Report: DHS and Geo-covariates data integration
Case study on Bangladesh survey 2014

Yichun Wang

For more information, please contact the author (email: yichun.wang@gmail.com).

Contents

Contents 2
A. Background Introduction 3
B. Setting and steps of the project 4
C. Methodologies 5
1) Locating and integrating geo-covariates from sources 5
2) Identify DHS variables for analysis 12
3) Predictive Models 17
3-1) Random Forest 18
3-2) Xgboost 18
4) Variable consideration 19
5) Adding individual characteristics and disaggregate data 20
6) Bayesian Geo-statistical model 21
D. Results and conclusions of Bangladesh data integration 22
1) Key observations 34
2) Bayesian Geo-statistical model 34
E. Summary and further research 39
Reference 40

A. Background Introduction

In September 2015, heads of Member States at UN agreed to the Sustainable Development Goals (SDGs) and to full implementation of the agenda by year 2030. The SDGs include ending poverty and malnutrition, improving health and education, empower women and eliminate gender disparity, and building resilience to natural disasters and climate change.

To monitor the indicators in these areas, household survey data such as Demographic and Health Surveys (DHS) (http://dhsprogram.com), Living Standard Measurement Surveys (LSMS) and Multiple Indicator Cluster Surveys (MICS) (http://mics.unicef.org) have been used more extensively to estimate national level statistics of related indicators.

However, increasingly, people see the need to link the household survey data to other data sources that provide larger context on the services and levels of development to better understand the distribution of inequality. To decrease inequities and accelerate progress towards SDGs, we need more detailed understanding of what the inequality is associated with, rather than just urban-rural, male-female, rich-poor divide. We need to integrate other reliable data sources to understand how unequal access to infrastructure, natural resources, difference in climate, etc. affect indicators related to poverty, living standard, education, health, an attitude towards women. This project is an attempt to form a methodology that can help to see how data from outside sources can be integrated and used, and how statistical tools can be used to help us achieve better understanding and tracking of where disadvantaged people are located.

With sharp increase in publicly available geo-spatial data as well as new and more accessible geographic information system (GIS) technologies and methods, the research community has started using geographic covariates when tracking malaria, vaccination coverage, poverty mapping (ref. 7, 8, 9). Data with features derived from remote sensing data such as night-light composite data, enhanced vegetation index data, aridity data, human foot print index, are considered environmental and physical metrics likely to be associated with human welfare. More sophisticated data on traveling road condition and travel time to cities also have potential impact. Raster files of these geo-spatial data at different grid levels can be downloaded from internet for all to use.

The DHS Program routinely collects geographic coordinates of the primary sampling units (PSU, also known as cluster) in most surveyed countries. Although not all GIS data are released to the public, there are rich resources of survey data over time from DHS, with random displacement added to protect the privacy of the survey respondents. Using GIS, researchers can link DHS data with routine health data, health facility locations, local infrastructure such as roads and rivers, and environmental conditions. In 2015, the DHS Program conducted interviews with geospatial experts to obtain guidance for curating a list of geo-covariates. In the meantime, DHS solicited users through the website for research interest. The result is 22 geographic covariates from 15 different data sources that have been selected. For details see Ref. 5.

Since September 2017, DHS Program has been providing set of geospatial covariates in addition to survey cluster Global Positioning System (GPS) data collected during each survey. These covariate datasets are available for download in DHS Program website (https://dhsprogram.com/data/available-

datasets.cfm) and also in Spatial Data Repository (https://spatialdata.dhsprogram.com/covariates/) (ref 5).

This project started by a search for existing data sources and software resources, as well as experiment on a system of methods. While most of the research is on tracking or mapping affected people in the countries, we wanted to find out the structure of variation. Variation on household poverty level has two components: between cluster variation, e.g., how poverty level differs between geographic locations (of the PSU); and within cluster variation, e.g., how poverty level differs within the same geographic location among individual households. One can imagine that a geographic location that is far from cities, with little infrastructure (road or night-light), and bad environmental conditions such as high aridity and low vegetation index, would generally be poor. But not every household in the same location would be equally poor. Total variation is the sum of the two variations.

Geographic disaggregation is of lots of interest, because in this way governments and related agencies know where to target help. Disaggregation on individual characteristics such as age, sex, education is of interest too; it is especially insightful to understand the disparity of different groups after controlling their geographic features. For example, it is often that "better clusters" have people with higher education attainment. So, we would like to look at the effect of education attainment after we control the features of the clusters. Proper modeling would also enable us to identify indicators where age, sex or education has significant disparity and investigate the pattern in relation to geo-location related features.

B. Setting and steps of the project

Our approach has three stages:

1) We examine the cluster level variation explained by geo-covariates for various indicators for Bangladesh, to understand how infrastructure, natural environments affect the level of the clusters indicators.

2) We add individual household/member characteristics such as age, sex, education, to the model to understand how/whether these subgroups have different levels in similar infrastructure and natural environments.

3) Finally, we experiment with Bayesian geo-statistical models, considered as state of art analytic techniques in the literature.

At the end of this study, in addition to concrete observations on the indicators and data from Bangladesh, we would also provide R-source code and data documents that establish the technical procedures to achieve data integration of survey data with GIS coordinates data. The observation on Bangladesh is a case study which can serves as a guide on how to proceed with other survey data.

