OpenStreetMap Road Network Analysis for Poverty Mapping

Luda Zhao
Stanford University
ludazhao@cs.stanford.edu

Paula Kusumaputri
Stanford University
paulaksp@cs.stanford.edu

1. Introduction

Eradicating worldwide poverty by 2030 is the top goal on the United Nations sustainable development agenda, but accurate measures of poverty metrics are severely lacking in many parts of sub-Saharan Africa. Recent work by Xie et. al. have shown that highly accurate predictions of poverty indicators can be made by analyzing satellite images using convolutional neural networks and performing transfer learning to capture relevant high-level image features[3]. Another work [8] has utilized anonymized mobile phone network data to capture wealth distribution of particular regions. Since reliable poverty data in developing countries is typically scarce and sparse, we hope to use large-scale publicly available data to infer informative, socioeconomic indicators. A computational approach to extracting accurate and reliable measures from public data will lead to more informed policy-making decisions. In this paper, we hope to extends these works by incorporating another source of publicly available data – geographical and typological labels of roads and other infrastructure.

We believe there exists a strong correlation between the quality of countrys road network and a countrys economic development. A study cited in [1] found that those in poorer communities spend more travel time to reach the nearest road/transportation. Inadequate road access reduces the ability for those living in the areas to access infrastructure such as education, health facilities, transportation and participate in the market economy. As such, the poor continue to be isolated and stays below the poverty line. Previous analysis [3] has shown that CNN architecture is learning key poverty indicators using unsupervised learning, such as neurons that activates specifically on roads. Based on these studies of unstructured extraction of road information, we believe that analyzing road networks in a structured data form in developing countries might reveal good insights into the poverty problem.

2. Datasets

2.1. OpenStreetMap data

OpenStreetMap(OSM) is an open-source mapping project that provides volunteer-supplied cartographic information around the world. In addition to having the advantage of being open-source, OSM has used extensively by humanitarian teams in health outreach projects, especially in sub-Saharan Africa, which leads it to have the most complete coverage of road network conditions out of any major mapping projects, even when compared to commercial map engines such as Google maps[7]. OSM encodes streets as ways representing a collection of nodes. Each node has an ID, longitude/latitude coordinates, and tags, or the metadata associated with each node/ways. One of the main advantage of OpenStreetMaps is the availability of metadata associated with road information, including information on whether it is paved, the capacity of the road, and the category of the road (primary/secondary/tertiary). As these road typologies are influential in characterizing the overall health of the road network, we will use these tags extensively to build predictive poverty-mapping models.

2.2. Poverty Surveys

As used in [2] and [3], we use data from two surveys: consumption expenditure from the Living Standards Measurement Study (LSMS) and household asset index from the Demographic and Health Surveys (DHS) as our indicators of poverty for Uganda, Tanzania, Nigeria, Malawi, and Rwanda. LSMS is a house-

OSM statistics	Nigeria	Tanzania
Number of Nodes	6,963,532	6806013
Number of Nodes with tags	248,462	177494
Number of Ways	475,696	723999

Table 1: OSM statistics: Nigeria, Tanzania

hold survey program initiated to assist policy-makers, while the DHS collects nationally representative data on health outcomes in developing countries, and both are commonly used metrics for poverty in countries where indicators such as yearly income or household wealth datas are not systematically recorded and available.

2.3. CNN Satellite Image Features

There has been ongoing efforts by Xie et. al. to building predictive models of poverty by using neural networks to extract image features from publicly available satellite data[3]. To evaluate the country-level poverty map, we use satellite image data which were randomly sampled near the DHS/LSMS survey for the 5 countries (Uganda, Tanzania, Nigeria, Malawi, and Rwanda). These 4096-dimensional image feature vector extracted from satellite imagery by the VGG-F CNN covers a 10km by 10km area centered on a cluster. Their work have suggested that some of their image features corresponds roughly to the presences of road networks, which partly provided our intuition in using OSM data.

