due to both (a) the availability of geographically explicit (municipality) census data, and (b) the relative infrequency of country-wide surveys available in Mexico today (promoting the usefulness of the derived model; (Ruggles *et al.* 2003)). Census data is collected from the 2010 Population and Housing Census conducted by the Instituto Nacional de Estadística, Geografía e Informática (INEGI), distributed by IPUMS (Ruggles *et al.* 2003). The census was conducted during May and June of 2010, and was the result of a 10% sample (N=11,938,402). Each household in the census contains information on both (a) the municipality in which the household exists, and (b) the relative weight of that household; these variables allow for the generation of municipality-scale aggregations of the underlying information. In the case of discrete variables, we calculate the weighted percentage of households which have a given response (i.e., piped water).

A set of 52 variables are selected to explore, as shown in table 1. Variables were selected to be representative of a wide range of different socioeconomic characteristics, and are broadly characterized as being a member of "Sociodemographic", "Infrastructure", or "Wellbeing" categorizations. Sociodemographic factors include - for example - the average number of children in a household, % of the population that is single, or the % of the population that is disabled. Broadly, we theorize that these factors will have lower overall correlations with factors in satellite imagery, but note that there may be relationships between landscape characteristics and social phenomena - i.e., a higher average age may be associated with environmental conditions that attract retirement communities. The second category of variables represent infrastructure characteristics captured by the census - for example, the percent of households that use electricity for cooking purposes, or own a television. We anticipate a higher level of correlation with these variables, given their direct correlation with land use types (i.e., urban environments are more likely to have such infrastructure in place). Finally, variables measuring wellbeing include factors about individuals such as their level of schooling. We expect some correlations to be present in these variables, i.e. given the presence of schools, but note that key features may be obscured by the 30

Table 1. Summary of accuracy of estimates for each variable tested. * denotes cases in which the range of r^2 values across the 5 model tests was less than 0.25; ** denotes cases with ranges less than 0.10.

Sociodemographics			**Infrastructure & Wellbeing**		
Name	**MAE**	**R2**	**Name**	**MAE**	**R2**
Avg. Chld. Survive**	0.03	0.88	% Elec.	0.02	0.54
Avg. Inc.*	39874	0.17	% No Piped Water	0.09	0.27
Earned Inc.	266.12	0.6	% Sewage System	0.14	0.6
% Rural	0.18	0.44	% Elec. Cooking	0	0.6
% Owned Household	0.03	0.81	% Phone	0.04	0.84
% Married w/ Child.	0.03	0.68	% Cellphone	0.08	0.8
% Single Parent	0.01	0.61	% Internet	0.04	0.47
% Extended Family	0.02	0.76	% Burn Trash	0.08	0.84
Avg. # Families*	0.01	0.01	% Auto	0.07	0.76
% Single	0.02	0.57	% Hotwater	0.07	0.8
% Married	0.02	0.54	% Computer	0.04	0.68
Avg. Children Born*	0.09	0.51	% Washer*	0.07	0.86
Tot. Pop.*	35072	0.45	% Refrig.**	0.05	0.91
Avg. Yrs Since Birth*	0.34	0.84	% TV**	0.03	0.91
Avg. Dead Children*	0.01	0.88	% Radio*	0.03	0.89
% Foreign Born	0	0.42	Avg. Rooms	0.16	0.81
% Employed	0.04	0.33	% No Toilet	0.04	0.3
% Disabled**	0.01	-0.05	% Any School	0.02	0.58
% No Health Cov.	0.09	0.54	% Literate*	0.02	0.85
Avg. Age*	2.15	0.31	% Less than Primary*	0.02	0.88
Int. Migrants	1064	0.66	% Primary*	0.01	0.87
% No Meal Last 30	0.03	0.51	% Secondary	0.02	0.65
Avg. Family Size	0.27	0.6	%University	0.01	0.39
Avg. # Children	0.02	0.84	% Wood Cooking*	0.09	0.84
Avg # Children<5**	0.01	0.87	% Trash Collection	0.09	0.82
Avg. Age of Death	0.01	0.6	Avg. Years of School*	0.26	0.87
Mean r^2		0.55	**Mean** r^2		0.72

meter resolution imagery leveraged.

Landsat satellite imagery is leveraged in this study, selected due to its' open, public nature, availability of historic imagery, and global scope (Woodcock *et al.* 2008). To promote transferability to other sensors, we specifically use the red, green and blue bands of the Landsat 5 TM level-1 data product (USGS 2019). We generate a cloud-free annual composite using monthly imagery from the year 2010, for each municipality (Google Earth Engine 2021). Broadly, we suggest the relatively coarse resolution of Landsat information (30 meters) will provide a helpful baseline as to the level of accuracy that may be expected from this style of approach.

A small number of pre-processing steps are employed before imagery is introduced into the model. First, we crop each image to the bounds of a given municipality. Second, we standardize the pixel values, and apply a brightness factor of 2 in order to facilitate model optimization (i.e., mitigating the potential for gradient decay due to the presence of large quantities of small pixel values) and visualization. Finally, we convert each image to a tensor representation for input into the model.

