Constructing a wealth index for socio-economic modelling in South Africa

Dr. Matthew Thomas

HIV Inference Group Meeting

10^{th} November 2022

Introduction

- Accurate estimates of socio-economic well-being are vital to set policies and address inequality.
  - Reduce poverty or improve public health.
  - Target deprived areas.
- Significant advancements in census data collection, challenges remain to produce reliable estimates of poverty in developing countries.
- DHS Program conduct nationally representative surveys in over 90 countries around the world.
- DHS surveys collect data on a wide range of monitoring and impact evaluation indicators in the areas of wealth.
- These questions can be extracted and used to study national and sub-national socio-economic status.

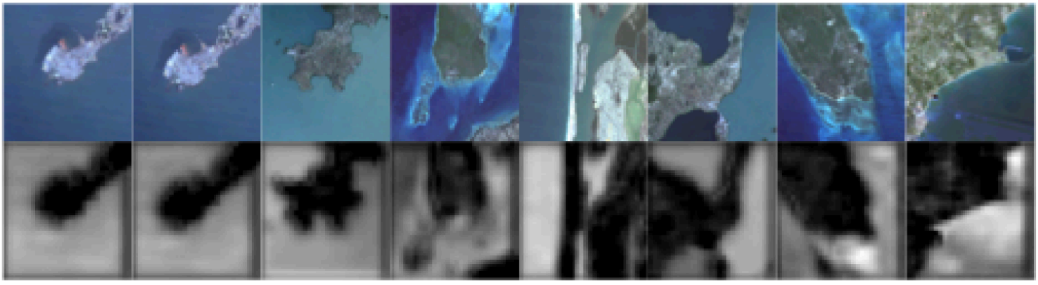
INTRODUCTION

- Yeh *et. al* (2020) used 43 DHS surveys across to construct a wealth index between 2009 and 2016 in 23 countries across sub-Saharan Africa.
- Combined with satellite images, used deep learning methods to predict socio-economic status in three year bands (2009-2011, 2012-2014, 2015-2017).

Table: Summary of wealth related questions used in Yeh et al. (2020). Questions are categorised into asset variables and non-asset variables.

Asset variables		**Non-asset variables**
Has electricity	Has motorcycle/scooter	Water quality
Has television	Has car/truck	Toilet quality
Has radio	Has mobile telephone	Floor quality
Has refrigerator	Number of bedrooms per person	

Introduction



Introduction

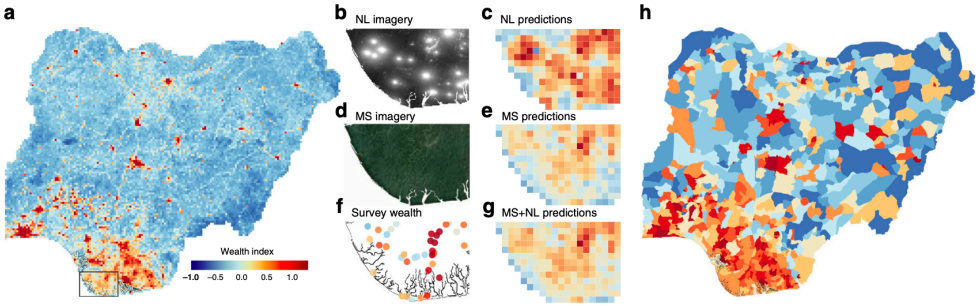


Fig. 6 Spatial extent of imagery allows wealth predictions at scale. a Satellite-based wealth estimates across Nigeria at pixel level. **b**, **d** Imagery inputs to model over region in Southern Nigeria depicted in box in **a**. **f** Ground truth input to model over the same region. **c**, **e**, **g** Model predictions with just nightlights (NL) as input, just multispectral (MS) imagery as input, and the concatenated NL and MS features as input. In this region, the model appears to rely more heavily on MS than NL inputs, ignoring light blooms from gas flares visible in **b**. **h** Deciles of satellite-based wealth index across Nigeria, population weighted using Global Human Settlement Layer population raster, and aggregated to Local Government Area level from the Database of Global Administrative Areas.

Principle Component Analysis

- Yeh *et al.* (2020) and other studies have used Principal Component Analysis (PCA) to construct a household level wealth index.
- PCA is a method of dimensionality reduction and it has many uses (Image compression, Financial modelling etc.)
- The aim of dimensionality reduction is to reduce the dimensionality of a data set to produce a more manageable reduced data set.
- Useful when attempting to analyse a large data set, with a significantly large number of variables.