3. Approach

3.1. Data Preprocessing

We obtained the OSM extracts of two African countries, Nigeria and Tanzania. We chose these countries because of the relative abundance of survey data as well as their large sizes, which offsets one of the main challenges of this task – scarcity of data.

The statistics regarding the Nigeria and Tanzania subset of the OSM are shown in Table 1.

Upon filtering through all available tags, we decided on the candidates that we felt would potentially be helpful in our task – the road typology(primary/secondary/tertiary) and the physical conditions of the road(paved/unpaved). Road with different typologies are essential in measuring difference in road access, where paved/unpaved roads differ significantly in their ease of transport and reflects the general well-being of the infrastructure.

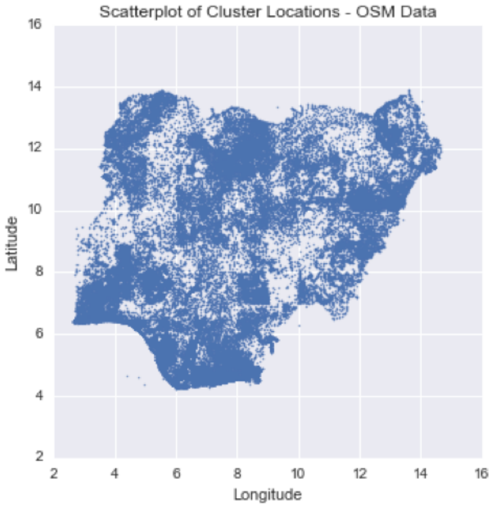


Figure 1: Distribution of OSM nodes, Nigeria

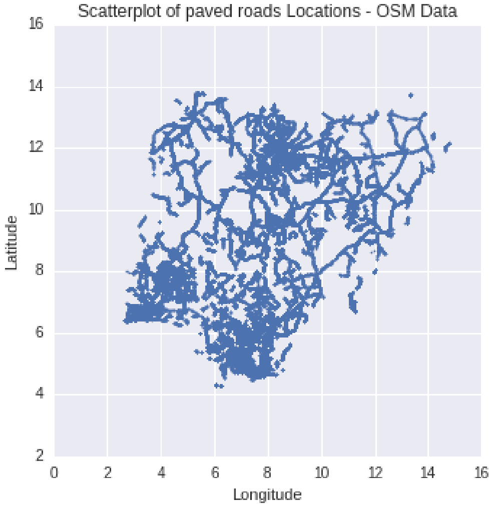


Figure 2: Distribution of OSM ways(paved roads), Nigeria

3.2. Feature Extraction

Our goals are to extract features from the OSM dataset that have the most predictive power with regarding to poverty measures such as the household asset index. Given the tags we identified previously, we come up with the following two metrics we wanted to construct, given the coordinates of a particular household cluster.

3.2.1 Closest distance to nearest road

From the coordinate location of a survey location, we computed the shortest Euclidean distance to the nearest road. We computed the following values:

- Closest Distance(km) to road, any type
- Closest Distance(km) to paved roads
- Closest Distance(km) to unpaved roads
- Closest Distance(km) to Primary roads
- Closest Distance(km) to Primary OR Secondary roads

3.2.2 Total distance of roads in proximity

We also calculated the total sum of roadage in the proximity of an area gives some indictation of the relative robustness of transport infrastructure in the area. We calculated the following metrics for $r = 1km$ and $r = 5km$ for all survey locations:

- Total distance(km) of roads within r km, any type
- Total distance(km) of primary roads within r km
- Total distance(km) of primary AND secondary roads within 1km
- Total distance(km) of paved roads within r km
- Total distance(km) of unpaved roads within r km

Both categories relied on a central routine we called ”getClosestNodeToCoordinate”, which takes a survey location as parameter and output the closest node in the OSM. However, with around 7 million nodes for each countries and thousands of locations for each of the metrics, a naive search through all possible nodes was time-prohibitive. Thus, we decided to use a KD-Tree structure to dramatically sped up computation time.

