Machine Learning Project Proposal

Abiola Oyebanjo

Vishali Sairam

Jean Pierre Salendres

March 2, 2021

1 Introduction

Countries across Africa and Asia are experiencing unprecedented growth in urban settlements. Much of this growth belies the traditional urban-rural dichotomy, prevalent in western countries. Growth of peri-urban areas, small towns, or big villages are all examples of this non-dichotomous relationship.

In our project, we hope to use satellite imagery and other datasets to supplement widely available data from the Demographic and Health Surveys. 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 along the many levels of the rural to urban lifestyle scale. 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). The Built Environment specifically has been found to affect outcomes like chronic diseases, malaria, and depression - all of which can be linked to living in an increasingly 'urban' environment. 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.

2 Research Objectives

In line with this, our objectives for this project are:

- Analysing urbanization in Rwanda, Tanzania and Uganda by creating an 'urban continuum' measure
- Using the above, we want to understand how built environment plays a role in:
  - Maternal Mortality
  - Cardiovascular Disease

We expect that by moving to gradient-based or a continuous measure of the built environment, we will get a more precise understanding of how gradual changes in the urbanization level affect specific health outcomes. Our contribution is unique in that we would like to use night-sky luminosity as a proxy for urbanization, obtained through satellite data, to expand the binary urbanization scale in the health data. This will allow us to pin-point effects on health outcomes and further use machine learning techniques to understand the variation and interaction between our variables.

3 Literature Review

Multiple sources of literature have characterized the varied nature of urban transformation in developing countries. Djikstra et al (2018) develop a new measure of urbanization using spatial data and differential measures of population grid to show that population share in rural areas is radically different in many Asian and African countries. The population share in rural areas as defined by the degree of urbanisation is similar to the share reported based on national definitions in most countries in the Americas and Europe, but radically different in many African and Asian countries. A similar methodology is followed by Galdo, Li and Rama (2020), but in the case of India. Applications of satellite sensing have been used extensively to study economic well-being and poverty in Africa. Georganos et al (2019), employ satellite-derived VHR land-use/land-cover (LULC) datasets and couple them with the DHSWealth Index (WI) to provide a city scale wealth index of Dakar.

Night-time luminosity satellite data has been linked to studies about poverty and economic development in several studies including Chen and Nordhaus (2011) and Bruederle (2018). Lastly, we place our study in the increasingly researched literature linking environmental change and urbanization to specific health outcomes across Africa and Asia. Montgomery (2008) uses data from urban samples of 85 Demographic and Health Surveys to find that the neighborhoods of relatively poor households are more heterogeneous than is often asserted.

However, we are not aware of existing literature going beyond economic or health metrics using broad categories and diving into changes in specific health outcomes, such as maternal mortality and cardiovascular disease. More importantly, the literature provides examples of global or continental analyses rather than more focused country-studies that may exploit context-specific variation in our urbanization and health outcomes data.

4 Data and Method

Given the satellite and demographic data, we hope to primarily use random forests and logistic regressions to test effects on maternal mortality and cardiovascular disease. Random forests and linear regressions have been used extensively while studying health and demographic outcomes. Decision Trees tell us about the likelihood of observing certain health outcomes conditional on belonging to a particular sub-group of the population, and account for the fact that the relationship between a health outcome and explanatory variables may be complex and not linear.

4.1 *Health Outcomes Data (Dependent Variables)*

Machine learning techniques require that we extract and harness data large enough to detect false positives, increase statistical power and mount predictive capacities. As earlier highlighted, we intend to combine transnational data from multiple years in three East African countries (Rwanda, Tanzania and Uganda). They constitute three of the six countries in the East African Community (EAC). The main data source is the Demographic and Health Surveys (DHS) Program which collects representative national and sub-national data on population, health, HIV, and nutrition through more than 400 surveys in over 90 countries. The DHS data includes households, geographical and geo-spatial data that will be matched by their cluster (district) ID.

We are focusing on cluster-level data for a granular representation of urbanization in the DHS data from 2009 to date. There is a data collection round at least once every two to three years. To protect respondent confidentiality, DHS-GPS data is randomly displaced by their GPS latitude/longitude positions. Urban clusters are displaced up to 2 kilometers,and rural clusters are displaced up to 5 kilometers, with 1 percent of the rural clusters displaced up to 10 kilometers. GPS coordinates allow us to get continuous surface maps of ownership of health-related assets and compare their spatial distribution. The focal dependent variable is maternal mortality where we expect some heterogeneity across rural/urban residences. Measures for mortality include *neonatal mortality, post natal mortality, under-five mortality, infant mortality, child mortality, early neonatal deaths and stillbirths.*Furthermore, we consider the evident correlation between urbanization and global warming/climate change. In this regard, we intend to explore the implication of urban changes (indexed by air pollution) on the incidence of chronic cardiovascular diseases such as anemia and obesity *(overweight)*

Exploratory results from DHS data shows an average coverage of 20,000 households per survey. We expect missing data for each variable of interest - and will deal with them based on how they affect the mean or median of each outcome.

4.2 *Urbanization Satellite Data*

Available night-time light (NTL) satellite data from the Defense Meteorological Satellite Program (DMSP) - Operational Linescan System (OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS) enables the use of night-sky luminosity to monitor various types of human activity. One challenge is that DMSP data is available to the public only until 2013 and the new generation VIIRS data starts in 2012 and continues until the present.

In our project, this data can serve as a proxy for urbanization levels. However we would like to look at a longer period than each of the satellites offer. Therefore, we will use data combining both satellites' data resulting in a 1992-2018 data-set. The methodology and ready-to-use harmonized data comes from Li, Zhou, and Zhao.

The harmonized NTL time series data will allow us to create an urbanization scale for the three countries in our project: Rwanda, Tanzania, and Uganda. Our new scale will provide greater variation in the urbanization scale as compared to the binary urban-rural data coded into the DHS data used for health outcomes.

Our Machine Learning project will be able to use night-time luminosity data, and the resulting urbanization scale, to increase the urbanization-level accuracy of the DHS geo-tagged health outcomes data. Furthermore, it will allow us to create better predictive models along a range of key health outcomes in the three East African nations selected. Urbanization scales look different across countries and therefore these country-specific variations in urbanization can be more precisely analyzed and integrated into machine learning algorithms.

5 Bibliography

Bruederle, A., Hodler, R. (2018). Nighttime lights as a proxy for human development at the local level. PloS one, 13(9), e0202231. https://doi.org/10.1371/journal.pone.0202231

Dijkstra, Lewis et al. 2020. "Applying the Degree of Urbanisation to the Globe: A New Harmonised Definition Reveals a Different Picture of Global Urbanisation." Journal of Urban Economics (September): 19–21.

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

Li, X., Zhou, Y., Zhao, M. et al. A harmonized global nighttime light dataset 1992–2018. Sci Data 7, 168 (2020). https://doi.org/10.1038/s41597-020-0510-y

Montgomery, Mark. 2008. "The Urban Transformation of the Developing World." Science Advances Vol. 319,(1): 86.