Predicting Livelihood Indicators from Community-Generated Street-Level Imagery

Jihyeon Lee, 1 Dylan Grosz, 1 Burak Uzkent, 1 Sicheng Zeng, 1 Marshall Burke, 2 David Lobell, 2 Stefano Ermon 1

1Department of Computer Science, Stanford University
2Department of Earth Science, Stanford University
{jihyeon,dgrosz,buzkent}@cs.stanford.edu

Abstract

Major decisions from governments and other large organizations rely on measurements of the populace’s well-being, but making such measurements at a broad scale is expensive and thus infrequent in much of the developing world. We propose an inexpensive, scalable, and interpretable approach to predict key livelihood indicators from public crowd-sourced street-level imagery. Such imagery can be cheaply collected and more frequently updated compared to traditional surveying methods, while containing plausibly relevant information for a range of livelihood indicators. We propose two approaches to learn from the street-level imagery: (1) a method that creates multi-household cluster representations by detecting informative objects and (2) a graph-based approach that captures the relationships between images. By visualizing what features are important to a model and how they are used, we can help end-user organizations understand the models and offer an alternate approach for index estimation that uses cheaply obtained roadway features. By comparing our results against ground data collected in nationally-representative household surveys, we demonstrate the performance of our approach in accurately predicting indicators of poverty, population, and health and its scalability by testing in two different countries, India and Kenya. Our code is available at https://github.com/sustainlab-group/mapillarygcn.

Introduction

In 2015, all member states of the United Nations adopted 17 Sustainable Development Goals, including eliminating poverty, achieving good health, and stimulating economic growth by 2030 (United Nations 2015). To evaluate countries’ progress toward these goals, national governments and international organizations conduct nationally-representative household surveys that collect information on a range of livelihood indicators from households distributed throughout a given country. These surveys, such as the Demographic and Health Surveys (DHS) Program, provide critical insight into local economic and health conditions (DHS Program 2019). However, they are costly and time-consuming, particularly when surveying remote populations linked by poor infrastructure. As a result, surveys are conducted infrequently or may only capture an extremely small proportion of households (Yeh et al. 2020). Satellite images and machine learning have been proposed as an alternative (Ayush et al. 2020b,a; Uzkent, Yeh, and Ermon 2020; Uzkent et al. 2019; Jean et al. 2016), but high-resolution imagery is expensive and obscures details of the ground level.

We introduce a scalable, interpretable approach that uses street-level imagery for livelihood prediction at a local level. We utilize Mapillary, a global citizen-driven street-level imagery database (Neuhold et al. 2017). Although Mapillary cannot match the consistent quality of commercial imagery, its widespread and growing availability in developing regions make it an appealing candidate as a passively collected data source for predicting livelihood indicators. In eight months, the Mapillary dataset doubled in size from 500 million images to 1 billion, with users capturing and verifying imagery from mobile devices (Solem 2019; map 2019).

We show how to capture information from Mapillary imagery to accurately predict livelihood indicators in India and Kenya, some of the most populous and economically diverse countries in the world. We present two complementary approaches: (1) The first creates representations for multi-household clusters by segmenting street-level images and aggregating informative objects, trains models, and interprets them using the most predictive features. (2) While the strength of the first approach is interpretability, we also propose a second to learn the relationships between images and leverage the inherent spatial structure of clusters by representing them as graphs, where each image is a feature-rich node connected by edges based on spatial distance.

Our approaches predict three indicators — poverty, population, and women’s body mass index (BMI, a key nutritional indicator) — in villages and urban neighborhoods. They achieve high classification accuracy and strong $R^2$ scores for regression in India and Kenya. Our method is a cheap, scalable, and effective alternative to traditional surveying to measure the well-being of developing regions.

Related Works

Recent research has explored the usage of passively collected data sources as cheaper alternatives to door-to-door or paper forms of data collection, to augment or eventually replace expensive household surveys. Proposed

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Remote Sensing Data

Remote-sensing imagery from satellite or aircraft has been used to predict road quality in Kenya (Cadamuro, Muhebwa, and Taneja 2019), land use patterns in European cities (Albert, Kaur, and Gonzalez 2017), and economic outcomes in Africa and India (Jeant et al. 2016; Yeh et al. 2020; Pandey, Agarwal, and Krishnan 2018; Ayush et al. 2020a). However, these approaches face challenges, such as poor generalizability to other locations or indicators (Head et al. 2017; Bruederle and Hodler 2018) and lack of nuance as local scenes are not visible from space. Street-level imagery provides greater detail (e.g. people) and local information.