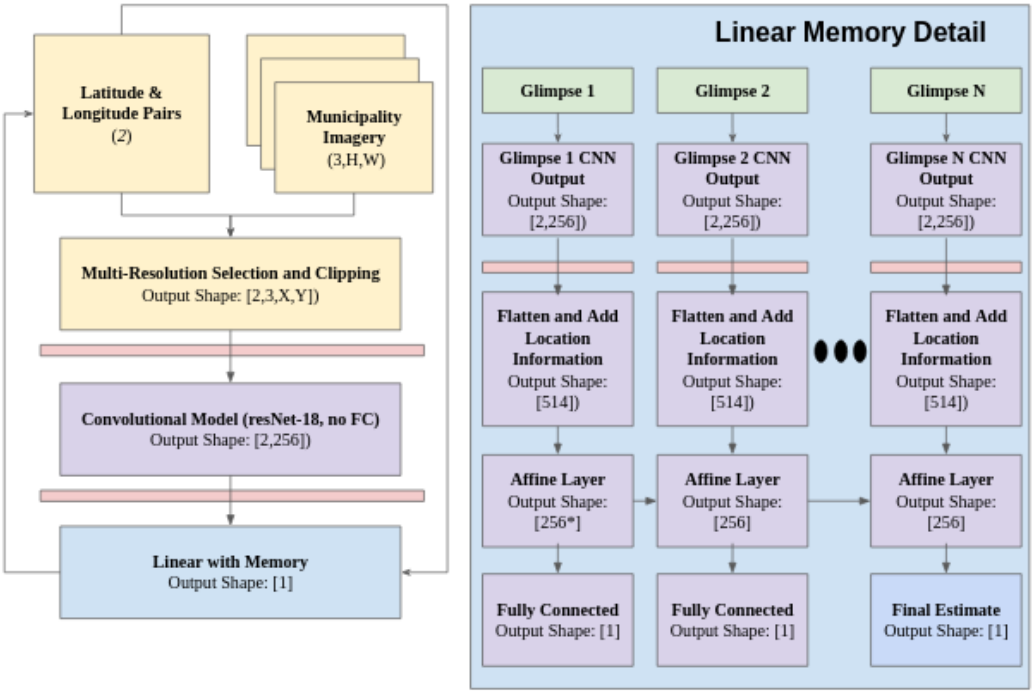

Figure 2. Overall Model Architecture. This figure illustrates each stage of the model architecture employed, inclusive of our implementation of geographic attention. Horizontal red lines indicate ReLU activation layers. *During the first glimpse, the affine layer takes as input a vector of [514 + 256], with the additional 256 elements being intitialized as 0s. In subsequent glimpses, this input shape changes to [514 + 256], with the additional 256 elements representative of the hidden state from the previous glimpse.

2.2. *Methods*

2.2.1. Overall Model Architecture

We leverage an identical modeling procedure for each of the 52 census variables tested, applying a multi-glimpse based, recurrent model of visual attention not previously applied in the context of satellite imagery (Mnih *et al.* 2014). To mitigate the small-N nature of our dataset, we apply a transfer learning approach to the convolutional layer within the architecture, using weights pre-trained on ImageNet that have previously been shown to be effective in the context of satellite imagery (Goodman *et al.* 2020, Runfola *et al.* 2021, Jean *et al.* 2016a). The multi-glimpse architecture we employ is summarized in Fig. 2, and detailed further in the following section.

2.2.2. Formal definition of the Convolutional Network

Convolutional neural networks (CNNs) rely on a set of convolutional layers, each with a defined filter which is used in the convolutional process to produce features (generally represented as tensors) representative of elements of importance within the image (Lecun *et al.* 2015). To formally define a CNN for the purposes of describing how our multi-glimpse model is implemented, first let $X = \{X_{w,h,c,i=1}, ..., X_{w,h,c,i=n}\}$ represent a set of n input images with width w, height h, and channels c.

Additionally, let $F_l = \{F_{k,k,c,j=1}, ..., F_{k,k,c,j=f}\}$, where F is a set of filters to be used in the convolutional process within layer l, k are the filter dimensions, c is the channels to which a filter will be applied, j the index of the filter, and f the total number of

filters. Weights for each filter, for each convolutional layer, are defined in $W_{j,l}$, with index j and l representing the filter and layer, respectively. Following this, the output of any given layer can thus be obtained by:
$Y_{j,l} = X_i \otimes W_{j,l}$

In most contexts, filter dimensions F_l become iteratively smaller as layer l increases, at which point an affine (or, fully connected) layer is utilized to produce a final score for a given input X_i. This final affine layer most commonly takes the form of a multi-layer neural network in which all nodes are connected to all other nodes in the following layer.

2.2.3. Multi-Glimpse Geographic Attention

A core challenge in the use of satellite imagery for the estimation of Mexican census information with convolutional models is the highly variable spatial dimensions of regions of interest (N=2358 municipalities), which range in area from 2.21 km^2 to 72,417.9 km^2. To mitigate this challenge, we apply a model which incorporates a multi-glimpse geographic attention approach- conceptually, allowing the model to iteratively apply convolutions to sampled, similarly-sized (in terms of w and h in the above notation) regions of each municipality, and training the model to bias samples towards regions that are most relevant (i.e., ignoring large stretches of desert). This multi-glimpse procedure is implemented following a number of steps, in which:

(1) Within each municipality, a parameterized distribution of latitude and longitude coordinate is sampled, generating latitude-longitude pairs for each glimpse.
(2) Two images are clipped from the municipality-scale Landsat cloud free mosaic based on the latitude and longitude selected in step 1, with one image representing a coarse-scale subset, and the second a zoomed-in region.
(3) For each glimpse, these images are passed into a resNet-18, pretrained on ImageNet.
(4) A linear layer with a memory function takes in each glimpse sequentially to produce the final estimate of a value for a municipality census variable; glimpses are moved throughout the image on the basis of a hidden element within this linear layer.
(5) The true values for each municipality are contrasted to the estimated aggregations, and the total difference (mean absolute error) is used in a backpropogation procedure to update the parameters across the neural network.

In step one of the above procedure, latitude and longitude pairs are sampled from a parameterized gaussian distribution, in which the mean coordinates are constructed as parameters which are updated during the training process. The selection of a gaussian distribution encourages the first samples in the training process to be biased towards the center of the image; we provide more thoughts on the rationale and implications of this choice in our discussion. Samples are clamped to the minimum and maximum coordinates of a given municipality (with coordinates normalized to a -1, 1 range to facilitate sampling across all municipalities). This is formalized in notation as:

$$\begin{aligned} l_{i,x} &= \mathcal{N}(\mu_x, \sigma = 0.1) \\ l_{i,y} &= \mathcal{N}(\mu_y, \sigma = 0.1) \end{aligned} \tag{1}$$

The parameters μ_x and μy are themselves estimated as the output of a small linear network which takes as input the hidden node values of the convolutional layer of the

previous image - i.e., what (if any) features are detected in the previous glimpse. This allows for a dynamic strategy in which each glimpse is conditioned on the nature of the features detected in the previous glimpse - i.e., if an urban area is in a first glimpse, we may parameterize so as to preference moving a short or far distance away for the next glimpse, contingent on what tends to perform best. This broadly allows for geographic attention to different areas within a municipality, irrespective of the size of a given municipality.

Once sampled, for each latitude and longitude pair two images are generated, with initial image dimensions based on the size of the input municipality. The first image is selected with a centroid of the selected latitude and longitude, and image dimensions X and Y equal to (in pixels):

$$X, Y = min(int(min(H, W)/5), 50) \tag{2}$$

where H is the height of the satellite image of the target municipality, and W is the width. A second, zoomed in image is then sampled from the same area, following past research indicating such multi-resolution information can promote higher accuracy in recurrent attention models (Mnih *et al.* 2014). This second image retains the same centroid as the first, but has dimensions 75% smaller than the first image. In practice, this equation enables us to take larger windows of pixels for larger municipalities, while scaling to smaller windows for smaller cases; the scaling factor in equation 2 - set to "5" - determines the relative size between cases. [1]

These images are then fed forward into a pre-trained resNet18 model, with the output vectors of the final convolutional layer saved into two vectors, one for each image. The fully connected layer of the resNet18 is removed, instead keeping the 256 length feature vector associated with each input image, thus generating two 256-length vectors for each municipality, one for each scale of imagery.