Principle Component Analysis

- This new reduced data set should retain as much information in the original data set as possible and remove any redundant information.
- PCA is a linear dimensionality reduction technique

$$y_i = a_1 x_{i,1} + a_2 x_{i,2} + \ldots + a_n x_{i,n}$$

- The elements a_j are summation weights that we apply to data to
- The simplest case for choosing $a_1 = 1$ and $a_2 = \ldots = a_j = 0$

$$y = a_1 x_1$$

- This would not be a great choice because it will not contain much information about the original data set.
- There needs to be a criterion placed on the choice of a_j such that we retain as much information about the original input data $\boldsymbol{x}$.

DATA

- Survey data was obtained from the 2016 DHS survey in South Africa with fieldwork conducted between June and November 2016.
- The survey contains responses to questions regarding the household's assets, such as electricity, a refrigerator, a car and other items.
- Variables were categorised into asset variables (such as household has a washing machine, or a refrigerator), and non-asset variables (such as the source of drinking water).

Table: Summary of wealth related questions. Questions are categorised into asset variables and non-asset variables.

Asset variables		**Non-asset variables**
Has electricity	Has mobile telephone	Source of drinking water
Has television	Has car/truck	Type of toilet facility
Has radio	Has watch	Main floor material
Has refrigerator	Has animal drawn cart	Number of household members
Has motorcycle/scooter	Has boat with a motor	Number of rooms used for sleeping
Has a computer	Owns livestock	Type of cooking fuel
Electricity connected to the mains	Has microwave oven	Share toilet with other household
Vacuum cleaner or floor polisher	Has bicycle	
Has electric/gas stove	Has washing machine	
Has telephone (landline)		

DATA

- The extracted asset variables had binary responses with ”yes” recoded as 1 and ”no” recoded as 0.
- Following Yeh et al. (2020), Responses to:
  - ”Source of drinking water”
  - ”Main floor material”
  - ”Type of toilet facility”

  were recoded between 1-5, with 5 representing the highest quality and 1 being the lowest quality
- Responses to Type of cooking fuel were recoded, with non-polluting fuel recoded as 1 and polluting fuel recoded as 0 (according to the WHO definition of polluting fuels).
- An number of rooms per person was also calculated, by combining the variable number of rooms used for sleeping and number of people per household.
- Households with missing responses were removed, leaving a total of 11083 households in our data set to estimate an asset index.

PCA Analysis

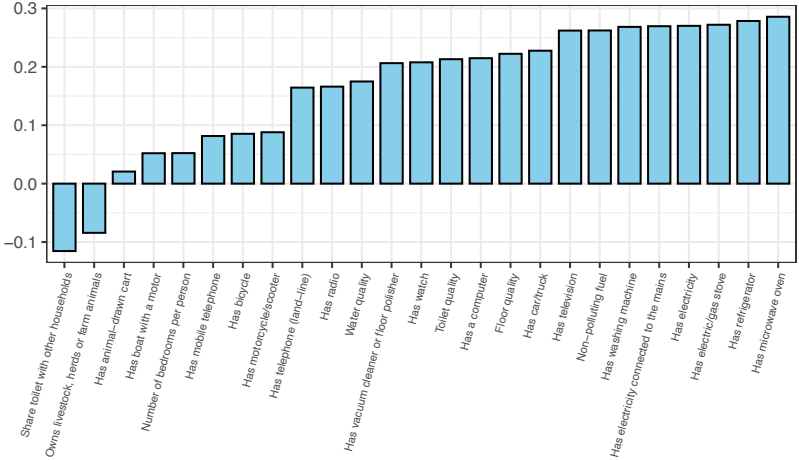


Figure: Weights from the first principal component from wealth index using method (i).

DATA

- To test the robustness of a wealth index we compare multiple configurations.
  - (i) Wealth index (Assets and non-assets)
  - (ii) Wealth index (Assets only)
  - (iii) Wealth index using variables from Yeh *et al.* (2020)
  - (iv) Sum of assets owned
- We perform a PCA for (i)-(iii) and extract the first principle component as the wealth indexes above and sum the number of assets owned (see Table 2) for (iv).
- Performed on the household level, and also aggregated to the cluster level.