The followings are the concrete research questions and issues related to selected indicators (such as household basic needs, education attainment, nutrition needs of under 5 children, maternal health,

women's economic empowerment, etc.) that we try to address by integrating survey data with geo-covariate data.

1) How much selected indicators vary because of the location of the household and the conditions that is associated with their locations as indicated in the geo-covariates.

2) Without data integration, we can only use urban/rural divide and administration areas to dis-aggregate geographically. However, with fast urbanization, infrastructure building with road and electricity, and climate change, we need to have better understanding of how these factors impact on the needs of those people in different geographic locations. Thus, a system of updating the geo-covariates reliably and timely is of great value.

3) By integrating the geo-covariate data with DHS survey data, we can attempt to build a predictive model that generates a map or gives an estimate of indicator levels of a local community (urban cluster or village) only based on their location.

4) After understanding the location's impact on the selected indicators, we can also incorporate individual variables such as sex of household head, age, and education attainment of each household member. This will allow us to understand, given the same geo-location and environment, which demographic individual characteristics have impacted individuals.

5) We can further disaggregate indicators by the groups that have major impact in the model. We can compare the variation at the individual level within each cluster with the variation at cluster level and try to understand the main source of variation.

Understanding the sources of variation are important. That can guide governments in better targeting of their interventions. The difference will imply different policy emphasis. If an indicator is affected more on its location through geo-covariate such as its travel time to the cities, or aridity of the areas, infrastructure intervention is more needed. But if an indicator has little variation sourced from its geo-location, more variation from household/individual characteristics, effort to change individual choice preferences or the culture itself might be the preferred approach.

C. Methodologies

1) Locating and integrating geo-covariates from sources

As already mentioned, since late 2017, DHS started assembling geo-covariates data and publish them for user's convenience. This list of geo-covariates was decided based on experts' discussion and users' feedback. We started this project with this list of variables. For the complete list, see table 1 in appendix.

Part of the work is to create and document the knowledge and techniques to locate related data sources and integrate them to the DHS survey data. All the geo-covariates data are stored in raster data format, which is one of the two types of digital format for GIS data.[1] The raster files are map projections, they

[1] More introduction on raster files can be found at https://en.wikipedia.org/wiki/GIS_file_formats

try to portray the surface of the earth or a portion of the earth on a flat piece of paper or computer screen. In a lay man's term, map projections try to transform the earth from its spherical shape (3D) to a planar shape (2D) so that maps can be made on flat layers (represented as two-dimensional image file). In short, they are files consist of columns and rows which is mapped into geo-coordinate system by the institutes that generate the data. The cells store information for each geo location determined by the coordinates (column and row). The size of the cell is the corresponding ground units. Because the structure of the format is identical to image files, "cells' are often called pixels and the size of the ground unites are called resolution. This type of data is also often called grid data, or a map.

Here are the steps we take to integrate the geo-covariates stored in a raster file to the DHS data:

1. Getting GIS coordinates from shapefile:
   dhsShapeData<-readShapePoints(paste(shapedata_folder, "BDGE71FL.shp" , sep=""))
   the coordinate reference system (crs) of the DHS shapefile: crs: +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0

2. Reading raster files downloaded from the source websites (taking SMOD raster file as an example):
   smodData<-raster(paste(geodata_folder,
   "GHS_SMOD_POP2015_GLOBE_R2016A_54009_1k_v1_0.tif", sep=""))

3. Checking the coordinates reference system[2]; if it is different from the coordinate reference system (crs) of DHS shapefile, we need to transform the shapefile:
   dhsShapeData2 <- spTransform(dhsShapeData, smodData@crs)

4. Using R-library facility to extract information from smodData:
   dhs_all2000 <- extract(smodData, # raster layer
   dhsShapeData2, # SPDF with centroids for buffer
   buffer = 2000, # buffer size, units depend on CRS
   df=TRUE) # return a dataframe
   for each DHS cluster, this returns all the smod data within 2km radius
   dhs_all2000<-as.data.frame(dhs_all2000)

5. Taking the average of the values extracted from 2km radius:
   smodData<-aggregate(dhs_all2000$GHS_SMOD_POP2015_GLOBE_R2016A_54009_1k_v1_0, by=list(dhs_all2000$ID), FUN=mean)

We decided not to integrate all the variables in the DHS list and instead we went for a simplified list (Table 1).

[2] A coordinate reference system (CRS) defines, with the help of coordinates, how the two-dimensional, projected map in GIS is related to real places on the earth. The decision as to which map projection and coordinate reference system to use, depends on the regional extent of the area one wants to work in, on the analysis and often on the availability of data. There are thousands of different projections used when people produce the raster files. Different map producing agencies may use different projection.