This vector (of dimensions [2,256]) is then fed into a recurrent linear layer with memory[2], alongside a vector of length 2 including the latitude and longitude information from where the images were generated. Figure 2 shows a zoomed-in example of the recurrent linear memory implementation used in this analysis. During the first glimpse, we take the [2,256] output from the resNet18 and flatten it (output size of 512), concatenating the information to the two latitude and longitude coordinates selected (for a size of 514), and then add 256 0s, which will be leveraged in future glimpses for memory (i.e., we initialize the first glimpse with no memory information). This 770 element vector is then passed into an affine layer with an output of 256 elements, which is in turn fed into (A) a final affine layer to generate the single estimate for a given value, and (B) a separate affine layer which estimates the new latitude and longitude coordinates for the next glimpse. A second glimpse is then taken, and the process is repeated. During the second glimpse, the affine layer's memory is updated to include the previous glimpse's affine layer output. In this implementation, after four glimpses are taken, the final estimate is generated based on all preceding steps, and is then used to update the network parameters.

[1] Of note, these images can still be of variable dimensions across different municipalities, but the scale of difference - and thus potential information loss due to implicit scaling - is far lower than in the source imagery. We discuss the implications of this further in the discussion.

[2] This follows previous literature (Mnih *et al.* 2014), and outperformed more complex LSTM-based implementations for the final layer of the architecture; we observed the same under-performance of LSTM in ad hoc tests.

2.2.4. Tuning, Training & Validation

Each model was fit across 200 epochs, with an early stopping if loss did not decrease across any 50 epoch interval. Adam optimization was employed, with a learning rate of .0003 (selected via a grid search). For each of the 52 variables in table 1, five models were fit using a monte carlo validation strategy, in which 5 random subsets of 25% of the data were withheld, and the remaining 75% used for training in each case (for a total of 260 models, 5 models for each variable). The r^2 and mean absolute error for the witheld test set each model iteration was recorded; the mean of these values across all five model iterations are presented, by variable, in table 1. Of note, these r^2 values may resolve to values less than 0 (i.e., negative values) in cases where the prediction is worse than simply predicting the mean.

3. Results

Across all model runs, 22 variables were estimated with an average accuracy (r^2) of 0.75 or higher (out of the 52 tested cases)[3]. The top-3 best performing variables - the percentage of households with televisions, refrigerators, and radios - each achieved average r^2 values of 0.89 or higher; further, these cases were very consistent across the five folds of the data, with little variance in r^2 values irrespective of the random split between training and testing data. Other consistent, strong performers include:

(1) Variables related to child health, including the average number of dead children in a household, number of children under 5, number of years since last birth, and total number of children.
(2) Education related variables, including the population with less than primary education, number of years of schooling, and literacy rates.
(3) A selection of other infrastructure-related variables, including the presence of a washing machine and wood fuel being used for cooking.

Contrasting to these cases, a number of variables performed consistently poorly, irrespective of model iteration. These included total population, average age, average income, number of families, and the worst performing case in this study, percentage of the population which is disabled.

A number of test cases had relatively high mean accuracy, but across the 5 k-fold samples had ranges of r^2 exceeding 0.25, indicating a high volatility as a function of the randomly drawn sample. The strongest of these cases were the percentage of households with a phone, percentages of households that have trash burned or collected, and the average number of rooms. Variables that performed poorly but had large ranges of r^2 across runs included the percentage of households with no piped water or toilet, percent employed, and percent of the population with access to the internet.

Figure 3 shows scatterplots for a selection of four variables of interest, indicating the strength of fit between the satellite imagery derived estimates (y axis) and census metrics (x axis) for the validation dataset. Each point is a municipality. These values are from a single fold of the validation process (4^{th}). Figure 4 shows the spatial distribution of the average prediction and estimated values for Earned Income, which had an accuracy similar to that found in much of the broader literature (R^2 of 0.6).

Figure 5 shows the relationship between absolute error and observed income. Of

[3] r^2 is the focus of this study, due to it's ubiquity in the target social sciences this methodology would support. MAE is also reported.

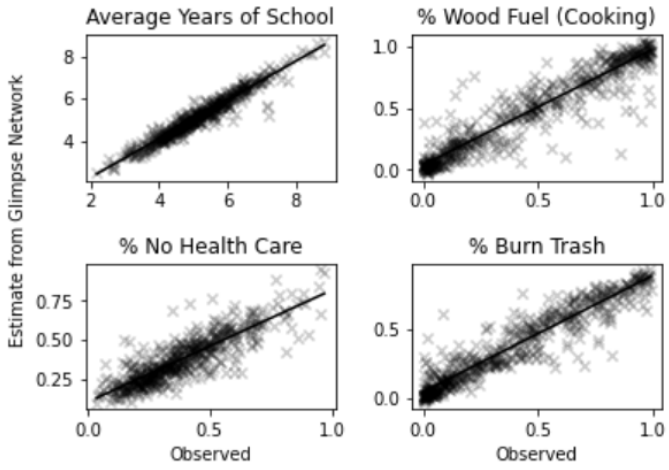

Figure 3. Scatterplots for a selection of examples. This figure shows the scatterplots for four exemplar variables, illustrating the tightness of the fit for each example case.

this selection of four variables, the strongest observed fit (r^2) is .001, indicating no prediction bias toward low income or high income areas. As a second test for bias along socioeconomic dimensions, table A1 provides correlations between all variables error and educational attainment; no strong correlations were observed.

Figure 6 shows the relationship between absolute error and the observed value for each of four values (i.e., a test for heteroskedacitiy), in which each point on the figure is a single municipality. For this selection of variables, there was no strong correlation between increasing observed values and the error in the estimate from satellite imagery (see appendix 1 for results across all variables).

Figure 7 illustrates the evolution of the loss function value (MAE) over epoch for each of four variables. Of note, % disabled represented one of the poorest performing variables in the analysis, with a r^2 value of -.05; conversely, the average number of surviving children ($r^2 = 0.88$) and average years of school ($r^2 = 0.87$) represent two of the best performing variables.

4. Discussion & Conclusion

4.1. *Discussion*

This article explores the ability of deep learning models to identify features in satellite imagery that may be correlated with a range of socioeconomic factors, and the degree to which we can use those features to make estimates. We specifically tested this at a municipality scale, thus focusing on large area estimation. We found that such an approach is effective for a wider range of variables than had previously been studied in the literature, but is not effective for every class of variable.