PCA Analysis

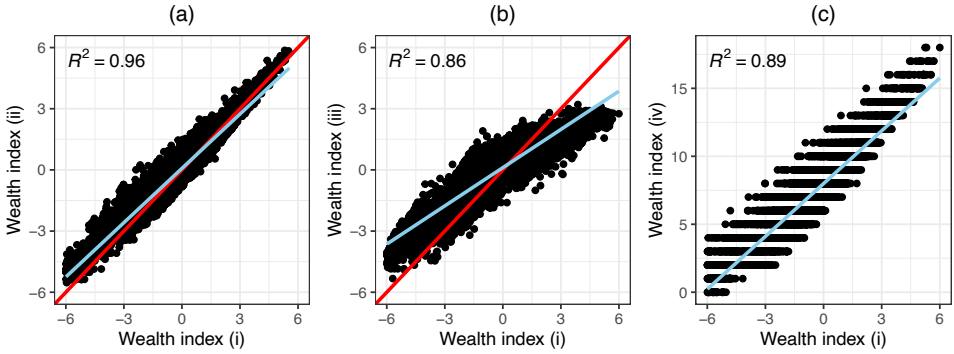


Figure: Comparison of the wealth indexes (described above) (a) method (ii) vs. (i), (b) method (iii) vs. (i), and (c) method (iv) vs. (i). Red line denotes the y=x, and blue line denotes the results of fitting linear regression to the scores combination. Associated R2 from each linear regression shown in the top left corner of each plot.

PCA Analysis

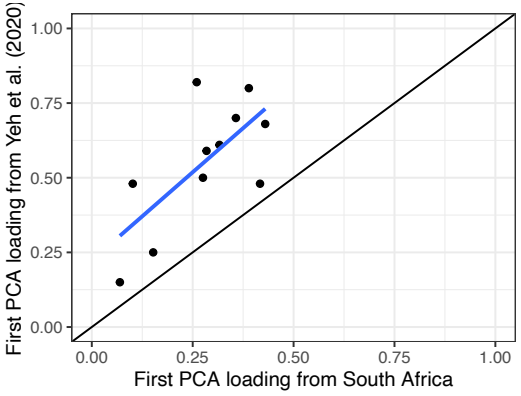


Figure: Comparison of the weights from the principal components analysis from Yeh *et al.* (2020) and the South Africa analysis

Wealth index

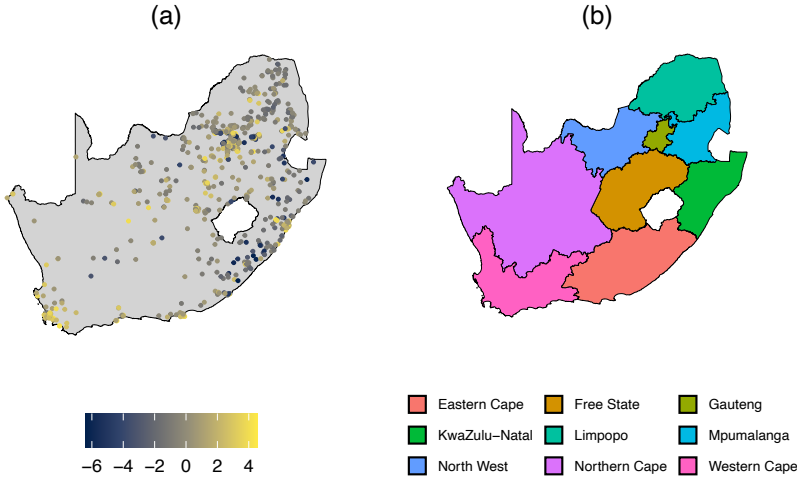


Figure: (a) Estimated wealth index at the cluster level for South Africa. Colours denote the scale of the wealth index and (b) Map of South African provinces.

