Environmental Links to Health Outcomes in Rwanda, Tanzania, and Uganda: Midterm Project Report

Vishali Sairam
192379@mpp.hertie-school.org

Abiola Oyabanjo
abiola.oyebanjo@fu-berlin.de

Jean Pierre Salendres
192796@mpp.hertie-school.org

March 30, 2021

Abstract

In our project, we hope to apply Machine Learning models to DHS data sets and satellite imagery to classify key health outcomes. The idea for our project stems from the fact that in most remote places, carrying out a full-blown demographic and health checks would be time consuming and expensive. In order to supplement these efforts, we hope to use easily available data on the built environment as an initial tool to classify possible health outcomes. By doing so, we hope to add to the growing literature on the effects of urbanization or urban poverty on health outcomes and measure the potential change to those who are transitioning from rural to urban lifestyles. Being in an urban area enables access to health services, and improved access to water sanitation and infrastructure, but it is also associated with more pollution, sedentary and stressful lifestyles and consumption of less nutritious food (Loutan, L. 2012). By creating a continuum of urbanization across three countries in East Africa - Rwanda, Tanzania and Uganda, we want to understand how changing levels of urbanization contribute to specific health outcomes.

In the course of this project, we want to use data on environmental factors to classify health outcomes. Our questions are the following:

- How can we best classify health outcomes using data on the environment that is readily available?
- For each health outcome, which method of classification works best and why?
- How can we use ML to conceptualize better the idea of a built environment?

We use two methods of measuring the built environment and test its effect on three binary health outcomes. We also compare the efficiency of three machine learning models (Logistic Regression, Support Vector Machines and Decision Trees) in classifying these outcomes.

1 Proposed Method

We use two methods to measure the built environment:

For the first measure, we pool data on the built environment from different sources. These include data on household characteristics from the Demographic and Health Surveys, climate and geographical data also from DHS, with nighlight data from the NOAA. For the second measure (future work), we follow from Yeh et al (2020) use publicly available satellite data from Google API. They train a convolutional neural network with this using nightlight data to create estimates of household consumption and assets, both of which are hard to measure in poorer countries.

We test these two methods on three health outcomes related to access. While the first measure requires survey data, the second measure requires only publicly available satellite data, thus making it a useful tool in certain

contexts.

Models

Our baseline model is the data on two types of environment fitted to a logistic regression. Model 2 is the same dataset fitted to a Support Vector Machine and Model 3 is fitted to a Decision Tree algorithm.

Logistic Regression (Baseline)

A Logistic Regression Machine Learning algorithm can be useful for us and the classification problem, since it is a predictive analysis algorithm based on the concept of probability. We want to evaluate the probability of child, maternal and general health outcomes based on variables of the built environment around the individual. As a proxy, night-time luminosity data allows us to understand where the individual can be placed on a rural-urban continuous scale based on the level of luminosity. Moreover, with many variables from the DHS data sets, we may build the machine learning model to understand which variables are important and how to integrate them into an algorithm. The code is in the following link: `https://github.com/abiola1864/ML4Development/blob/main/midterm/codes_for_each_method/logistic_regression.ipynb`

Initially, for logistic regression, the set taken was general health. We assess whether an individual reports to be anemic or not anemic based and we use this to learn how the environment around him can be linked, as measured by variables in the DHS data set.

Support Vector Machine

The Support Vector Machine (SVM) allows us to efficiently perform a non-linear classification called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The code is in the following link: `https://github.com/abiola1864/ML4Development/blob/main/midterm/codes_for_each_method/svm.ipynb`

Decision Tree

With a decision tree, it is possible to identify a set of rules with which we classify high risk and low risk health outcomes. Decision trees, unlike regression, do not assume independence between the various predictors which is likely to be the case with the DHS data. It is also particularly useful in our case since most of the feature values that we have are categorical and not continuous.