CPU runtime, closestNodeToCoord	Time(seconds)
Naive Search	9.12
KD-tree Search	0.021

Table 2: Comparison of naive search algorithm and KD-Tree search algorithm for nearestNeighbor query.

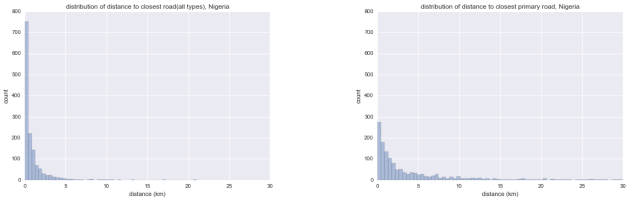
3.3. Background: KD-Tree for optimized NN-search

A kd-tree is a space partitioning binary tree that allows one to very quickly find the closest k point in a dataset to a given query point, given the dimensions of the data is small. Edges of the tree will correspond to subsets of space, and each node, v, in the tree will have two data-fields: the index of some dimension i_v, and a value m_v. Let S_v denote the subset of space corresponding to the edge going into a node v, and define $S_< = \{x : x \in S_v, x(i_v) < m_v\}$ and $S \geq \{x : x \in S_v, x(i_v) \geq m_v\}$.

Given a pair $[S, v]$, where $S = x_1, ..., x_n$ is a set of points, and v the current node(starting at root), we can insert the point into the kd-tree using a recursive algorithm:

- if $n = 1$, then store that point in the current node v. v will now be a leaf of the tree.
- Otherwise, pick a dimension $i1, ...d$. Let m be the median of the ith dimension of the points: $m = median[x_1(i), ..., x_n(i)]$. Store dimension i and median m at node v. Partition the set S into $S_<$ and $S_>$ according to whether the ith coordinate of each point exceeds m.
- Make two children of $v_<$ and $v_>$, and recurse on $[v_<, S_<]$ and $[v_>, S_>]$.

The tree will initially be balanced because we are using the medians of the coordinate values, and hence will have depth $logn$. Given a point s_{target}, if we want to find the closest point in our kd-tree structure to v, we will first go down the tree, find the leaf in which s_{target} would end up, and recurse upwards to find possible close neighbors. Given a small d, the kd-tree supports an amortized $O(logn)$ search of both nearestNeighbor search and radius search, which



(a) Distribution of distance to closest road, primary roads

(b) Distribution of distance to closest road, all roads

Figure 3: Sample distribution of extracted features

dramatically sped up our feature extraction efforts.

Since KD-tree does not work in spaces with non-Euclidean distance metrics, we converted latitude/longitude coordinates, based on a non-Euclidean distance system, to ECEF(Earth centered, Earth fixed)coordinates, which is based on a 3-D Euclidean coordinate system with the origin at the Earth's center. We then inputted these coordinates into the KD-tree, computed our queries, and converted them back to lat/long coordinates for analysis. Some sample distributions of features are shown here(Fig 3).

As a sanity check, we compared our collected data with data collected from the LHS surveys, which asked correspondent to estimate how far away they are to a major roadway system. As self-reported estimation of closest roads can have very large degrees of error, this validation was only a rough estimator of the accuracy of our metrics, and there were significant discrepancies between the two datasets. Nonetheless, the distribution of the closest road distance from OSM were similar to the survey data(see Fig 4).

3.4. Modeling

We used our features vectors to predict two key metrics providing by the DHS and LSMS datasets, respectively: asset index and average household consumption. For asset indices, we used a ridge regression linear model, where we minimized residual sum of squared with a penalty on high coefficients, which addresses the potential problem of overfitting with a small dataset.

$$\min_{w} ||Xw - y||_2^2 + \alpha ||w||_2^2$$

Where X is the matrix of feature vectors, y = the asset index, and w = coefficient vectors.

We used 10-fold cross validation with an internal linear search to tune our regularization parameters. We then used the best regularization parameter for each fold to obtain our training and testing accuracies, averaged over each trial of the 10-fold cross validation.