Table 1: Choice of geo-covariates for DHS survey data

Geo-Covariate Name	Variable Name in DHS database	Variable Definition	Data link	Update frequency
Travel_Times2015	Travel_Times	Travelling time (in minutes) to the nearest city of more than 50,000 people. (Resolution: 1km X 1km grid)	https://map.ox.ac.uk/wp-content/uploads/accessibility/accessibility_to_cities_2015_v1.0.zip	Previous data was made in 2000. New data is based on Open street map and google roads map from 2015.
SMOD2015	SMOD_Population_2015	1 = "rural cells" or base (BASE) 2 = "urban clusters" or low density clusters (LDC) 1,500-50,000 inhabitant/km2 3 = "urban centers" or high density clusters (HDC) >50,000 inhabitants/km2	http://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_SMOD_POP_GLOBE_R2016A/	This data package contains an assessment of the REGIO-OECD "degree of urbanization" model using as input the population GRID cells.
Buildup2015	BUILT_Population_2014	The percentage of building footprint area in relation to the total cell area. (Resolution: 1km X 1km)	http://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_BUILT_LDSMT_GLOBE_R2015B/	These data contain an information layer on built-up presence as derived from ad-hoc Landsat 8 collection 2013/2014 image collections, produced by means of Global Human Settlement Layer methodology in 2015.
Friction2015	-[3]	Calculated land-based travel speed for given geo-coordinate position that lies between 85 degrees north and 60 degrees south for a nominal year 2015	https://map.ox.ac.uk/wp-content/uploads/accessibility/friction_surface_2015_v1.0.zip	First time release.

[3] This is not available in DHS GC file, have been integrated from worldpop files.

Geo-Covariate Name	Variable Name in DHS database	Variable Definition	Data link	Update frequency
Nightlight2015	Nightlights_Composite	Average radiance composite from night time satellite image data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB)	https://data.ngdc.noaa.gov/instruments/remote-sensing/passive/spectrometers-radiometers/imaging/viirs/dnb_composites/v10//2015/SVDNB_npp_20150101-20151231_75N060E_v10_c201701311200.tgz	Temporal averaging is done on a monthly and annual basis. We choose to use the 2015 annual data "vcm-orm-ntl" (VIIRS Cloud Mask - Outlier Removed - Nighttime Lights)
Avi2015	Enhanced_Vegetation_Index_2015	Vegetation Indices (VI) are robust, empirical measures of vegetation activity at the land surface. They are designed to enhance the vegetation reflected signal from measured spectral responses by combining two (or more) wavebands, often in the red (0.6 - 0.7 μm) and NIR wavelengths (0.7 - 1.1 μm) regions.	https://e4ftl01.cr.usgs.gov/MOLT/	MOD13A3: Monthly 1km VI (data updated on a monthly basis). 201406 data downloaded.
Hfp2004	Global_Human_Footprint	An updated map (based on data from of anthropogenic impacts on the environment in geographic projection which can be used in wildlife conservation planning, natural resource management, and research on human-environment interactions.	http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic/data-download	Created in 2005 based on data from 1995-2004.
Aridity2000	Aridity	Climate data related to evapotranspiration processes and rainfall deficit for potential vegetative growth.	https://cgiarcsi.community/data/global-aridity-and-pet-database/	The Global-Aridity are modeled using the data available from WorldClim Global Climate Data from 1950-2000 (http://WorldClim.org)

Geo-Covariate Name	Variable Name in DHS database	Variable Definition	Data link	Update frequency
DroughtEpisode	Drought_Episodes	Provide a means of assessing the relative distribution and frequency of global drought hazard. Drought events are identified when the magnitude of a monthly precipitation deficit is less than or equal to 50 percent of its long-term median value for three or more consecutive months.	http://sedac.ciesin.columbia.edu/data/set/ndh-drought-hazard-frequency-distribution/data-download	Utilizing average monthly precipitation data from 1980 through 2000 at a resolution of 2.5 degrees, this data was created in 2005.
Density2015	All_Population_Density_2015	Number of inhabitants per cell (1km X 1km)	http://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/GHSL/GHS_POP_GPW4_GLOBE_R2015A/	Residential population estimate for 2015 provided by CIESIN GPWv4 were disaggregated from census or administrative units to grid cells,
aWealthIndex2011	Not on file	2011 estimates of mean DHS wealth index score per grid square, and associated uncertainty metrics.	http://www.worldpop.org.uk/data/summary/?doi=10.5258/SOTON/WP00020	Created in 2017
aIncome2013	Not on file	2013 estimates of income in USD per grid square, and associated uncertainty metrics.	http://www.worldpop.org.uk/data/summary/?doi=10.5258/SOTON/WP00020	Created in 2017
APP2013	Not on file	2013 estimates of mean likelihood of living in poverty per grid square, as defined by $2.50 a day poverty line, and associated uncertainty metrics.	http://www.worldpop.org.uk/data/summary/?doi=10.5258/SOTON/WP00020	Created in 2017

The last three variables are from World Bank, not included in DHS datasets, but added after research on relevant information. They contain information on wealth index from DHS, Household Expenditure from Poverty Probability Index (PPI)[4], and reported household income from Grameen Bank. World bank used these three indices, integrated them with mobile phone subscription data and remote sensing data built. For each index, World Bank built data-mining models and Bayesian geospatial models on the sample, and extended projections to the entire grid of the country. As these data reflect wealth, income and expenditure of the geolocations of the country, we think we should take advantage of this information whenever it is available. For any country, if similar grid data from World Bank is available, we think this direct estimation on wealth/income of location should be included in data integration.