Initially, we tested this model with the DHS Rwanda 2014-15 data set. In the next part of our project, we will extend this model to 2014-15 data sets of Tanzania and Uganda.The code is in the following link: `https://github.com/abiola1864/ML4Development/blob/main/midterm/codes_for_each_method/ch_imputed.ipynb`

2 Experiments

Data We use the DHS demographic and geographic datasets and data on nightlights to test the effect on three binary health outcomes:

- Child Health - as measured by weight/height deviation, signifying malnutrition. Numbers that are above and below two standard deviations are classified as wasting/stunting as per WHO standard.
- Maternal Health - as measured by whether the woman is anemic or not. We chose Anaemia because of its severity and importance in Sub-saharan Africa, especially pregnant women.
- General Health - as measured by whether the respondent requires money for treatment. Our initial variable of interest was coverage of health insurance, however due the number of missing observations, we decided to use this measure instead.

We chose these variables since they are specifically related to access and hence contingent in some way upon the type of environment the respondent is in. While (a) and (b) are related to access to nutrition, the second is related to access to finance. Table 1 outlines the outcome variables.

After assembling the dataset, we drop variables which are more than 75% missing (water available for washing hands, b, c). We also drop the wealth index and place of residence (urban/rural) due to the possible high correlation with other variables. Finally, in order to impute missing observations, which are present for the categorical data in the dataset, we use the MICE package using R

Outcome	Data Specifications
Stunting	N: 9395 1: 51% and 0: 49%
Anaemia	N: 22012 1: 20% and 0: 80%
Treatment Money	N: 44988 1: 52% and 0: 48%

Table 1: DHS Data and Specification

Method	Area under ROC	FPR	FNR
Child	0.54	0.05	0.64
Maternal	-	0.10	0.42
General	0.66	0.19	0.17

Table 2: Results of Decision Tree Classifier

with five iterations. We combine this with two different measures of Built Environment. We expect that by moving to a better measure of the built environment, we will get a more precise understanding of how gradual changes in the urbanization level affect specific health outcomes. The nature of maternal health dataset presents us with a "class imbalance" problem, that is, there are significantly more observations in one category than in the other. We also hope to explore this in the next part of the assignment.

Evaluation method: For the midterm report, we focus only Rwanda 2014 data.
Evaluation Parameters: We split the dataset into training, test and validation datasets with a 60, 20, 20 percent split. We first do a baseline model for each of the algorithms with the test and training model. Following this, we use a grid-search to determine the best parameters which we then test on the validation set. In the next part of the project, we hope to use a scoring rule to penalize higher false negatives in the data and we believe that this would give us better results.

Classification Accuracy / Performance Metrics

We use four evaluation conditions:

False Negative Rate: Our key evaluation condition is to decrease the presence of false negatives - for example classifying a child as healthy when he is in fact stunted.

False Positive Rate: This classifies a child as stunted when it is in fact healthy.

Area under Curve: For measuring the ability of our classifiers to discriminate between positive and negative classes.

Mean Square Error: For measuring the performance of the loss function of our distribution.

Experimental details:

1. Logistic Regression

We focus first on the general health outcome, which we measure through the binary variable anemic or not anemic. Child health and maternal health, using this method, will be added in future work. The individuals' environment as measured by dozens of variables from the DHS survey will be the input to the machine learning model. Visualization graphs of all relevant variables and their distribution in terms of our general health outcome (anemia DHS variable "v467c") can be seen in the code. We split train and test sets and used logistic regression code to train model and then predict on test set.

2. SVM The SVM Model has only been applied to general health currently. Future work will apply this model to the child health and maternal health branches of our health outcomes analysis. The data was split in two with a test size of 20 percent. We used SVM algorithm from the Sklearn to minimizes the loss between the hyperplanes and their dimensions. For this analysis, we focus on the effect of nightlife composite on anemia classification.

3. Decision Tree

The decision tree analysis was applied to the three health outcomes: general, child, and maternal heath. We train an initial classifier based on train and test data. We then use the grid search algorithm with scoring based on recall to obtain parameters to test on the validation set. The results of these are outlined in the table below. With a decision tree classifier, there is a high chance of over fitting, and we will try to introduced a points systems to address this in the next part of assignment.