Table 1 shows a synopsis of all tested variables. The strong performance of many infrastructure variables was not surprising, given those variables are the ones most likely to be correlated with physical structures visible from satellite imagery. For example, one of the best performing variables was the percent of households with access to a refrigerator; this is very likely to be tightly correlated with electrification, and thus a wide range of urban features (i.e., dense road networks) that are visible from imagery. Even more directly, the average number of rooms had a r^2 value of 0.81 -

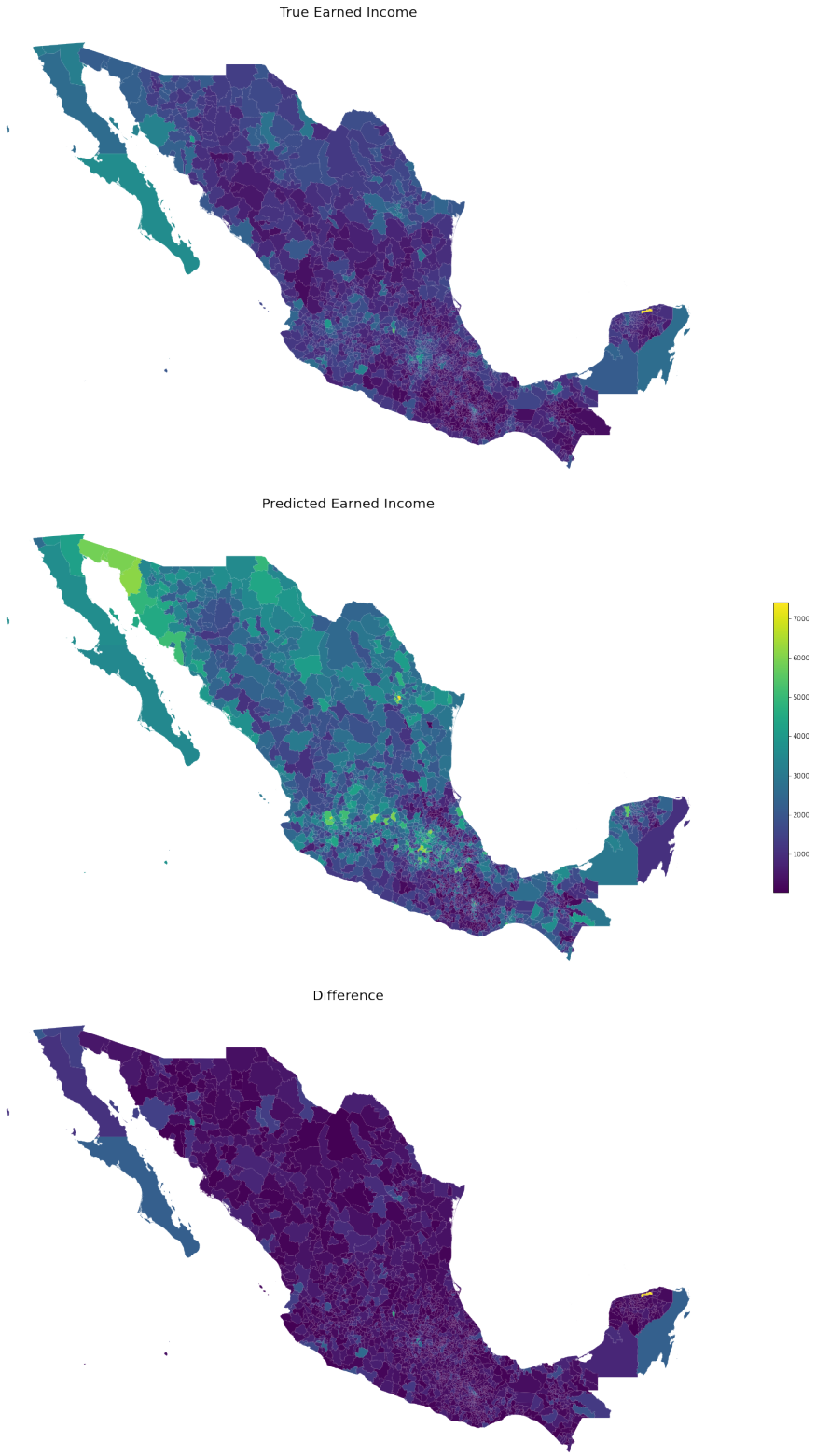

Figure 4. Observed and predicted values for the variable Earned Income. This figure shows the observed and predicted values for the variable Earned Income, a moderately-well performing variable with R^2 of 0.6.

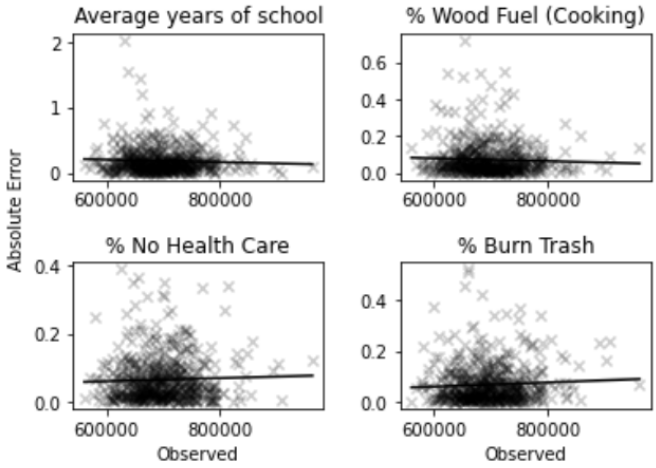

Figure 5. Correlation between absolute error and observed income. To explore if error may be biased towards low income areas, this figure shows the relationship between absolute error for a selection of four variables (y axis) and observed income values (x axis). For example, the figure labeled "Average years of school" suggests that the rate of error in our estimates for the average years of school individuals have obtained is not correlated with observed income level - i.e., there is no bias towards wealthy or poor areas. Of this selection of variables, the strongest observed fit (r^2) is .001.

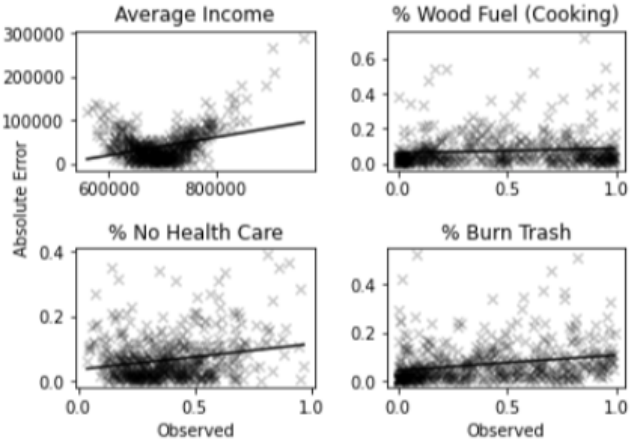

Figure 6. Correlation between absolute error and the observed value for four variables.