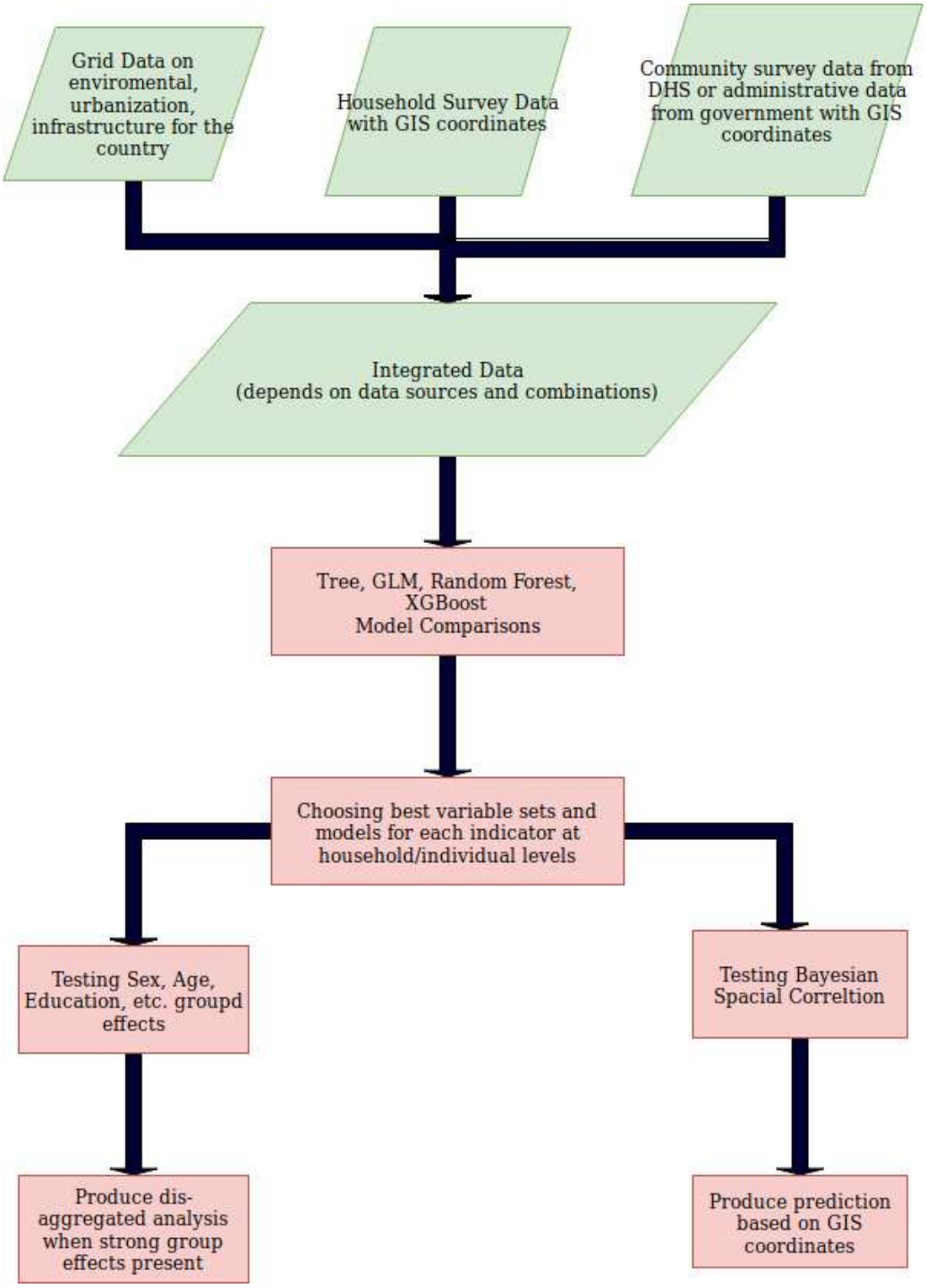
Even though for DHS location, the GIS coordinates have been masked with 2km and 5km random disturbance for urban/rural location respectively; algorithm to integrate the data must take this noise into account. DHS queried data within 2km and 5km radius of the published urban/rural location and take the average as the value for the DHS location. In this study, we decided to use 2km radius for all clusters. this will make it easy to apply the predictive model to any GIS coordinates without knowing whether this location is classified as rural or urban by the government administration. To make sure that this shortcut of data integration does not impair the model's goodness of fit, we compare models build on geo-covariates from DHS and from our own integration.

An illustration of the flow of the project is also presented in figure 1.

DHS also conducts community survey, in which a knowledgeable person from the PSU will answer questions such as how far are the nearest hospital, schools, market, and what kind of road the village is accessible to, whether there is cooperative or Grameen bank locally, etc. We choose 9 of the variables from this community survey data (see appendix for details). We conduct statistical tests to see the effect of these variables in addition to the geo-covariates in predicting indicator values.

[4] https://www.povertyindex.org/country/bangladesh

Figure 1. An illustration of project flow



2) Identify DHS variables for analysis

In June 2013, DHS organized a working group meeting to discuss the use of geographic data from DHS population-based surveys for spatial interpolation (ref 2). Several data-related factors need consideration when using DHS data for spatial interpolation; they include:

1. indicator is a robust measurement,
2. indicator is not a rare event,
3. indicator is spatially distributed,
4. indicator has specific reference period,
5. indicator is not temporally related, and
6. indicator relates to the current location of the respondent.

We choose a list of 13 indicators that measure the household needs, education attainment (segmented by age), childhood nutrition, and women's health care needs and decision power at home, all related to the SDG targets (Table 2).