3 Results

For the logistic regression model, we estimate the Jaccard Index (0.55), precision score(0.53), log loss (15.67),

Method	Training Set	Test Set
Linear	0.48	0.47
Nonlinear	0.34	0.34
Bayersian	0.59	0.59

Table 3: Mean Square ErrorResults of SVM Classifier: Nightlife Composite on Anemia Classification

classification report (see graph in code file), and confusion matrix(see matrix in code file). The results point to a model that does predict better than random chance but it is not a strong predictor for our general health outcome (anemic or not anemic). For example, we have 4337 individuals that are correctly predicted as anemic, but 3787 that are false positives. The precision for this is 0.53 and recall is 0.94, f1 score of 0.68.

For the SVM model, the preliminary result of the mean square error baseline for non-linear classification is 0.348. We adjusted the kernels via the randomized search that optimizing hyperparameters. The The goal of the optimization procedure is to find a vector that results in the best performance of the model after learning, such as maximum accuracy or minimum error. The result is better at 0.593 with parameter distributions of gamma (0.005, 0.1) and uniform(51, 10).

For the decision tree classifier, using a grid search to find maximum depth and minimum leaf samples with an 5 cross validation folds in fact increased the false negative rate. This is the the proportion of significance tests that failed to reject the null hypothesis when the null hypothesis is indeed false. This indicates that the initial models that we have were had over-fit the data. However, the false positive rates fell considerably in the child health and general health models. The area under ROC curve decreased from 0.85 to 0.66 for general health and from 0.60 to 0.54 for child health when we applied grid search algorithm. Our next aim is to introduce a scoring system which penalizes false negative results more than a false positive results, and we will explore these in the coming parts of the assignment.

4 Future work

I. Algorithm Choices and Parameter Adjustments In the first stage, we have done a preliminary evaluation of three models: logistic, SVM, and decision tree regression using data from Rwanda. These algorithms will continue to adjust (learn) their internal parameters based on the new data. However, we will also move into adjusting parameters that are not learned and have to be configured based on the specific project at hand. These "hyperparameters" can be modified through optimization strategies that we seek to explore, and may include grid search, random search, hill climbing and bayesian optimization. We need to evaluate iterations of new sets of parameters and evaluate them on the test set to find the parameters with the best metric score. For example, the tree depth in a decision tree model is a typical hyperparameter. Moreover, we want to keep observing differences in performance across child, maternal, and general health outcomes.

II. More learning: Tanzania and Uganda We will continue to assess the performance of these three models as data from Uganda and Tanzania is fed into the models. There are natural differences between countries that may impact the precise links of environment to health outcomes across the urban-rural scale (as measured by nighttime luminosity data). We want to understand changes and differences in model fit and make appropriate decisions in order to have a machine learning environment that is able to adjust to these country-particularities. Precisely, the user of the algorithm should better understand the reliability of health outcomes predictions in various countries. We will focus on night-time luminosity data as a proxy for the individual's location on the rural-urban continuous scale.

III. Incorporate Satellite Method of Daylight Environment Assessment This is a path that we want to explore. Based on the performance of our machine learning model that links local environment features (measured by DHS survey variables) to health outcomes (general, maternal, and child health), it may be possible to assess the predictive capacity of adding daylight environment observations from satellite imagery. This would involve neural network exercises.

5 Bibliography

DHS (2008). An assessment of the quality of data on Health and Nutrition in the DHS Surveys, accessed at https://dhsprogram.com/pubs/pdf/MR6/MR6.pdf

Loutan, L. (2012). Urbanization reshaping infectious diseases. International Journal of Infectious Diseases, 16, e20.

Pinchoff J, Mills CW, Balk D (2020) Urbanization and health: The effects of the built environment on chronic disease risk factors among women in Tanzania. PLoS ONE 15(11): e0241810. https://doi.org/10.1371/journal.pone.0241810

ProjectPro, How to optimize hyper parameters of a DecisionTree model using Grid Search in Python?, accessed at https://www.dezyre.com/recipes/optimize-hyper-parameters-of-decisiontree-model-using-grid-search-in-python

Yeh, Christopher et al. 2020. "Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-Being in Africa." Nature Communications 11(1): 1–11. http://dx.doi.org/10.1038/s41467-020-16185-w.

Xi Chen and William Nordhaus, "Using luminosity data as a proxy for economic statistics," Proceedings of the National Academy of Sciences (US), May 24, 2011, 108(21): 8589-8594