Street-level Imagery

Past studies have used street-level imagery to measure social or health outcomes, but the focus has been in urban areas, predicting perceived safety of American cities with Google StreetView (Fu, Chen, and Lu 2018; Naik et al. 2014), analyzing urban perception from Baidu Maps or Tencent Maps in Chinese cities (Gong et al. 2019; Yao et al. 2019), identifying urban areas from human ratings of StreetView images in India (Galdo, Li, and Rama 2020), or predicting crime (Andersson et al. 2017) and voting patterns (Gebru et al. 2017) in U.S. cities. Our goal is to create approaches for developing regions that lack infrastructure by leveraging a community-generated data source, with an additional challenge that crowdsourced data can be noisy and inconsistent.

Another study that utilizes street-level imagery is (Suel et al. 2019), which trains a CNN to predict the best and worst-off deciles for outcomes like crime and self-reported health (it performs poorly for predicting objective health). The authors trained the CNN on Google StreetView imagery from London and demonstrated that learning could transfer to three other cities of the UK (Suel et al. 2019). We try the same method in our experiments, training a CNN on Mapillary images. However, given the noisy quality of the dataset and more difficult task of predicting accurately throughout developing countries and not just within developed cities, we require more sophisticated approaches to make improvements. For example, our graph convolutional network learns from multiple images, representing the spatial structure between them in edges, and encodes useful semantic features in the nodes. Furthermore, the visual makeup of cities in the same country can be similar, and it is a harder task for models to generalize to other countries. By performing experiments in India and Kenya, which have different data distributions, we show our approach can scale across countries.

Lastly, there are works that use crowdsourced imagery, i.e. Flickr, to predict the ambiance of London neighborhoods (Redi et al. 2018) and “geo-attributes” such as GDP for gridded cells across the world (Lee, Zhang, and Crandall 2015). The latter is the most similar work since authors use a crowdsourced dataset to make predictions at a global, not solely urban, level. However, only 5% of Flickr’s imagery is geotagged (Hauff 2013). We found Mapillary is a more apt dataset for our work given its coverage of developing areas, exponential growth (Solem 2019), and geotagged imagery.

Datasets

First, we define the general problem of making predictions on geospatially located clusters. Assume there are \( N \) geographic clusters. A cluster \( i \) represents a circular area with center \( c_i = (lat_i, lon_i) \). There is a set of street-level images that fall within its spatial boundaries, \( X_i = \{ x_i^0, x_i^1, \ldots, x_i^{n_i} \} \), where each \( x_i^j \in \mathcal{X} \) is an image and \( n_i \) is the total number of images. Each cluster also has \( K \) surveyed variables of livelihood indicators, represented by \( y_i \). We aim to learn a mapping: \( \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{Y} \) to predict \( y_i \) given \( X_i \), where \( \mathcal{P}(\mathcal{X}) \) is the powerset of \( \mathcal{X} \) and \( X_i \in \mathcal{P}(\mathcal{X}) \). The regression task entails predicting the indicator directly. Classification involves predicting a binary label, where an indicator value \( \geq \) to the median results in a label of 1 and 0 otherwise. We perform these tasks over image sets \( X_i \in \mathcal{P}(\mathcal{X}) \) as clusters have a variable number of images.

We specialize the general problem to a dataset of street-level imagery and cluster-labels of indicators in India and Kenya. Each cluster represents a group of households within a 5km-radius area with label \( y_i \), which contains the index value and class label for \( K \) indicators. In India, \( K = 3 \), i.e. poverty, population, and women’s BMI. In Kenya, \( K = 2 \) as women’s BMI was not available. We discuss the data sources of \( \mathcal{X} \) and \( \mathcal{Y} \) in the following section.

Mapillary for Street-Level Images

The Mapillary API provides access to geo-tagged images and metadata (Mapillary 2019). The data is community-driven, and anyone can create an account and upload images with EXIF embedded GPS coordinates or identify the location on a map. We matched images to household clusters, and only clusters with \( n_i > 0 \) were kept. This resulted in 7,117 clusters and 1,121,444 images from India and 1,071 clusters and 156,756 images from Kenya. 98% of Mapillary images were available in high-resolution (2048 \( \times \) 1536 px) and the remaining 2% in lower resolution (640 \( \times \) 420 px).