Table 2: Household/Individual level measured indicators selected

Variable Names	Variable Name in DHS database	Target Population	Definition	R-code
Poverty	hr$HV271 (renamed: hr$WealthIndex)	All Households	Wealth index factor score (5 decimals). The wealth index is a composite measure of a household's cumulative living standard calculated based on a household's ownership of selected assets, such as televisions and bicycles; materials used for housing construction; and types of water access and sanitation facilities and generated using principal components analysis. Since around 17% of Bangladesh lives in extreme poverty, we use -0.87 as the cut off value. Any household with wealth index lower than -0.87 is considered to be poor.	hr$Poverty <- (hr$WealthIndex<= -0.87)
AccessElectricity	hr$HV206 (renamed: hr$ AccessElectricity)	All Households	Whether the household has electricity	hr$AccessElectricity <- !(hr$AccessElectricity == "No")
SafeSanitation	hr$HV205(renamed: hr$ SafeSanitation)	All Households	Type of toilet facility in the household. Safe includes following categories only: "Flush to piped sewer system", "Flush to septic tank", "Flush to pit latrine", "Ventilated Improved Pit lat", "Pit latrine with slab"	hr$SafeSanitation <- (hr$SafeSanitation %in% c("Flush to piped sewer system", "Flush to septic tank", "Flush to pit latrine", "Ventilated Improved Pit lat", "Pit latrine with slab")
BankCard	Hr$HV247(renamed: hr$BankCard)	All Households	Any member of the household has a bank account	hr$BankCard<- (hr$BankCard=="Yes")

Variable Names	Variable Name in DHS database	Target Population	Definition	R-code
CleanFuel	hr$HV226(renamed: hr$CleanFuel)	All Households	Type of cooking fuel.	hr$CleanFuel<- (hr$CleanFuel %in% c("Electricity", "LPG", "Natural gas", "LPG/Natural Gas", "LPG, natural gas", "Liquid gas", "Biogas"))
CleanWater	hr$HV201 (renamed: hr$CleanWater)	All Households	Main source of drinking water for members of the household. Clean water includes following responses: “Piped into dwelling”, “Piped to yard/plot”, “Public tap/standpipe”, “Tube well or borehole”, “Protected well”, "Bottled water", "Protected spring", "Rainwater”)	hr$CleanWater = (hr$CleanWater %in%c(“Piped into dwelling”, “Piped to yard/plot”, “Public tap/standpipe”, “Tube well or borehole”, “Protected well”, "Bottled water", "Protected spring", "Rainwater”)
SchoolAgeEducation-LowerSecondaryAge	pr$HV121(renamed: pr$SchoolAgeEducation)	All household members of age 12-14	Child household member attended school during current school year.	pr$SchoolAgeEducation<- (pr$SchoolAttendance=="Attended at some time")
SchoolAgeEducation-UpperSecondaryAge	pr$HV121(renamed: pr$SchoolAgeEducation)	All household members of age 15-17	Child/young adult household member attended school during current school year.	pr$SchoolAgeEducation<- (pr$SchoolAttendance=="Attended at some time")
SchoolAgeEducation-CollegeAge	pr$HV121 (renamed: pr$SchoolAgeEducation)	All household members of age 18-22	Child/young adult household member attended school during current school year.	pr$SchoolAgeEducation<- (pr$SchoolAttendance=="Attended at some time")

Variable Names	Variable Name in DHS database	Target Population	Definition	R-code
AdultEducation-Older	pr$HV109 (renamed: pr$Education)	All household members of age 36- 60	Educational attainment codes the education of the household member into the following categories: None, incomplete primary, complete primary, incomplete secondary, complete secondary, higher education. Must have completed secondary or higher education for inclusion here.	pr$AdultEducation<- (pr$Education %in% c("Complete secondary", "Higher"))
AdultEducation-Young	pr$HV109(renamed: pr$Education)	All household members of age 23- 35	Educational attainment codes the education of the household member into the following categories: None, incomplete primary, complete primary, incomplete secondary, complete secondary, higher education. Must have completed secondary or higher education for inclusion here.	pr$AdultEducation<- (pr$Education %in% c("Complete secondary", "Higher"))
CurrentlyWorking	pr$SH13(renamed: pr$CurrentlyWorking)	All household members of age 18-60	Whether the respondent is currently working.	pr$CurrentlyWorking<- (pr$CurrentlyWorking=="Yes")
Stunt	pr$HC70 (renamed: pr$Stunt)	All household members of age 0-5	Measure of weight index of child compared to WHO Child Growth Standards.	pr$Stunt<- (as.numeric(as.character(pr$Stunting))< -200)
Underweight	pr$HC71 (renamed: pr$UnderWeight)	All household members of age 0-5	Measure of weight index of child compared to WHO Child Growth Standards.	pr$Underweight<- (as.numeric(as.character(pr$Underweight))< -200)