We also incorporated the 4096-dimension satellite image features extracted by Xie et.al.[3] with our road features in order to attempt improving upon the results obtained with only satellite image features, with the same model + training procedure used as above.

4. Results

We ran the regression models on multiple subsets of the dataset, for the two tasks of predicting average household consumption from the LSMS survey and asset index from the DHS survey.

To predict consumption expenditures, we used a log consumption prediction instead of the raw consumption values provided from the survey. This is due to the skewed distribution of raw consumption values (dollars per cap per day) as shown in Fig. 5 a, whereas the log consumption in Fig. 5 b has a more normalized distribution. With some analysis, we also found some noise in the LSMS dataset, since new single household coordinates were sometimes added. To reduce noise, we ran the regression model by filtering out any single household clusters.

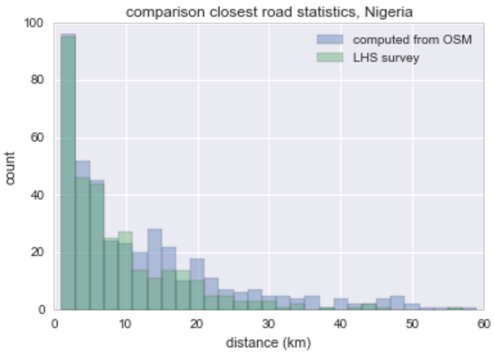
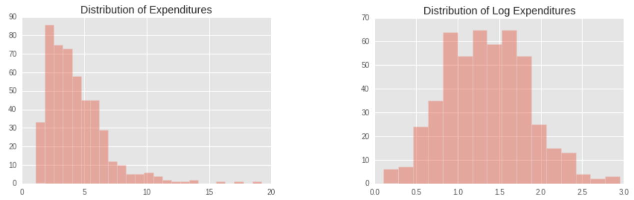


Figure 4: Comparison of distance-from-road metric from survey vs. data computed from OSM



(a) Consumption

(b) Log Consumption

Figure 5: LSMS Consumption Distribution

Prediction Task	Nigeria	Tanzania
Asset Index	0.376	0.465
Consumption	0.447	0.450

Table 3: r^2 value using only the OSM road features for the 2 tasks to predict log consumption expenditures.

4.1. Regression using Only OSM Features

First, we only used the extracted road features to predict average household consumption and asset index, measured at the cluster level for Nigeria and Tanzania, with results in Table 3.

We provided a comparison to the baseline of running the model with four columns of the LSMS survey data features: distance to nearest headquarter, market, road and population center. As shown in Fig 6, our road features shows an increase in r^2 of 0.25 for Nigeria and 0.28 for Tanzania.

4.2. Regression with Image Features Extracted

After predicting both consumption and asset index from only the OSM features, we hypothesized that we can improve upon the previous work [3] by appending the OSM road features to the previously extracted image features. We compared the model with enhanced dataset to a baseline of pre-existing 4096-dimensional image feature vectors.

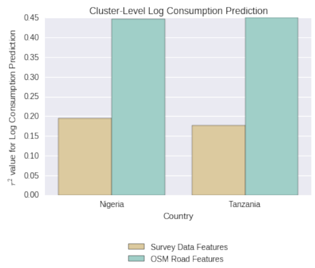


Figure 6: Consumption Prediction with Survey Features vs OSM road features

Asset Index		
Country	Image Features	Image+Road Features
Nigeria	0.672	0.689
Tanzania	0.566	0.611

Consumption		
Country	Image Features	Image+Road Features
Nigeria	0.410	0.410
Tanzania	0.535	0.536

Table 4: Results comparing baseline to image+road features to predict asset index and consumption.