Livelihood Indicators

We predict three varied livelihood indicators: poverty and population in both countries and a health-related measure in India. Each index is naturally continuous. The values are rescaled to be between -1 and 1 and used directly for regression. We split by the median to produce class labels. Figure 1 shows the overlap of Mapillary images and indicator labels, and Figure 2 shows the index distribution by country.

Poverty

We obtained wealth index values from the most recently completed surveys of the Demographic and Health Survey (DHS), 2015-16 for India and 2014 for Kenya. DHS data is clustered; households within a 5km-radius contribute datapoints individually but share the same geographic coordinates to preserve privacy. The index is calculated from assets and characteristics, such as vehicles, home construction material, etc. We consider poverty as the inverse of wealth.
Figure 1: **Left:** Overlap of Mapillary images with household clusters for India and Kenya. Each map also contains a close-up view of a cluster $i$, where yellow dots represent its images. **Right:** Distribution of number of images per cluster in either country.

Figure 2: **Left to right:** Distributions of wealth (poverty is defined as its inverse) and population over India and Kenya. Distribution of women’s BMI in India. Note that binary labels are generated from a median split, so classes are balanced.

**Population** Facebook’s High Resolution Population Density Maps consist of geo-located population density labels across the world. Their data is much denser than that of Mapillary Vistas, so we average the values within a 5km radius of a cluster’s coordinate to produce its label.

**Women’s BMI** Women’s body-mass-index (BMI) is an important indicator of human well-being. We compute BMI by dividing weight in kilograms by height in meters squared for the 697,486 samples in the DHS survey and average the values across all women in a cluster to produce its label.

**Methods**

Given this dataset constructed with geotagged Mapillary images and indexes, we propose methods to learn the mapping $P(\mathcal{X}) \mapsto \mathcal{Y}$. We focus on two learning paradigms: (1) image-wise learning where a model learns from a single image, $x_j \in \mathcal{X}_i$, sampled from cluster $c_i$, and (2) cluster-wise learning where a model learns from all images $\mathcal{X}_i$ in $c_i$.

**Image-wise Learning**

As did (Suel et al. 2019), we directly map each individual image in the cluster, $x_j \in \mathcal{X}_i$, to the label space. We refer to the model that learns this mapping as ResNet-ImageWise. As in Figure 3, image-wise predictions, $p_i$, are combined at test time using an aggregation strategy to produce final predictions, $y_i$. Each prediction is considered a vote, and the majority class is considered the final prediction for cluster $i$.

**Cluster-wise Learning**

**Learning from Cluster Level Object Counts** In this section, we propose a method to utilize object counts from street-level images in a cluster. Image level object counts are aggregated across the cluster to create cluster-wise object counts. Finally, we train a classifier or regression model on cluster-level object counts to predict indexes.

**Panoptic Segmentation on Mapillary Images** With the street-level imagery in Mapillary and panoptic segmentation dataset called Mapillary Vistas (Neuhold et al. 2017), we can train a network to segment street-level images. Mapillary Vistas has 28 **stuff** and 37 **thing** classes, where **stuff** refers to amorphous regions such as ”nature” and **things** are enumerable objects with defined shapes, like ”car.” It contains 25,000 annotated images with an average resolution of $\sim$9 megapixels captured at varying conditions, times, and viewpoints, i.e. from the windshield of a car, walking down a road, etc. These factors make Mapillary Vistas an ideal dataset to train a model to segment objects.

We use the seamless scene segmentation model proposed by (Porzi et al. 2019). The model consists of two main modules—instance segmentation and semantic segmentation—and the third module fuses predictions from both to generate panoptic segmentation masks. The instance segmentation module uses Mask-RCNN (He et al. 2017), and the semantic segmentation module uses an encoder-decoder architecture similar to the Feature Pyramid Net-
Figure 3: Overview of the proposed methods. Top: image-wise learning, which learns mapping $f_\tau$ from imagery as in (Suel et al. 2019). Middle: cluster-wise learning, which uses a panoptic segmentation model to create cluster-level representations and learns the mapping $f_t$. Bottom: GCN that represents clusters as a fully connected set of images. Some connections are left out in the above figure for the sake of clarity. Using Graph-Conv layers followed by a Graph-Pool layer, it learns mapping $f_g$.

work (Lin et al. 2017). Finally, ResNet50 is used to parameterize the backbone model. During training, the Mapillary images are scaled such that the shortest size is $1024 \times t$ pixels, where $t$ is randomly sampled from the interval $[0.5, 2.0]$. The authors report 50.4% mean IoU (intersection over union) score on the Mapillary Vistas test set. To be consistent with the trained model, we scale our Mapillary images such that the shortest size is represented by 1024 px.