Variable Names	Variable Name in DHS database	Target Population	Definition	R-code
Waste	pr$HC72 (renamed: pr$Waste)	All household members of age 0-5	Measure of weight index of child compared to WHO Child Growth Standards.	pr$Waste <- (as.numeric(as.character(pr$Wasting))< -200)
PowerHealth	ir$V743A (renamed: ir$DecisionHealth)	All women currently married, age 15-49	Response from women' questionnaire for who has final say regarding respondent's health care, included are responses: "Respondent alone", "Respondent and husband/partner"	ir$PowerHealth<- ir$DecisionHealth %in% c("Respondent alone", "Respondent and husband/partner")
PowerPurchase	ir$V743B (renamed: ir$DecisionPurchase)	All women currently married, age 15-49	Response from women's questionnaire for who has final say regarding making large household purchases, included are responses: ("Respondent alone", "Respondent and husband/partner")	ir$PowerPurchase <- ir$DecisionPurchase %in% c("Respondent alone", "Respondent and husband/partner")
PowerVisitFamily	ir$V743C (renamed: ir$DecisionVisitFamily)	All women currently married, age 15-49	Response from women's questionnaire for who has final say regarding visits to family or relatives, included are responses: ("Respondent alone", "Respondent and husband/partner")	ir$PowerVisitFamily <- ir$DecisionVisitFamily %in% c("Respondent alone", "Respondent and husband/partner")
ProfessionalHelp	ir$M3A,ir$M3B,ir$M3C (renamed: ir$ProfessionalAssitance1, ir$ProfessionalAssitance2, ir$ProfessionalAssitance3)	All women who have delivered birth in the last 5 years	The type of person who assisted with the delivery of the child, included are 3 types of country-specific health professionals.	hr$ProfessionalHelp <- (grepl("1", hr$ProfessionalAssitance1) | grepl("1", hr$ProfessionalAssitance2) \| grepl("1", hr$ProfessionalAssitance3))

Our goal in this section is the following:
we want to understand the total variation of each indicator, defined as individual level standard deviation (household level or personal level). We also decompose this *variation into two components: cluster level variation, and individual level variation.*

$$Y_{ji} = \alpha + e_i(s_i) + p_{ji}$$

Where index i represents cluster i, index ji represents the j-th household in the cluster i. Y_{ji} is the observation of indicator value of *j*-th household (person) in cluster *i* in the sample survey. Here, e_i is the cluster effect, with location (s_i), and p_{ji} is the individual variation from the cluster they live in. We assume that a household/individual value Y_{ji} is influenced by the cluster they live in, and their household situation. The cluster effect $e_i(s_i)$ can be modeled in two parts: a Bayesian Geospatial model with spatial correlation, and effects of geo-covariates based on the characteristics of the location, such as population density, infrastructure development level, climate related information. Variation at cluster level can be represented at with two components: Var(spatial) and Var(geoCovariates). While p_{ji} can be modeled by household (personal) characteristics and individual variation.

So total variation observed on individuals , $SD_t^2 = \frac{\sum_{ji}(Y_{ji}-\bar{Y})^2}{n}$, can be estimated as sum of variation at cluster level, $SD_c^2 = \frac{\sum_i(\acute{e}_i-\acute{Y})^2}{m}$, and variate at personal(household) level , $SD_p^2 = \frac{\sum_i\sum_{ji}(Y_{ji}-\acute{e}_i)^2}{n}$. Here $\acute{Y}$ is the sample average, and $\acute{e}_i$ is cluster average, n and m are the total sample size and number of clusters.

$$SD_t^2 = SD_c^2 + SD_p^2$$

SD_c^2 is the cluster level variation, and part of it can be explained by the cluster's location and geo-covariates.
SD_p^2 is the individual level variation within the cluster, and part of it can be explained by individual characteristics of the household or person, such as age, sex, education, etc.

3) Predictive Models

Several pioneering works have been done in using the geo-covariate to predict various geo-located statistics. Random Forest, Artificial Neural Network, Bayesian Geo-statistical modeling are popular methods (ref 7,8,9,10). While Artificial Neural Network (ANN) is a popular new method, numerous case studies (comparison) did not show it superior to Random Forest or Bayesian Geo-statistical modeling, and the method is often considered a black box method for being hard to understand, we opted out the ANN methods. Instead, because of the growing interest in a new method called Xgboost as a competitor to Random Forest, we explore this new method.

Here we introduce the Random Forest and Xgboost methods and leave the Bayesian geo-statistical model for a later section. We also will use the conventional Decision Tree and GLM models in our model comparisons.)

3-1) Random Forest[5]

Random Forest is an ensemble algorithm based on decision trees. Decision tree is a statistical procedure that recursively partition the multi-dimension space of the explanatory variables, one variable at a time, based on the value of the response variable. The entire data is the root of the tree, with no partition. Each partition creates two new offspring-nodes of the parent node, resulting a tree-like data structure. Objective function (e.g., information entropy) is used to decide where to partition, and whether to partition. The advantage of the decision tree method is that it does not assume any form of relation, is invariant under any monotone transformation of the variables, and allow high level interactions between explanatory variables where interactions exist.

However, when the number of explanatory variables grows, decision tree can be too complicated and easily influenced by special cases; and is considered too easy to over-fit the training data and not suitable for new data. Random forest was created to address this shortcoming, by taking sub-samples of the training data and building trees for each sub-sample, and at the end taking average of all the trees as the prediction model.

Summary:

- *While not state of the art, it is still an efficient algorithm.*
- *It runs efficiently on large data bases.*
- *It can handle thousands of input variables without variable deletion.*
- *It gives estimates of what variables are important in the classification.*
- *It offers an experimental method for detecting variable interactions.*

3-2) Xgboost[6]

For a long time, Random forest was the most powerful predictive model in the machine learning community, flexible and robust. Until recent years, another method called gradient boosting became a powerful challenger. Like Random Forest, it builds trees on sub-samples. But its strategies are to build smaller trees with less depth and branches. Surprisingly, the results can be over-fitting – i.e.., fitting the sample data used in building the model exceedingly well and missing outside sample miserably. Xgboosting method allows several model parameters to fine tune the balance between precision in model building and robustness in outside sample fitting.