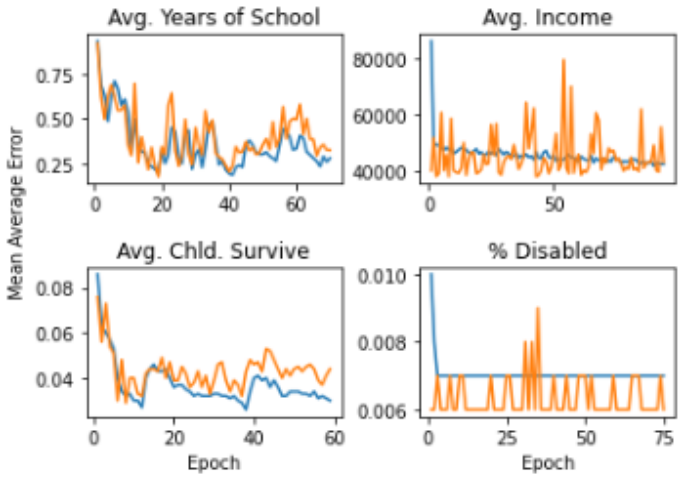

Figure 7. Relationship between loss function value (MAE) and epoch during training runs for each of four variables. This figure illustrates the training pattern for each of four variables, with validation loss in orange and training loss in blue.

likely representative of the density of buildings discernible in a given area.

There are a small number of cases that performed more poorly than anticipated. Two of these - the % of households with no piped water ($r^2 = 0.27$) and the percent of households with no toilet ($r^2 = 0.3$) were particularly surprising, given that many such households are likely to be located in rural regions that would have clearly discernible features (i.e., a lack of roadways, increased cropland, no power lines, and other factors). Notably, this is not due to a lack of variability in the measured variable; some municipalities in Mexico have nearly complete coverage with water systems, while others have only a small percentage of households with waterline access (less than 2% in extreme cases). This suggests some important limitations of this study, which will be further explored in section 4.1.3.

Social factors that were not directly related to infrastructure generally performed more poorly than their infrastructure counterparts, as was anticipated. Average age, average income, number of families, and the worst performing case in this study, percentage of the population which is disabled all had r^2 scores well below what may be useful in practice. However, variables associated with children specifically tended to outperform expectation. The average number of children surviving in a household ($r^2 = 0.88$), average number of children born ($r^2 = 0.84$), and average number of children under five ($r^2 = 0.87$) were among the best performing variables across the full study. Given the coarse resolution of the Landsat imagery utilized, we hypothesize that suburban and urban infrastructure may be tightly correlated with positive child outcomes; however, further study is required to better understand the explicit correlations being identified for these specific variables (see section 4.1.3).

4.1.1. Error and Low Socioeconomic Status Areas

Because the methods proposed here are likely of highest use in relatively low-income or sparsely populated areas where survey instruments are less commonly deployed, we follow much of the existing literature (Jean *et al.* 2016a, Runfola *et al.* 2021) in exploring if their is any bias in error towards less affluent regions (i.e., is there more error in areas with lower socioeconomic status?). These tests are shown for

a small selection of variables in figure 5, and table A2 in the appendix provides a synopsis for all variables. These findings show no clear bias towards any particular socioeconomic group - i.e., errors appear to be consistent irrespective of the magnitude of measurement across all variables. Further, this seems to be true for tests using both income and educational attainment as the social dimension along which bias is explored.

4.1.2. Training Across Variables

We leveraged a single set of hyperparameters to train all variables in our dataset; while these hyperparameters were identified as a part of a large adhoc set of tests, they are unlikely to be optimal for all cases. Figure 7 shows the mean absolute error across epochs of training for a selection of four different variables with different rates of error ($r^2 = 0.87$, 0.17, 0.88 and -.05 for the average years of school, income, number of surviving children, and % of the population which is disabled, respectively). As this figure shows, the two poorly performing cases (average income and % disabled) quickly converged on a solution, from which relatively little (income) or no (% disabled) improvement was made. This suggests that tailored learning rates or other strategies may be able to contribute to improving the estimates for these cases. We expect that average income, in particular, could benefit from tuning of model hyperparameters given that learning continues to occur throughout the training process (contrasted to the % disabled case, in which no further gains are made after the first few epochs). Additionally of note, in the two strong performing cases (average years of school and children surviving), the optimal solution was found early in the training process - around or before epoch 20.

In addition to standardized hyperparemeters, some of the architectural model choices made could also have a differential impact across variables. The most prominent example of this is in the use of a parameterized gaussian distribution to sample the latitude and longitude location for each "glimpse". By using a gaussian distribution (rather than a uniform distribution), we intentionally bias the initial selection of image tiles towards the center of each municipality. This bias has the advantage of ensuring that the initial glimpse will have very few "no data" pixels - i.e., pixels that may fall outside of the municipality in question. However, it may also have a detrimental effect on variables that require imagery more likely to exist on the edges of municipalities. As urban centers tend to be located closer to the centers of municipalities in Mexico, this could particularly cause bias in variables sensitive to information on cropland that may be more frequently located on the edges of municipal regions.

A second example of this is the selection of the scaling factor for images, a value which controls the window size used for glimpses. Here, larger municipalities leverage larger window sizes, with the exact scale being determined by dividing the height and width of the municipality by a scaling factor (5), and constructing a square region based on the minimum of either height or width (as detailed in equation 2). This scaling factor helps to dramatically reduce the variation in size across municipalities, but can come with some drawbacks that may be detrimental to use. For example, if a scaling factor was set to '1', the glimpses would always take in the full image of a municipality; as the value increases, the size of the glimpse relative to the municipality decreases. Thus, bigger scaling factors result in more efficient computation, at the cost of smaller glimpse sizes. Identifying the optimal glimpse size for a given variable could improve model accuracy in cases where relevant image features are located in diffuse locations across a municipality, but would do so at increased computational costs.

4.1.3. Limitations & Future Directions

This study has a number of key limitations, and concomitant directions for future work. As alluded to in the context of water access earlier in the discussion, a core limitation is in our inability to examine what features are being identified as correlates for each variable. This inhibits our ability to understand and improve model performance for both strong performing variables for which we would like to identify if features are extendable to other contexts (i.e., is education correlated with the presence of features that may be school buildings?) as well as poor performing variables (i.e., what features are being missed that may be useful for estimating water access?). In addition to precluding many diagnostic strategies, this directly impedes our ability to generalize to other contexts - i.e., domain adaptation - as we do not know what identified correlates may be data dependent, and which may be reflective of true infrastructure correlates with socioeconomic factors. Future work which focuses on visualizing the key features being leveraged within each prediction would be of high value in translating these approaches into practical use.