Cluster Level Object Counts Using the seamless scene segmentation model, we segment every image $x^i_j \in X_i$ in a cluster with the hypothesis that the 65 different roadway features provide useful information. The authors of (Ayush et al. 2020b) correlated object detections from satellite imagery with poverty in Uganda and used them as interpretable features. We expected to discover patterns as well, such as more bike racks appearing in high-wealth areas. Each image $x^i_j \in X_i$ maps to a set of object detections $o^i_j$, where $o^i_j \in \mathbb{R}^{65}$. We then sum the number of instances for each class, or $\sum_{j=0}^{n_i} o^i_j$. To avoid bias towards clusters with many images, we append a feature with the total number of images in a cluster, or $n_i$. Each cluster is represented by a feature vector $z_i \in \mathbb{R}^{66}$. We map these interpretable embeddings to the label space and refer to the models as Obj-ClusterWise.

Graph Convolutional Networks

The methods thus far process images independently. We propose Graph Convolutional Networks (GCN) (Such et al. 2017) to learn relationships between images, representing a cluster as a graph, where image-based features serve as nodes connected by edges encoding their spatial distance. Each graph has a matrix $\mathcal{V}$ for nodes and $\mathcal{A}$ for edges. Our task is to learn the mapping: $f_g: (\mathcal{V}_i, \mathcal{A}_i) \rightarrow y_i$. We refer to these models as GCN. Since we model image connections with scalars, the GCN uses a convex combination of the adjacency and identity matrix to create a filter $\mathcal{H}$ that convolves $\mathcal{V}$ before passing the output through a ReLU activation and Dropout (Graph-Conv). Graph Embed Pooling (Graph-Pool) is the corollary for MaxPooling and treats the Graph-Conv output as an embedding matrix, reducing $\mathcal{V}$ and $\mathcal{A}$ to a desired number of vertices.

Node Representations Each image in a cluster is represented by a node, which is composed of CNN features from ResNet-ImageWise, detected object counts ($o^i_j$ in Obj-ClusterWise), or the combination of the two. The node representations for any cluster $i$ is $V_i \in \mathbb{R}^{n_i \times d}$, where $n_i$ is the maximum number of images per cluster and $d$ is the size of feature vector for each image.
Modeling Connections Between Nodes There is an inherent spatial structure to clusters as Mapillary images uploaded by users are geo-tagged and captured while driving or walking on roads. We take advantage of this structure by connecting the image nodes. We initialize $A$ as the normalized inverse distance between two images in a cluster. That is, let $d_{ij}^k = \text{distance}(x_i^k, x_j^k)$ be the spatial distance in kilometer unit between two images $x_i^k$ and $x_j^k$ in cluster $i$. Let the maximum distance between any two images in any one cluster is $d_{\text{max}}$. In this case, for $a_{ij}^k \in A$, $a_{ij}^k = 1 - \frac{d_{ij}^k}{d_{\text{max}}}$. This way, we construct matrix $A$ using a scalar for each edge and we get: $A_i \in \mathbb{R}^{n_i}$ for any cluster $i$.

Experiments
We perform extensive experiments on our dataset consisting of Mapillary images and ground-truth indexes. As our work is the first to utilize Mapillary for indicator prediction, we build baselines to benchmark our model performance. The metrics are classification accuracy and the square of the Pearson correlation coefficient (Pearson’s $r^2$) for regression. For each country, we randomly sample 80% of the clusters as the training set and the remainder is the validation set.