XgBoost method has won many recent machine learning competitions, the algorithm itself is more resource efficient on large scale data, it is worth exploring this method, in comparison to the popular random forest. If the prediction powers are similar, we will choose the one that is simpler to learn and run.

To compare the models, we use the root mean squared errors, RMSE in the total error and cluster level error. Let $\hat{Y}_{ji}$(model) be the ji-th predictive value of the model for the DHS sample indicator, and let:

[5] From https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
[6] From http://dmlc.cs.washington.edu/xgboost.html

$$RMSE_t^2(\text{model}) = \frac{\sum_{ji}(Y_{ji} - \widehat{Y_{ji}}\,(model))^2}{n}, \text{ for total error.}$$

$$RMSE_c^2(\text{model}) = \frac{\sum_{i}(\bar{e}_i - \widehat{Y_{ji}}\,(model))^2}{m}, \text{ for cluster error.}$$

Remember that $\acute{e}_i$is the cluster average of indicator values. When the model input is all at cluster level (as the geo covariates are), $\hat{Y}_{ji}$ is constant for all households/individuals in the same cluster.

The model can be random forest, Xgboost. We also used more familiar modeling method, such as, Generalized Linear Model (GLM) and decision tree model (Tree). The smaller RMSEs the better the model estimates are.

4) Variable consideration

In addition to the modeling methods, we also need to consider what data to use. Although DHS provides geo-covariate information on surveys where the GIS coordinates of the PSUs are provided, there are a few benefits of mastering the procedures to integrate the data ourselves.

Firstly, we can update the information whenever needed. For example, at the end of 2017, the Malaria Atlas Project (University of Oxford, United Kingdom) released the global map of accessibility to cities based on data of the year 2015 from two major sources: Google Road Map and Open Street Map. It is a major update of information from their previous release in 2008, based on data from 2000. Data integration released by DHS is based on the earlier version. In today's world when infrastructure building expands rapidly, and data collection and technology improves in even faster pace, the ability to integrate the most recent and effective data is an important skill to master and pass on to NSOs and researchers.

Secondly, since the Geo-Covariate variables are provided as gridded data that covers the whole country, the data can be extracted for any given location. We can calculate the model prediction on any geo-location in the country. This means we can extend our modeling out of the DHS sample location to the whole country. For consistency, it is much better we build the model on the data obtained in the same way.

For many surveys, DHS also have community survey questionnaire that is answered by knowledgeable person from the PSU. The following 10 variables are from this dataset (the BDSQ71FL folder):

1) MainRoad: Main access road to this village (All weather road, Seasonal road, Waterway, Path, Other, Do not know, Missing)
2) DMarket: Distance to the weekly market
3) DPS: Distance to primary school
4) DGHS: Distance to Girl's high school
5) DBHS: Distance to Boy's high school
6) DHealthN: Distance to health facility
7) Grameen: Grameen Bank member
8) MotherClub: Mother's club or ladies associations
9) Cooperative: Cooperative society

10) Residence: Urban vs Rural

If this set of information is useful in improving the model result, effort to integrate this information from National database will be worth it. We explore this area in our study. With this in mind, we carried out comparison on variable sets, presented in Table 3.

Table 3. Variable sets

Short name for set of variables	Explanation
GC	Geo-covariates by DHS
GCUD	Geo-covariates updated by raster files from internet, excluding the three variables on poverty and income from World Bank
GCUDALL	Geo-covariates updated by raster files from internet, including variables from World Bank.
SQ	Information on service for PSU from DHS community survey data.
SQGC	Combination of GC and SQ variables
SQGCUD	Combination of GCUD and SQ variables

The goal of model comparison is to find out which statistical models and which set of variables works well to predict the response variables.

5) Adding individual characteristics and disaggregate data

Once we have chosen the model and geo covariates set to use, we have built a framework to explain the variation of the indicator values on cluster level. Now, adding household/individual characteristics, such as sex, education of household head, and sex, education and age of individuals into the model we have chosen from previous section, will help us see the disaggregation of some key cohorts.
We use the model based on geo-covariates as a base line, calculated two R-square type of statistics, (R_t^2, R_c^2) =), each represent the R-square of the model at total and cluster level.

$SD_t^2(Model) = \frac{\sum_{ji}(\hat{Y}_{ji}(model)-\bar{Y})^2}{n}$, where $\hat{Y}_{ji}(model)$ is the predicted value from the model.

And

$SD_c^2(Model) = \frac{\sum_{i}(\hat{Y}_{ji}(model)-\acute{Y})^2}{m}$, when the model input variables are all cluster level, $\hat{Y}_{ji}(model)$ is the same for all households/individuals in the same cluster.

Then

$$R_t^2(model) = \frac{SD_t^2(model)}{SD_t^2},$$

And

$$R_c^2(model) = \frac{SD_c^2(model)}{SD_c^2}.$$

But when individual/household characteristics are added to the model input, recalling our model on the indicator value Y_{ji} has cluster component ($e_i(s_i)$) and individual component (p_{ji})

$$Y_{ji}=\alpha+e_i(s_i)+p_{ji}$$

Then we can include sex, education, age as part of p_{ji} of the model one at a time, and calculate R^2(model+Sex), R^2(model+Age), R^2(model+Education). Comparing the new R^2-s with R^2(model) can tell us if each of the newly added individual characteristic makes difference. When the new R^2 has big increase compared to R^2(model), it means households (individuals) of different cohorts have different levels of indicator values.