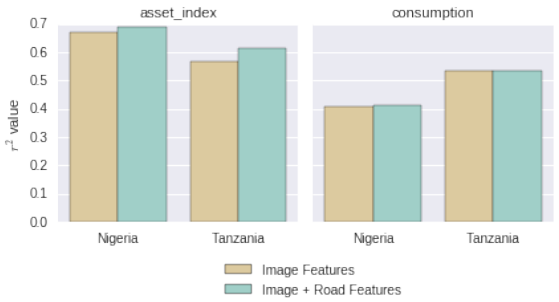


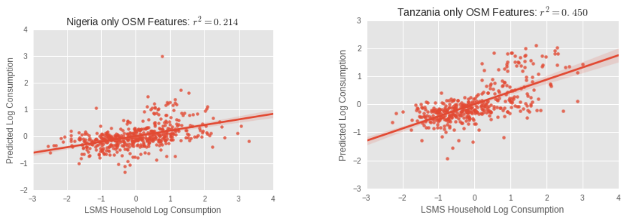
Figure 7: Comparison between the baseline image features with enhanced feature set.

The model with additional OSM features performed the asset-index prediction task with an increase in r^2 from the baseline by 0.02 for Nigeria, and an increase by 0.04 for Tanzania (see Table 4). On the other hand, running regression for consumption prediction task does not seem to show any improvement when comparing how the model performs with the enhanced vs baseline feature.

5. Analysis

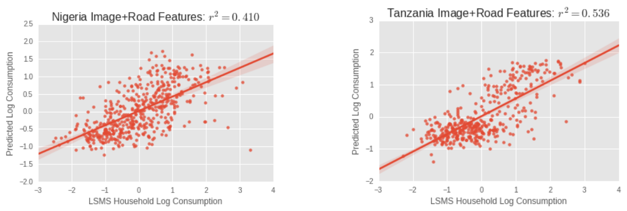
5.1. Principal Component Analysis

In order to understand the detailed relationship between the extracted features and the poverty metrics that we want to predict, we used Principle Compo-



(a) Nigeria OSM Features (b) Tanzania OSM Features

Figure 8: Consumption Prediction using Only OSM Features



(a) Nigeria Image+OSM Features (b) Tanzania Image+OSM Features

Figure 9: Regression Result for Consumption Prediction with Image+Road Features.

nent Analysis(PCA) to visualize and discover trends within our data. Using PCA on the matrix of extracted features, we found the two most significant principal components and plotted the projection of each data point onto them for Nigeria. The data points were colored from brown(low) to green(high) based on their asset index(10).

We see that the top two components seem to explain the poverty discrepancy well, with a clean clustering of poorer regions on the top left(lower x, higher y), while the richer locations tend to spread out on the lower right side(higher x, lower y). These results are significant since no other labels from the dataset was used – only the distance metrics extracted from OpenStreetMap.

By examining the two principal components itself, we came up with the following observations:

- **The first component(x-axis in the scatter plot)**: Factors that strongly correlates positively with the first component are the total distance metrics, which capture the road density surrounding the region(See Fig 11). Combined with the clustering patterns, we can infer that higher road density is correlated with a higher asset index, which is reasonable given the importance of infrastructure robustness in the economic well-being of a location.

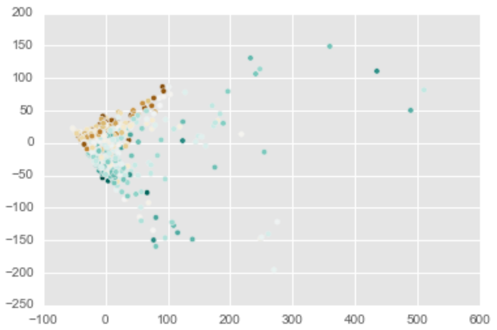


Figure 10: PCA analysis – Nigeria

- **The second component(y-axis in the scatter plot)**: The feature that strongly correlates positively the second component is the *distanceToUnpavedRoad* metric, while the feature that strongly correlates negatively with the second component is the *distanceToPavedRoad* metric(See Fig 12). Combined with from the observation that poorer regions have significantly higher y scores than richer regions, this gives the natural interpretation that the presence + distance of unpaved vs. paved roads differentiates the poorer and richer regions.