Baselines
For each cluster, we predict a local (geographic) average of neighboring clusters as a baseline. This simulates how often we have access to district-level statistics about livelihood indicators. We approximate the district-level values as the mean of the 1000 clusters closest to $c_i$. In other words, the baseline predicts the $\hat{y}$ from $Y_{\text{sort}} \subseteq Y$, where $Y_{\text{sort}}$ represents a subset of clusters sorted by distance to $c_i$ in increasing order where $|Y_{\text{sort}}| = 1000$.

Implementation Details
For ResNet-ImageWise, all images $X$ are resized to $224 \times 224$. We train a ResNet34 (He et al. 2015) model initialized with weights pretrained on ImageNet (Russakovsky et al. 2015) to learn a mapping $f_r$ from image to $y_i^r \in \mathbb{R}^K$. There is one classification and one regression label for each indicator ($K = 6$). We train with batch size 128 and learning rate 0.001 (after trying 0.1, 0.01, and 0.0001) for 50 epochs on a NVIDIA 1080TI GPU with 40G of memory.

We train different models as part of Obj-ClusterWise, which map the object counts feature to the label space. For the classification task, we use a 3-layer Multi-Layer Perceptron (layer size 256, ReLU activation, learning rate of 0.001), Random Forest (300 trees), Gradient Boosted Decision Trees (300 boosting stages), and k-Nearest Neighbors ($k = 3$). The same models are used for regression.

For the GCN, we feed the representation of image nodes via $Y$ and node connections via $A$ into two Graph-Conv layers of size 64 (each followed by a ReLU activation) followed by a Graph-Pool of size 32. These operations are followed by another pair of Graph-Conv layers of size 32 and a Graph-Pool of size 8. The resulting output is then fed into a fully connected dense layer of size 256 and a final output layer of size 2 for binary classification or size 1 for regression. We trained with a batch size of 256 and learning rate of 0.0001 on a NVIDIA 1080TI GPU. We use Adam optimizer (Kingma and Ba 2014) for all the experiments in this study.

Pooled Models for Transferability
We also train models that pool data from both India and Kenya to assess whether learning transfers. We train a Obj-ClusterWise model and a GCN model in two experimental settings. In the first, we randomly sample 50% of clusters from the training datasets of both countries, train the model, and evaluate on India and Kenya separately (50% Pooled). In the second setting, the models learn from all training clusters from both countries and evaluate on India and Kenya separately (100% Pooled).

Predictions on Livelihood Indicators
Poverty
ResNet-ImageWise, which is the baseline method from (Suel et al. 2019), achieves 74.34% accuracy in India and 73.71% in Kenya. Models that use cluster-wise representations perform better than the baselines. Obj-ClusterWise leverages the semantic information of object counts, obtaining 75.77% in India and 77.34% in Kenya. The GCN makes further improvements, achieving 81.06% in India, learning from both the visual features and spatial relationships between images. The 100% Pooled GCN achieves the best accuracy in India. As in Figure 2, Kenya is more left-skewed than India in its wealth distribution, providing additional examples of low-wealth clusters. For regression, wealth does not persist over large areas and can shift between clusters, so the baseline does especially poorly. Obj-ClusterWise methods achieved $r^2$ scores of 0.52 to 0.54 in India and the best score of 0.50 in Kenya, suggesting the semantic information encoded by object counts is useful.

Women’s BMI
BMI is an example of a health-related indicator that may not be obvious from imagery, but our models still learn effectively and improve over the baselines. BMI was not available in the 2014 DHS survey for Kenya, so we report results in India. The GCN achieved 89.56% classification accuracy and 0.57 $r^2$. We note that neither high BMI nor low BMI is desirable, so we consider for future experiments changing the index to difference from a healthy BMI, which is between 18.5 to 24.9 (DHS Program 2019).

We observe the method from (Suel et al. 2019) is not adequate and make significant improvements on most tasks by leveraging cluster-wise representations. The GCN, which is designed to learn relationships between images through feature-rich nodes and spatial distance-based edges, attains the best performance on most tasks. Obj-ClusterWise models perform comparably, relying only on object counts, sug-
Table 1: Results on predicting wealth, population, and BMI. The method in (Suel et al. 2019) serves as an additional baseline. To the best of our knowledge, we are the first to present a scalable, interpretable pipeline to make predictions from only crowdsourced street-level imagery across India and Kenya.