A second major limitation is in the number of glimpses employed. Here, we limit the model to four glimpses in order to promote reasonable computational speeds that might be employed for all of Mexico across the dozens of variables we explored. While not explicitly tested, it is likely that increasing the number of glimpses allowed would improve model performance - especially for variables that may need to detect features that are not common on the landscape. In particular, variables such as water access may be correlated closely with water features that may or may not be captured in four glimpses; increasing the number of glimpses would increase the probability such features are observed. Future work may even explore dynamic numbers of glimpses contingent upon the size of the municipality under consideration, allowing for more glimpses if the region to be studied is larger.

A third major limitation is the resolution of the Landsat satellite sensor (30 meter spatial), and our choice to leverage only 3 bands (red, green and blue). As it pertains to spatial resolution, past research has shown that for the estimation of income, features that may only be detectable using high resolution imagery (i.e., close to one meter) may be key (Jean *et al.* 2016a). While the work presented in this article is notably different than past work in that it focuses on large area estimation, it is still likely that overall accuracy is being driven downward in some cases due to an inability to discern meaningful features with coarse data. Future research which explores smaller scopes with multiple resolutions could help to shed light on the relative importance (or lack thereof) of higher resolution instruments.

We intentionally chose to limit this analysis to visible bands, motivated by the common need to implement models on sensors with fewer bands, and the concomitant value of using the weights provided in this paper as starting points for transfer learning to those sensors. For example, a researcher could use the weights we identified for income in this piece as a starting point for fine tuning a model leveraging high-resolution, 3-band satellite imagery by another imagery provider. However, by limiting the bands we leveraged to only the three visible bands, we have also limited the potential accuracy of our approach. Further research into the explicit bands that are most useful for sensing socioeconomic factors would be of high value, and something that has been called for by the community for some time Liverman and Cuesta (2008), Kugler *et al.* (2019), Hunter *et al.* (1999).

4.1.4. Contributions

This paper makes two key contributions to the literature. First, we present a multi-glimpse strategy for estimating values across large areas (municipalities in Mexico) which have dramatic variation in their geographic scopes. This strategy provides a computationally scalable solution to large area estimation, in which users can choose to add additional glimpses (and thus minimize information loss) according to their own computational capacity. Further, it biases the selected subsets of large areas chosen for computationally intensive stages of the model towards areas that are most likely to contain meaningful features of relevance.

The second major contribution is an improved understanding of the degree to which satellite imagery contains features that are correlated with socioeconomic outcomes. Of the 52 variables tested, we show that eighteen[4] can be estimated with r^2 values greater than 0.8; to the authors knowledge, none of these 18 variables have been tested in the literature to date. Acknowledging that many of the variables with strong correlations may themselves be measuring similar phenomenon, this nonetheless suggests that the measurement of socioeconomic variables using satellite data has the potential to expand well beyond the narrow subset of topics that have been explored to date, especially when considering the limited resolution of the satellite data employed in this study (30 meters).

4.2. *Conclusion*

Our ability to understand the trajectory of human development is significantly impeded by a lack of accurate socioeconomic data in many of the most vulnerable regions of the world. Today, international providers of climate and development aid are already beginning to turn to satellite-based estimates of socioeconomic status to better inform the allocation of aid in these so-called "data deserts" Burke *et al.* (2021). In order to both (a) enable these applications, and (b) understand the limits of their capabilities, it is important for researchers to explore the types of information contained in satellite images - and, what socioeconomic variables it may be able to provide proxy information for.

In this paper, we introduce a multi-glimpse attention-based architecture for the estimation of census variables aggregated to highly variable spatial scopes ('large area estimation'). The architecture leverages a recurrent layer, in conjunction with a resNet18 convolutional layer. We tested this approach using Landsat 5 imagery and a 5-fold randomization strategy for each of 52 census variables, establishing the ability of this architecture to identify features correlated with a wide range of socioeconomic factors. We found this approach can consistently achieve accuracy (r^2) greater than 0.8 for many infrastructure variables (i.e., rooms in household, access to refrigeration), and additionally tended to perform well for variables related to child health and education. Conversely, we found weak support for the use of satellite imagery to detect features related to sociodemographic variables such as the proportion of the population which was disabled, marriage status, employment, and nationality.

Broadly, these findings illustrate that satellite based estimates of social factors have

[4]Avg. Children Surviving, % Owned Household, Avg. Years Since Birth, Average Dead Children, Average number of Children, Average Number of Children less than 5, % Phone, % Cellphone, % of households that burn trash, % of households with a washing machine, refrigerator, radio or TV, the average number of rooms, % of population that is literate, has less than primary school, primary school, % of households using wood to cook, which have trash collected, and the average years of schooling.

considerable untapped potential for a wide range of variables that had not previously been tested. While some of these are intuitive in nature - i.e., measuring the average number of rooms in a household - many are less clear in their relationship with satellite data. For example, we show that numerous metrics associated with child welfare are tightly correlated with features visible in satellite imagery. This is likely due to interrelationships between infrastructure that is visible (i.e., roads, buildings) and underlying drivers of child wellbeing, but the exact features being detected are unknown at this time. Future research directions which help shed light on the explicit features being detected for individual variables of interest could greatly aid in scaling this approach.

Declaration of Interests Statement

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available at https://figshare.com/s/3a556aefb0bf8137912d, DOI: 10.6084/m9.figshare.20482527.

Funding

This work was funded by the Department of Homeland Security Center for Accelerating Operational Efficiency, award ID 17STQAC00001-03-03.

Acknowledgements

The authors acknowledge William Mary Research Computing for providing computational resources and technical support that have contributed to the results reported within this paper. Specific thanks go to Dr. Eric Walter and Matt Kennedy for their ongoing support of our activities. URL: https://www.wm.edu/it/rc

References

Andersson Magnus, O.H. and Archila, M.F., 2019. How Data-Poor Countries Remain Data Poor: Underestimation of Human Settlements in Burkina Faso as Observed from Nighttime Light Data.

Ayush, K., *et al.*, 2021. Efficient poverty mapping from high resolution remote sensing images. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, 12–20.

Babenko, B., *et al.*, 2017a. Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. 11.

Babenko, B., *et al.*, 2017b. Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. *arXiv preprint arXiv:1711.06323*.

BenYishay, A., *et al.*, 2017. Indigenous land rights and deforestation: Evidence from the Brazilian Amazon. *Journal of Environmental Economics and Management*, 86, 29–47.

Brewer, E., *et al.*, 2021. Predicting Road Quality using High Resolution Satellite Imagery: A Transfer Learning Approach. *PLoS One*.

Brewer, E., Lv, Z., and Runfola, D., 2023. Tracking the industrial growth of modern china with high-resolution panchromatic imagery: A sequential convolutional approach. *arXiv preprint arXiv:2301.09620*.

