EconoClusters

Efficient Definition of SES in SSA using only 4 assets

# Introduction

The table below links the clustering results from 33 countries, as detailed in our accompanying paper (link to come). The project's overall goal was to derive meaningful socio-economic status (SES) clusters of households using the DHS data. In addition, we did so by using a minimal set (4 or 5) of SES/asset variables rather than the number of relevant variables collected in the DHS survey (> 30). The impetus for this comes from our studies of trauma and injury in low-resource settings in sub-Saharan Africa. We have found (REF) that SES status significantly predicts health outcomes post-injury. Still, collecting data for research purposes on trauma victims is challenging and time-constrained, and assessing SES status must be very efficient (can only ask 4-5 questions about SES to the injury victim or their family members in the context of, particularly, severe trauma).

The user has access to several resources, including

1. Results of creating clusters based on a small subset of variables for 33 countries here.
2. Example of more extensive output by country with annotation for help in interpretation here.
3. Access to code used to both process the original DHS data and used to run the clustering analyses.

# Methods

The EconomicClusters algorithm was applied to 33 SSA countries with available Demographic Health Survey (DHS) data from 2010 or later. The EconomicClusters ran weighted k-medoids clustering on all combinations of four asset variables that characterized each country's most distinct economic groups of five. Cluster optimality was based on the maximum average silhouette width (ASW), an average measure of each observation's similarity within their cluster versus neighboring clusters**.** We validated this by analyzing cluster associations with established DHS proxies correlated with SES, including child mortality, public/private healthcare reliance, and women’s educational attainment. These evaluations are included in the linked documents.

# Key Messages

The DHS data provides a rich source of information about SES, but having available algorithms to define SES groups parsimoniously has been underdeveloped. Studies of acute trauma and injury require parsimonious characterization of SES, preventing extensive questionnaires such as the DHS. We proposed an updated, generalizable process for identifying asset variables that classify households into ranked asset groups. This framework permits the succinct identification of important asset variables in trauma registry data collection that will be used and continually updated to identify and support at-risk groups for trauma and injury.

# Using the output files

Before we go into detail, these country-level files can be helpful in several ways:

1. The output provides candidate sets of variables (about) equivalent to their ability to group (cluster) subjects by SES. We rank these sets by our measure of cluster performance, called the average silhouette width (ASW), where bigger means better clusters (more similarity among households in the same cluster, more significant differences among households in different clusters). If you are designing your study, these variable sets can be used, informed by your knowledge of the local context, to select which SES/asset variables to include on a future survey where determining SES status is of interest.
2. We also report which variables are consistently chosen among the variables that result in the best-performing clusters. This can also aid in selecting a variable list for a future study.
3. Provide the asset variable patterns in each cluster but only among the top-performing variable set. These can help interpret and/or rank the clusters from lowest to highest SES.
4. We show the statistical relationship of how we rank the clusters (based on the average number of assets in the cluster) to the validation variables and provide a formal inference of this relationship. The greater the magnitude of the association (in the expected direction) and the smaller the p-value, the stronger the evidence for the proposed SES ranking of clusters.

# Example Output: Cameroon

Below, we provide each section of the output and some annotations to help interpret and use it.

## Header Info

Economic Clustering Summary Report: Cameroon 2018

**Data Summary**

**Country Code-year:** CM18 (DHS code)

**Number of observations:** 11710 (total number of households in original survey data)

**Number of variables used:** 32 (number of potential asset variables available for clustering households)

**Variables used in the algorithm:** hv206, hv207, hv208, hv209, hv211, hv225, hv227, hv230b, hv232, hv232y, hv237, hv243a, hv243b, hv244, hv246, hv247, hv252, sh121h, sh121j, sh121k, sh121l, sh121m, sh121o, sh121s, sh122h, sh123a, water, toilet, floor, roof, cookfuel, wall (the DHS variable names of those 32 potential clustering variables).

## Top 10 Variable Sets (Sorted by ASW)

Below are the 10 asset variable sets that result in the 10 most optimal clustering configurations. There is no meaningful ordering of the variables within each set. The ASW column reports the measure of cluster performance that goes from -1 (worst) to 1 (perfect).

| Set | ASW | Var.1 | Var.2 | Var.3 | Var.4 |
| --- | --- | --- | --- | --- | --- |
| 1 | 0.9476 | has television | has mobile telephone | mixer | cable |
| 2 | 0.9381 | has television | has mobile telephone | cable | laptop computer |
| 3 | 0.9319 | has electricity | has television | mixer | cable |
| 4 | 0.9301 | has television | mixer | cable | roof |
| 5 | 0.9245 | has electricity | has television | cable | laptop computer |
| 6 | 0.9222 | has electricity | has mobile telephone | mixer | cable |
| 7 | 0.9217 | has television | has mobile telephone | has bank account | cable |
| 8 | 0.9202 | has television | cable | laptop computer | roof |
| 9 | 0.9184 | has electricity | has television | has mobile telephone | cable |
| 10 | 0.9183 | has television | has refrigerator | has mobile telephone | cable |