Analysis: Effect of Number of Images

We analyzed how many images are necessary for a cluster to be sufficiently representative. We took random samples of images, from 50 to 200 (the maximum size), from each cluster. We then trained an MLP model for 100 epochs with a learning rate of 1e-3 and evaluated it on classification in each country. As shown in Figure 4, more images led to increased accuracy, but surprisingly a relatively small number of images, even 50, was enough to achieve good accuracy.

Interpretability using Object Counts

Interpretability of the predictions is an important aspect to consider for estimates from machine learning models to be adopted by policymakers and practitioners (Ayush et al. 2020b; Murdoch et al. 2019). Currently, the DHS constructs the wealth index by conducting principal component analysis (PCA) on hand-collected characteristics of each household, including assets (i.e., televisions, bicycles, etc.), materials for housing construction, and types of water access (DHS 2016). We demonstrate how models can learn from features visible from the road, offering a much cheaper but still effective method of index estimation. Moreover, many existing models trained from passively collected data sources (Jean et al. 2016; Cadamuro, Muhebwa, and Taneja 2019; Sheehan et al. 2019) are accurate but not inherently interpretable. We examine which objects are important for a given index and visualize decision trees to help end-user organizations understand the model. Although the GCN model performs strongly, CNN features are difficult to interpret compared to semantic objects. The 100% Pooled Random Forest model performs close to the best with only object counts. Thus, we choose this model for interpretation.

Feature Importance As shown in Figure 5, evidence of development and growth, such as vehicles, traffic lights, street lights, and construction (barrier walls), were important...
in poverty prediction. Instances of terrain (e.g. dirt or exposed ground alongside a road) were also informative, most likely because they suggest a lack of urbanization. For population, infrastructure was salient, i.e. rail tracks and bridges, along with transportation modes, i.e. bicyclist, motorcycle, and truck. For women’s BMI, the most important features were streetlights, manholes, and billboards, which indicate the existence of services and development.

Visualizing Decision Trees We visualize decision trees to help end-user organizations understand how the model makes decisions. We train a decision tree regressor with a max depth of 3 on poverty prediction and then visualize it using the method of (Parr and Grover 2020) (Figure 6). At each node, moving to the left means the cluster has less than the threshold number of that object (right means greater than or equal to the threshold). Final predictions are at the leaves, where \( n \) is the number of clusters with that prediction. We demonstrate how to use the trees by tracing three paths.

Path 1: As expected, a cluster with few street lights will be predicted as low wealth, indicated by how most leaves have negative index values. Billboards are an informative feature, as few instances in the cluster leads to a negative estimate while more means positive. This is intuitive as storefronts and advertisements signal economic activity.

Path 2: Few buildings and barrier walls leads to a low-wealth prediction. These construction-type barrier walls are salient (used in the second level of the tree), indicating sites of development and growth for organizations to monitor.

Path 3: Alternatively, many buildings but few barrier walls can be found in cities of varying levels of development. Thus, street lights become an important factor, as many signal urbanization and lead to a positive wealth estimate.

Conclusion

In this work, we present a novel approach to make predictions on poverty, population, and women’s body-mass index from street-level imagery. In spite of the inconsistent quality of such a large-scale, crowd-sourced dataset, we achieve strong performance in predicting indicators that have not previously been the focus of machine learning methods, such as women’s BMI, a key nutritional indicator. We demonstrate how our method is scalable, making predictions across India and Kenya. We present two approaches: (1) cluster-wise learning, which represents household clusters from detected object counts and enables interpretability and (2) a graph-based approach that aims to capture the spatial structure of a cluster, representing images as feature-rich nodes connected by edges based on spatial distance. We hope that our method can be employed as a cheap but effective alternative to traditional surveying methods for organizations to measure the well-being of developing regions.
Statement of Broader Impact

International aid organizations and governments must be able to measure a populace’s well-being to understand where assistance is needed, but traditional ground-based survey methods used to collect data on well-being are expensive and time-consuming. Thus, several years or even decades will pass between surveys in some countries, inhibiting the ability of many organizations to understand or improve well-being. Although recent innovations in machine learning offer promising approaches to faster and more inexpensive livelihood measurement, the adoption of these new approaches by organization on the ground has been inhibited by their lack of interpretability: model-based features are typically not semantically meaningful, and practitioners on the ground are understandably hesitant to substitute these predictors for ones they understand.