Burke, M., *et al.*, 2021. Using satellite imagery to understand and promote sustainable development. 3. Available from: https://doi.org/10.1126/science.abe8628.

Cadamuro, G., Muhebwa, A., and Taneja, J., 2018. Assigning a Grade: Accurate Measurement of Road Quality Using Satellite Imagery. *arXiv*. Available from: http://arxiv.org/abs/1812.01699.

Castro, D.A. and Álvarez, M.A., 2023. Predicting socioeconomic indicators using transfer learning on imagery data: an application in brazil. *Geojournal*, 88 (1), 1081–1102.

Chen, D., 2017. Temporal poverty prediction using satellite imagery.

Cheng, G., Han, J., and Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. 10.

Daoud, A., *et al.*, 2021. Using satellites and artificial intelligence to measure health and material-living standards in india. *arXiv preprint arXiv:2202.00109*.

Dian, R., *et al.*, 2018. Deep Hyperspectral Image Sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (11), 5345–5355.

Fibæk, C.S., Keßler, C., and Arsanjani, J.J., 2021. A multi-sensor approach for characterising human-made structures by estimating area, volume and population based on sentinel data and deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 105, 102628.

Fibæk, C.S., *et al.*, 2022. A deep learning method for creating globally applicable population estimates from sentinel data. *Transactions in GIS*.

Goodman, S., *et al.*, 2019. GeoQuery: Integrating HPC systems and public web-based geospatial data tools. *Computers and Geosciences*, 122, 103–112.

Goodman, S., BenYishay, A., and Runfola, D., 2020. A convolutional neural network approach to predict non-permissive environments from moderate-resolution imagery. *Transactions in GIS*, 25 (2), 674–691. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12661.

Google Earth Engine, 2021. Simple Cloud Score. 5.

He, K., *et al.*, 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (9).

Head, A., *et al.*, 2017. Can human development be measured with satellite imagery? *Ictd*, 17, 16–19.

Hofer, M., *et al.*, 2020. Applying artificial intelligence on satellite imagery to compile granular poverty statistics. *Asian Development Bank Economics Working Paper Series*, (629).

Hu, W., *et al.*, 2019. Mapping Missing Population in Rural India: A Deep Learning Ap-proach with Satellite Imagery. *dl.acm.orgPaperpile*, 19, 353–359. Available from: https://dl.acm.org/doi/abs/10.1145/3306618.3314263.

Huang, L.Y., Hsiang, S.M., and Gonzalez-Navarro, M., 2021. *Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs*. National Bureau of Economic Research.

Hunter, L.M., *et al.*, 1999. People and Pixels: Linking Remote Sensing and Social Science. *Contemporary Sociology*, 28 (3).

Irvin, J., Laird, D., and Rajpurkar, P., 2017. *Using satellite imagery to predict health*. Technical report, Stanford University, Department of Computer Science.

Jarry, R., *et al.*, 2021. Assessment of cnn-based methods for poverty estimation from satellite images. *In*: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*. Springer, 550–565.

Jean, N., *et al.*, 2016a. Combining satellite imagery and machine learning to predict poverty. *Science*, 353 (6301), 790–794. Available from: http://science.sciencemag.org/.

Jean, N., *et al.*, 2016b. Combining satellite imagery and machine learning to predict poverty.

Science, 353 (6301), 790–794.
Jean, N., *et al.*, 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33, 3967–3974.
Kondmann, L. and Zhu, X.X., 2020. Measuring changes in poverty with deep learning and satellite imagery.
Krizhevsky, A., Sutskever, I., and Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60 (6), 84–90.
Kugler, T.A. and Fitch, C.A., 2018. Interoperable and accessible census and survey data from IPUMS. *Scientific Data*, 5.
Kugler, T.A., *et al.*, 2019. People and Pixels 20 years later: the current data landscape and research trends blending population and environmental data. *Population and Environment*, 41 (2).
Lecun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature*, 521 (7553), 436–444.
Leyk, S., *et al.*, 2017. Internal and International Mobility as Adaptation to Climatic Variability in Contemporary Mexico: Evidence from the Integration of Census and Satellite Data. *Wiley Online LibraryPaperpile*, 23 (6). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.2047.
Liverman, D.M. and Cuesta, R.M.R., 2008. Human interactions with the earth system: People and pixels revisited. *Earth Surface Processes and Landforms*, 33 (9).
Lv, Z., *et al.*, 2023. pyshore: A deep learning toolkit for shoreline structure mapping with high-resolution orthographic imagery and convolutional neural networks. *Computers & Geosciences*, 105296.
Marty, R., *et al.*, 2019. Assessing the causal impact of Chinese aid on vegetative land cover in Burundi and Rwanda under conditions of spatial imprecision. *Development Engineering*, 4.
Miller Runfola, D. and Napier, A., 2016. Migration, climate, and international aid: examining evidence of satellite, aid, and micro-census data. *Migration and Development*, 5 (2), 275–292.
Miller Runfola, D., *et al.*, 2017. A multi-criteria geographic information systems approach for the measurement of vulnerability to climate change. 22, 349–368.
Mnih, V., *et al.*, 2014. Recurrent models of visual attention. *In*: *Advances in Neural Information Processing Systems*. vol. 3.
Mossoux Sophie, M.K.H.S. and Canters, F., 2018. Mapping population distribution from high resolution remotely sensed imagery in a data poor setting.
Nawrotzki, R.J., *et al.*, 2016. Domestic and International Climate Migration from Rural Mexico. *Human Ecology*, 44 (6), 687–699.
Ni, Y., *et al.*, 2020. An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction. *IEEE Geoscience and Remote Sensing Letters*, 18 (9), 1545–1549.
Perez, A., *et al.*, 2019. Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. *arXiv preprint arXiv:1902.11110*.
Perez, A., *et al.*, 2017a. Poverty prediction with public landsat 7 satellite imagery and machine learning. 11.
Perez, A., *et al.*, 2017b. Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654*.
Ruggles, S., *et al.*, 2003. Ipums-international. *Historical Methods*, 36 (2), 60–65.
Ruggles, S.J., 1995. Sample Designs and Sampling Errors in the Integrated Public Use Microdata Series. *Historical Methods*, 28 (1).
Runfola, D., Stefanidis, A., and Baier, H., 2021. Using Satellite Data and Deep Learning to Estimate Educational Outcomes in Data Sparse Environments. . *Remote Sensing Letters*.
Runfola, D., Stefanidis, A., and Baier, H., 2022a. Using satellite data and deep learning to estimate educational outcomes in data-sparse environments. *Remote Sensing Letters*, 13 (1), 87–97.
Runfola, D., *et al.*, 2020a. GeoBoundaries: A global database of political administrative boundaries. *PLoS ONE*, 15 (4).