WEALTH INDEX

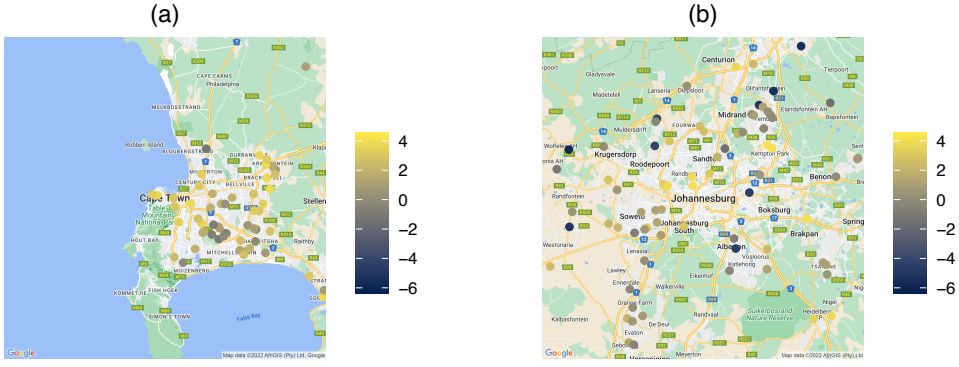
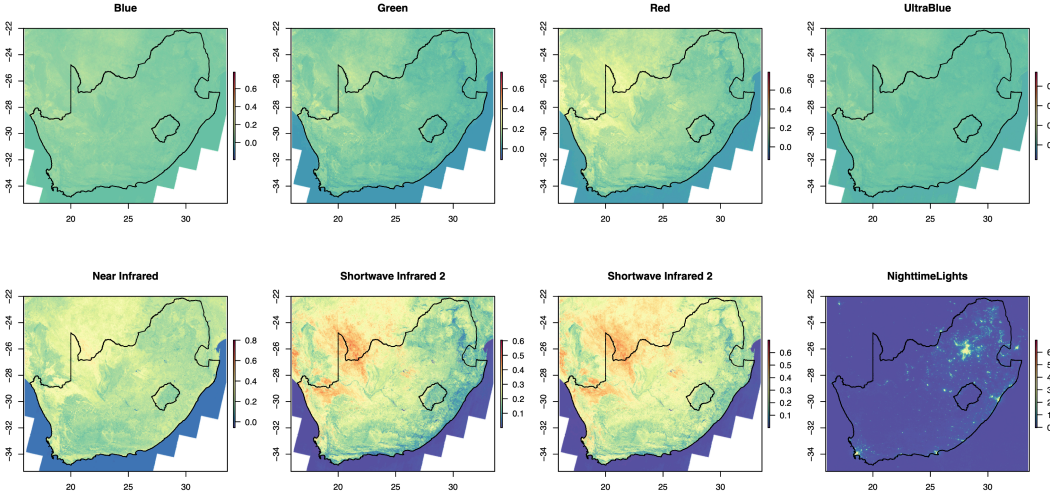


Figure: Maps of the estimated wealth index at the cluster level for (a) Cape Town and (b) Johannesburg. Colours denote the scale of the wealth index.

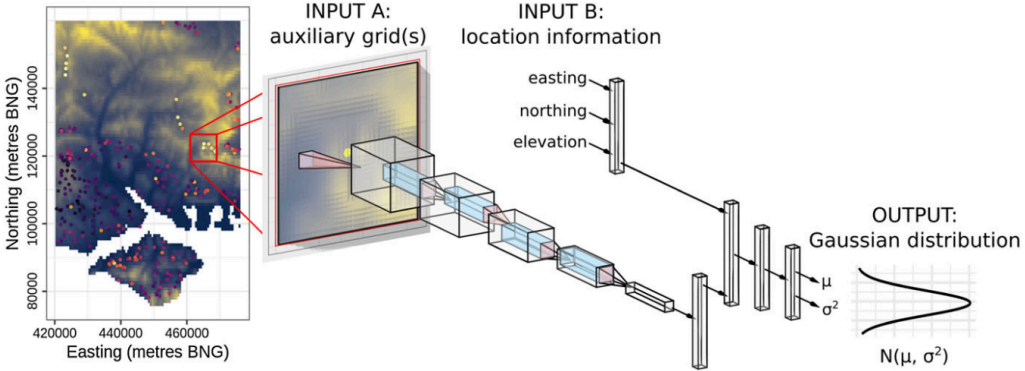
Summary

- Created a wealth index, which describes the relative wealth of regions in South Africa.
- Wealth index was robust to the variables included and it was shown that those with more assets had a higher wealth index.
- Generally lowest scores lived outside of the populated cities such as Cape Town and Johannesburg.
- However, we also found that there was considerable variation, even within the previously identified wealthier cities.
- Only used survey data, where we have estimated wealth at a relatively small number of points and cannot give us a complete picture of socio-economic status across South Africa.
- Next stage is to use (Bayesian) deep learning models, modelled from night-time lights data and multi-spectral imagery.

SUMMARY


Blue
Green
Red
UltraBlue
Near Infrared
Shortwave Infrared 2
Shortwave Infrared 2
NighttimeLights

SUMMARY


INPUT A:
auxiliary grid(s)
INPUT B:
location information
easting
northing
elevation
OUTPUT:
Gaussian distribution
μ
σ²
N(μ, σ²)
Northing (metres BNG)
80000
100000
120000
140000
420000
440000
460000
Easting (metres BNG)

Questions