This work presents the first use of new crowdsourced street-level imagery from Mapillary for the task of predicting key livelihood indicators at a local level. We show how this relatively cheap and growing data source generated by citizens can be used to train models that learn to predict a variety of indicators in spite of the imagery’s inconsistent and noisy quality. We then demonstrate how to visualize the models to understand what features are important in predictions. We show how objects detectable from the road can provide fine-grained, accurate estimates as well for a much cheaper cost and how certain objects signal development, growth, or economic activity.

While our livelihood labels derive from high-quality nationally representative data and are unlikely to be biased, there are possible biases in our crowd-sourced imagery because users upload images captured while driving, so coverage may be better alongside developed roads. Additionally, users do need a camera or a phone with camera to take the photos. However, since anyone can upload images and Mapillary supports past versions of Android and iOS, there is a diverse range of imagery uploaded from rural and developed areas across the country. Furthermore, we ensure user privacy as Mapillary blurs out faces, and our ground data contain anonymized information on dozens of households per village and thus are not individually traceable.

Our research establishes new benchmarks for performance on the task of predicting livelihood indicators from crowd-sourced street-level imagery. We demonstrate our approach’s ability to make predictions on poverty, population density, and even a health-related measure that may not be obvious or intuitive from imagery, i.e. women’s body-mass index (a key nutritional indicator). We also show our method scales to all of India, one of the most populous and economically diverse countries in the world. As Mapillary offers coverage throughout the developing world, we believe our pipeline can be generalized both geographically and to other livelihood-specific tasks, amplifying the efficacy of a range of humanitarian and sustainability organizations in need of this type of data.

Acknowledgments

This research was supported by the Stanford King Center on Global Development and NSF grants #1651365 and #1733686.

References


Galdo, V.; Li, Y.; and Rama, M. 2020. Identifying urban areas by combining human judgment and machine learning: An application to India.


Head, A.; Manguin, M.; Tran, N.; and Blumenstock, J. E. 2017. Can human development be measured with satellite imagery? In *ICTD*.


Appendix

We first show tree visualizations for each indicator in India and Kenya. Then, we visualize the output predictions of the Obj-Clusterwise model by indicator in either country.

Visualizations of Decision Trees

We train a 100% Pooled decision tree classifier and regressor with a max depth of 3 to predict poverty in India and Kenya. We then use the visualization tool presented by (Parr and Grover 2020) to visualize the tree. Each node displays the object class name and threshold that determines how to split the node (left child means $<$ the threshold, right child $\geq$ the threshold). Leaves represent predictions, where $n$ is the number of clusters assigned to that leaf.

For the classification trees, the histograms show the feature space distribution for a single feature, with the colors indicating the relationship between feature space and target class. For example, for the “car” histogram in Figure 7a, we see that the yellow bars are clustered at the lower end, which is intuitive for low wealth. The histogram gets proportionally shorter as the number of clusters that reach the node decreases, and the leaf size becomes smaller as well. (Parr and Grover 2020) motivates the use of pie charts to show quickly an indication of the strong majority category through color and size of slice.

For the regressor, feature-target space is shown with a scatterplot of feature vs. target. Regressor leaves use a strip plot to show the distribution explicitly (the number of dots is number of clusters assigned to the leaf), and the leaf prediction is the distribution center of mass, or mean, which is denoted with a dashed line.

Figure 7: Decision tree visualization for wealth prediction. Recall that we consider poverty to be the inverse of wealth. Classification on top, regression on bottom.
(a) India. Classification on top, regression on bottom.

(b) Kenya. Classification on top, regression on bottom.

Figure 8: Decision tree visualization for population density prediction. Classification on top, regression on bottom.

Figure 9: Decision tree visualization for women’s BMI prediction in India. Note that women’s BMI data was not available in the 2014 Kenya DHS survey. Classification on top, regression on bottom.
Visualizations of Predictions

Here we map our classification predictions across the country using the **Obj-Clusterwise** model. Correct predictions are green. False positives (predicted as "high" area for indicator but actually "low") are marked in red, while false negatives (predicted as "low" but actually "high") are colored purple.

Figure 10: India - Map of poverty classification predictions.

Figure 11: Kenya - Map of poverty classification predictions.

Figure 12: India - Map of population density classification predictions.

Figure 13: Kenya - Map of population density classification predictions.
Figure 14: India - Map of predictions for women’s BMI classification predictions.