6. Conclusion + Future Work

Using data provided by OpenStreetMaps, we were able to extract features which characterize road access in remote regions across two countries in sub-Saharan Africa, which we were able to use in effectively predicting key poverty metrics collected from surveys. We demonstrated that our metrics were significantly better at predicting these poverty metrics than using road access metrics collected by surveys. In addition, we showed that our features can be combined with the satellite image features from Xie et. al. effectively to further boost prediction accuracy, which is remarkable as both dataset derives solely from publicly available data. We were able to use PCA to interpret our features and analyze the different specific features that were influential in predicting poverty measures.

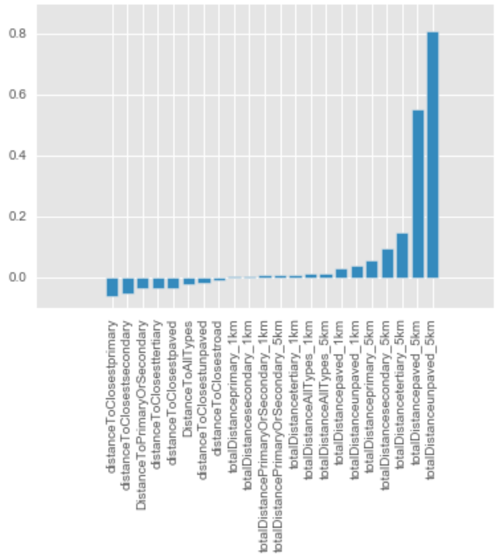


Figure 11: Correlation of each feature with 1st principal component

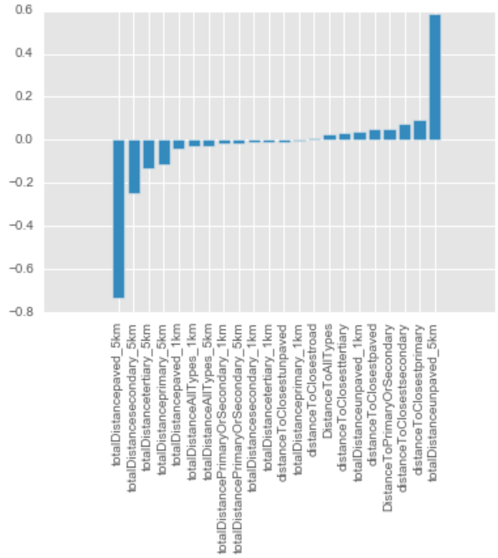


Figure 12: Correlation of each feature with 1st principal component

In the future, we hope to perform more robust analysis of road features in terms of their ability to explain variance in data. As our feature selection process was mainly guided by heuristics and intuition, we also plan to use more sophisticated feature selection methods for road access. Finally, we hope to generalize results to more African countries in order to fully present this technique as a broadly generalizable technique. This work has great potential for governments and NGOs to use on the task of poverty mapping, either directly or indirectly by complementing other techniques.

References

[1] http://www.3ieimpact.org/media/filer_public/2014/09/18/47_evaluation_of_road_access_impacts_edmonds.pdf

[2] Combining satellite imagery and machine learning to predict poverty. Science Paper. Draft

[3] Xie, Michael, et al. ”Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping.” arXiv preprint arXiv:1510.00098 (2015).

[4] https://sustainabledevelopment.un.org/content/documents/1767Poverty%20and%20sustainable%20transport.pdf

[5] ”Open Street Map Gets the Details in Africa - Development Seed.” Development Seed. N.p., n.d. Web. 03 May 2016.

[6] ”Got a Road? The Importance of a Good Road Network.” AfricaCan End Poverty. N.p., n.d. Web. 03 May 2016.

[7] Humanitarian OpenStreetMap Team. (n.d.). Retrieved June 06, 2016, from https://hotosm.org/about

[8] Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. ”Predicting poverty and wealth from mobile phone metadata.” Science 350.6264 (2015): 1073-1076.