Runfola, D., *et al.*, 2022b. Deep learning fusion of satellite and social information to estimate human migratory flows. *Transactions in GIS*, 26 (6), 2495–2518.

Runfola, D., *et al.*, 2020b. Exploring the socioeconomic co-benefits of global environment facility projects in Uganda using a Quasi-experimental Geospatial Interpolation (QGI) approach. *Sustainability (Switzerland)*, 12 (8), 1–13. Available from: www.mdpi.com/journal/sustainability.

Runfola, D., *et al.*, 2017. A top-down approach to estimating spatially heterogeneous impacts of development aid on vegetative carbon sequestration. *Sustainability (Switzerland)*, 9 (3).

Russakovsky, O., *et al.*, 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115 (3), 211–252.

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. *In*: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Tan, Y., *et al.*, 2020. Combining residual neural networks and feature pyramid networks to estimate poverty using multisource remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 553–565.

Tiecke, T.G., *et al.*, 2017. *Mapping the world population one building at a time*. Available from: https://arxiv.org/abs/1712.05839.

Tingzon, I., *et al.*, 2019. Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

USGS, 2019. *Landsat Levels of Processing*. Available from: https://www.usgs.gov/land-resources/nli/landsat/landsat-levels-processing.

Woodcock, C.E., *et al.*, 2008. Free Access to Landsat Imagery. *Science*, 320 (5879), 1011a–1011a.

Wu, P. and Tan, Y., 2019a. Estimation of economic indicators using residual neural network resnet50. *In*: *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 206–209.

Wu, P. and Tan, Y., 2019b. Estimation of poverty based on remote sensing image and convolutional neural network. *Advances in Remote Sensing*, 8 (4), 89–98.

Wu, S., *et al.*, 2017. A new approach to compute CNNs for extremely large images. *In*: *International Conference on Information and Knowledge Management, Proceedings*. vol. Part F131841.

Xia, G.S., *et al.*, 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (7), 3965–3981.

Xie, J., *et al.*, 2019. Scale-free convolutional neural network for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57 (9).

Xie, M., *et al.*, 2016a. Transfer learning from deep features for remote sensing and poverty mapping. *In*: *30th AAAI Conference on Artificial Intelligence, AAAI 2016*. AAAI press, 3929–3935.

Xie, M., *et al.*, 2016b. Transfer learning from deep features for remote sensing and poverty mapping. *In*: *Proceedings of the AAAI conference on artificial intelligence*. vol. 30.

Xu, H., *et al.*, 2021. Few-Shot Object Detection via Sample Processing. *IEEE Access*, 9.

Yao, X., *et al.*, 2016. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (6).

Yeh, C., *et al.*, 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11 (1), 2583.

Zhao, X., *et al.*, 2019. Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, 11 (4), 375.

Zhong, P., *et al.*, 2017. Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (6).

Zhou, B., *et al.*, 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (6).

Appendix A. Supporting Tables

Relationship between absolute error and socioeconomic status			
Name	**AE r^2**	**Name**	**AE r^2**
Avg. Child. Survive	.00339	% Elec.	.0004
Avg. Inc.	.00022	% No Piped Water	.00029
Earned Inc.	.00084	% Sewage System	.00046
% Rural	.00062	% Elec. Cooking	.00191
% Owned Household	.00001	% Phone	.00113
% Married w/Child	.00021	% Cellphone	.00103
% Single Parent	.00137	% Internet	.00028
% Extended Family	.00037	% Burn Trash	.00131
Avg. # Families	.0004	% Auto	.00014
% Single	.00009	% Hotwater	.00006
% Married	.00025	% Computer	.00039
Avg. Children Born	.0017	% Washer	.00001
Tot. Pop.	.00071	% Refrig.	.00001
Avg. Yrs Since Birth	.00058	% TV	.00029
Avg. Dead Children	.00269	% Radio	.0006
% Foreign Born	.00238	Avg. Rooms	.00005
% Employed	.0015	% No Toilet	.00008
% Disabled	.00037	% Any School	.00001
% No Health Cov.	.00006	% Literate	.00002
Avg. Age	.00037	% Less than Primary	.00278
Int. Migrants	.00192	% Primary	.0000
% No Meal Last 30	.00045	% Secondary	.00197
Avg. Family Size	.00146	% University	.00122
Avg. # Children	.00255	% Wood Cooking	.00017
Avg. # Children <5	.00208	% Trash Collection	.00006
Avg. Age of Death	.00066	Avg. Years of School	.02246

Table A1. Strength of fit (measured by r^2) between the absolute error (estimate - observed) for each variable and the observed education attainment. These values provide a test of if there are correlations between observed errors in our model and socioeconomic status, using educaiton as a proxy variable. No strong correlations were found across the tested variables.

Relationship between absolute error and observed values			
Name	**AE r^2**	**Name**	**AE r^2**
Avg. Child. Survive	0.03182	% Elec.	0.43876
Avg. Inc.	0.03057	% No Piped Water	0.9224
Earned Inc.	0.22593	% Sewage System	0.02845
% Rural	0.03314	% Elec. Cooking	0.61326
% Owned Household	0.06917	% Phone	0.03513
% Married w/Child	0.0605	% Cellphone	0.00007
% Single Parent	0.01547	% Internet	0.30256
% Extended Family	0	% Burn Trash	0.08431
Avg. # Families	0.62212	% Auto	0.01642
% Single	0.0002	% Hotwater	0.0684
% Married	0.00604	% Computer	0.23987
Avg. Children Born	0.12343	% Washer	0.00259
Tot. Pop.	0.55597	% Refrig.	0.06109
Avg. Yrs Since Birth	0.03739	% TV	0.14222
Avg. Dead Children	0.12416	% Radio	0.10442
% Foreign Born	0.64598	Avg. Rooms	0.00243
% Employed	0.0182	% No Toilet	0.34869
% Disabled	0.65407	% Any School	0.04586
% No Health Cov.	0.03223	% Literate	0.12165
Avg. Age	0.12399	% Less than Primary	0.04845
Int. Migrants	0.54389	% Primary	0.01585
% No Meal Last 30	0.19925	% Secondary	0.36044
Avg. Family Size	0.07325	% University	0.52603
Avg. # Children	0.01597	% Wood Cooking	0.01367
Avg. # Children <5	0.01643	% Trash Collection	0.01069
Avg. Age of Death	0.33677	Avg. Years of School	0.02246

Table A2. Strength of fit (measured by r^2) between the absolute error (estimate - observed) and the observed value for each municipality, by variable. Findings suggest weak correlations (i.e., heteroskedasticity) for the majority of variables, with the % Disabled, total population, international migrants, and % piped water being notable exceptions.