## Distributions of the SES variables chosen in top 10 cluster configuration sets and the proportion of times they were included in these top configurations (NOT SURE ABOUT THIS SO CORRECT)

This table provides some basic information on the marginal distribution of the levels of each variable, as well as the % of time they were included in the top cluster configuration sets (a measure of how important the variable is to clustering households).

| Variable | Description | % in Top Sets | Distribution |
| --- | --- | --- | --- |
| hv208 | has television | 90% | Binary, 52% (1/yes) |
| hv206 | has electricity | 40% | Binary, 64% (1/yes) |
| hv243a | has mobile telephone | 60% | Binary, 85.6% (1/yes) |
| sh121m | mixer | 40% | Binary, 12.7% (1/yes) |
| sh121o | cable | 100% | Binary, 38.3% (1/yes) |
| hv209 | has refrigerator | 10% | Binary, 22.3% (1/yes) |
| hv247 | has bank account | 10% | Binary, 16.3% (1/yes) |
| sh122h | laptop computer | 30% | Binary, 12.3% (1/yes) |
| roof | roof | 20% | Categorical, 0 (finished) = 84.1%,   1 (natural) = 13.1%,   2 (other) = 0%,   3 (rudimentary) = 2.8% |

## Patterns that define clusters

# Cluster Configuration (Set #1)

| Cluster Group | has television | has mobile telephone | mixer | cable | Proportion (%) |
| --- | --- | --- | --- | --- | --- |
| 1 | 0 | 1 | 1 | 1 | 11.32 |
| 1 | 1 | 0 | 1 | 1 |  |
| 1 | 1 | 1 | 1 | 1 |  |
| 2 | 1 | 0 | 0 | 0 | 13.97 |
| 2 | 1 | 0 | 1 | 0 |  |
| 2 | 1 | 1 | 0 | 0 |  |
| 2 | 1 | 1 | 1 | 0 |  |
| 3 | 1 | 0 | 0 | 1 | 26.71 |
| 3 | 1 | 1 | 0 | 1 |  |
| 4 | 0 | 1 | 0 | 0 | 35.68 |
| 4 | 0 | 1 | 0 | 1 |  |
| 4 | 0 | 1 | 1 | 0 |  |
| 5 | 0 | 0 | 0 | 0 | 12.32 |
| 5 | 0 | 0 | 1 | 0 |  |

## Relationship of cluster order and validation variables.

First, we only provide these results for the top cluster configuration set (that with the highest ASW). There are three tables per country based on other validation variables, but we will illustrate how to interpret these tables with child mortality. As one goes lower in the rows, the clusters are estimated to be poorer, and as one goes farther to the right in columns, the outcomes are worse (higher % of household child mortality). The p-value relates to a chi-square association test (IS THIS TRUE OR IS THIS BASED ON THE ORDERED LOGISTIC REGRESSION?). In this case, one can see a clear dose response in that the clusters with lower estimated asset wealth have higher child mortality. The p-value indicates that this relationship is significant.

**Using Children Deceased (Sorted by proportion of 0%)**

(higher SES as one goes down)

| Cluster ID/Children Deceased | 0% | 1-33% | 34-66% | 67+% |
| --- | --- | --- | --- | --- |
| 1\* | 1,638 (89.8%) | 119 (6.5%) | 55 (3.0%) | 12 (0.7%) |
| 3 | 3,364 (85.9%) | 386 (9.9%) | 132 (3.4%) | 34 (0.9%) |
| 2 | 1,833 (80.3%) | 321 (14.1%) | 105 (4.6%) | 25 (1.1%) |
| 4 | 3,966 (75.0%) | 954 (18.1%) | 288 (5.4%) | 77 (1.5%) |
| 5 | 926 (67.7%) | 261 (19.1%) | 135 (9.9%) | 46 (3.4%) |
| Total | 11,727 (79.9%) | 2,041 (13.9%) | 715 (4.9%) | 194 (1.3%) |
| \*The chi-squared p-value is 0 | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Survey Data Time | Number of Observations | Total Variables in Algorithm | Highest ASW Asset Subset | ASW |
| [Angola](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/AO1516_summaryscript.docx) | 2015 | 15857 | 21 | Has television, has refrigerator, has computer, type of roof development | 0.9541 |
| Benin | 2017 | 14028 | 18 | Has television, has mobile telephone, has bank account, has vcd/dvd player | 0.8996 |
| [Burkina Faso](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/BF1718_summaryscript.docx) | 2021 | 13248 | 23 | Has television, has wardrobe/bookcase, type of floor development, type of roof development | 0.8959 |
| [Burundi](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/BU1617_summaryscript.docx) | 2016-2017 | 15977 | 23 | has electricity, has mattress, type of toilet facility, type of roof development | 0.9289 |
| [Cameroon](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/CM18_summaryscript.docx) | 2018 | 11710 | 32 | has television, has mobile telephone, has mixer, has cable | 0.9476 |
| Chad | 2014 | 17193 | 17 | Has animal-drawn cart, owns land useable for agriculture, owns livestock, herds or farm animals, type of cookfuel used | 0.8641 |
| Comoros | 2012 | 4355 | 21 | Has television, has electricity, has refrigerator, has telephone (land-line) | 0.8835 |
| [Congo Democratic](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/CD1314_summaryscript.docx) | 2013-2014 | 18171 | 19 | Has television, has electricity, has mobile telephone, has chair | 0.9058 |
| [Ethiopia](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/ET19_summaryscript.docx) | 2019 | 8663 | 21 | Has television, has refrigerator, has mobile telephone, has electric mitad | 0.9556 |
| [Gabon](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/GA12_summaryscript.docx) | 2012 | 8924 | 31 | Has car/truck, has mobile telephone, has hifi, type of roof facility | 0.9865 |
| Gambia | 2019-2020 | 6544 | 32 | Has television, has dvd/vcd player, type of floor development, type of roof development | 0.9511 |
| Ghana | 2019 | 17819 | 27 | Has electricity, has computer, has freezer, type of roof development | 0.9306 |
| Guinea | 2018 | 7899 | 25 | Has electricity, has television, has refrigerator, has mobile telephone | 0.9448 |
| [Ivory Coast](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/CI11_summaryscript.docx) | 2016-2017 | 9677 | 23 | has television,has mobile telephone, owns cable subscription, owns cd/dvd player | 0.8964 |
| [Kenya](https://github.com/ahubb40/EconoCluster/blob/main/EC%20Summary%20Reports/KE20_summaryscript.docx) | 2022 | 37838 | 27 | has television, mobile phone used for financial transaction, has dvd player, type of roof development | 0.9268 |
| Lesotho | 2014 | 9402 | 20 | Has electricity, has car/truck, has mobile telephone, has battery or generator | 0.9017 |
| Liberia | 2019-2020 | 9006 | 20 | Has television, has cupboard, has mattress, type of roof development | 0.8745 |
| Madagascar | 2021 | 20494 | 22 | Has television, has electricity, has bed, had a dvd/divx player | 0.9118 |
| Malawi | 2015-2016 | 26109 | 21 | Has television, has electricity, has mobile telephone, type of floor development | 0.9160 |
| Mali | 2018 | 9494 | 24 | Has refrigerator, has mobile telephone, has mosquito bed net for sleeping, has wardrobe/bookcase | 0.9215 |
| Mozambique | 2011 | 13699 | 19 | Has television, has electricity, has refrigerator, has mobile telephone | 0.9141 |
| Namibia | 2013 | 9573 | 26 | Has electricity, has refrigerator, has mobile telephone, has microwave | 0.9016 |
| Niger | 2012 | 10505 | 20 | Has television, has electricity, has vcr/dvd, anything done to water to make safe to drink | 0.9532 |
| Nigeria | 2018 | 40403 | 25 | Has refrigerator, has mobile telephone, has electric iron, has fan | 0.8879 |
| Rwanda | 2019 | 12948 | 26 | Has television, has mattress, has cupboard, type of roof facility | 0.9930 |
| Senegal | 2019 | 4538 | 24 | Has television, has computer, has refrigerator, type of roof facility | 0.9219 |
| Sierra Leone | 2019 | 13141 | 18 | Has television, has electricity, has refrigerator, has mobile telephone | 0.9363 |
| South Africa | 2016 | 11083 | 20 | Has electricity, has refrigerator, vacuum cleaner or floor polisher, has washing machine | 0.9411 |
| Tanzania | 2022 | 15660 | 29 | Has television, has refrigerator, has bed, has cd/dvd player | 0.9076 |
| Togo | 2013-2014 | 9532 | 20 | Has television, has electricity, has cd/dvd reader, type of roof facility | 0.9165 |
| Uganda | 2016 | 19519 | 29 | Has television, has electricity, has cassette/cd/dvd player, has bed | 0.9252 |
| Zambia | 2018 | 12537 | 23 | Has electricity, has refrigerator, type of floor development, type of roof development | 0.9287 |
| Zimbabwe | 2015 | 10356 | 34 | Has television, has mobile telephone, has computer, has dish/decoder | 0.9285 |