6) Bayesian Geo-statistical model

"*Models of point-referenced data that include a spatially-structured random effect are commonly called geostatistical models. Geostatistics is the specific area of spatial statistics that studies these models.*"[7]

When we model the cluster effect based on the geo-covariates, we make an implicit assumption that all the spatial effect on the indicator value take the form:

$$e_i(s_i) = F(GC_i) + \in_i$$

Where $\in_i$ are random errors, independently and identically distributed. But this assumption can be challenged, due to the spatial position of the clusters. It is reasonable to assume that clusters that are close by are more related than clusters that are far away. $\in_i$ might not be independent of each other.
A very common assumption on the distribution of $\in_i$ is that it is a random effect with a multivariate Gaussian distribution, called Gaussian Field. In this distribution, an assumption on conditional distribution makes the correlation matrix of the multivariate Gaussian distribution much more manageable, and it is called the Gaussian Markov Random Field. This computable model becomes popular in geostatistical modeling. This type of random field is linked to stochastic partial differential equation (SPDE).

[7] See [Cressie, 1993] for a good introduction to spatial statistics**.**

For multi-variate Gaussian distribution, the covariance matrix describes all the dependence structure of $\in_i$. Here we choose the simplest form, called the Matérn covariance function. We represent the geo-statistical model with the following notation, assuming we observe Y_{ji} on location S_i:

$$Y_{ji}|s_i \sim N$$

Where σ_e is the dispersion parameter (standard deviation under the Gaussian distribution assumption) of the assumed conditional distribution of Y_{ji}.
The spatial dependence of Y_{ji} is modeled in $\in_i$ which is a random effect following the Gaussian Field.

$$\in_i \ GF(0, \Sigma)$$

In many situations we assume that we have an underlying Gaussian Field but cannot directly observe it and instead observe data with measurement error, i.e., $Y(s_i)=e(s_i)+\in_i$.

For a Gaussian Field, the joint distribution of the random effect is determined by its correlation matrix, Σ. Naturally, we assume stronger spatial dependence (correlation coefficients) when the points are spatially close and weaker dependence when they are far way. The stationary and isotropic Matérn correlation function fits the requirement and is a popular choice in geo-statistical model:

$$Cor_M\left(\in (s_i), \in (s_j)\right) = \frac{2^{(1-\upsilon)}}{\Gamma(\upsilon)}\left(\kappa\|s_i - s_j\|\right)^{\upsilon} K_{\upsilon}\left(\kappa\|s_i - s_j\|\right)$$

The Matérn covariance function is $\sigma_{\in}\ Cor_M$ ($\in (s_i)$,$\in (s_j)$), where $\sigma_{\in}$ is the marginal variance of the process. The parameters υ is normally determined by alpha, the smoothness constant in the corresponding SPDE. For our project, we choose alpha to be 2, corresponding to maximum smoothness. κ can be estimated and it represents the strength of spatial dependence. ρ=1/κ is called the range parameter, it is the distance where correlation between two locations decline to 10%. The bigger κ or the smaller ρ values indicate a weak spatial dependence. When the spatial dependence is weak, it is reasonable to abandon the geo-spatial models and keep the naïve predictive model we build previously. When the spatial dependence is strong, the model prediction improves with geo-statistical models.

In recent years, the technology and demand on modeling spatial data has resulted workable software. The R-INLA package provides actual computation tools to estimate the spatial dependence via the SPDE approach.

D. Results and conclusions of Bangladesh data integration

To try out the method with Bangladesh data, we randomly chose 65% (390 out of 600) of the clusters to the training set and 35% (210 out of 600) to the testing set. The training sets are used to build the models. Afterwards we apply the model to the testing set. We can then calculate $RMSE_t(model)$ and $RMSE_c(model)$ on both training data and testing data.

We build the four models (GLM, Tree, Random Forest, XGBoost), using the training set, based on different variable sets (GC, GCUD, GCUDALL, SQGC, SQGCUD). We also apply each model on the testing set as cross validation. At the end, we want to identify the model and variable set that gives reasonable predictive results and are easy to use and update.

So, to illustrate the four different model results, we choose Clean Fuel as the response variable (Table 4). For other response variables, the results are presented in the appendix.

Table 4: Model comparison ---Root Mean Squared Error of four models (GLM, Tree, Random Forest, XgBoost) on six sets of variables.

CleanFuel-ALL		**SD_t**	**SD_c**							
		36.36%	31.38%							
Short name for set of variables	Group	$RMSE_t$ (GLM)	$RMSE_c$ (GLM)	$RMSE_t$ (Tree)	$RMSE_c$ (Tree)	$RMSE_t$ (Random Forest)	$RMSE_c$ (Random Forest)	$RMSE_c$ (pooled, Random Forest)	$RMSE_t$ (XGBoost)	$RMSE_c$ (XGGoost)
GC	Training	26.27%	17.94%	22.84%	12.29%	22.33%	10.92%		25.16%	16.37%
	Testing	25.69%	17.40%	26.94%	19.18%	25.04%	16.42%	13.11%	26.05%	17.94%
GCUD	Training	25.78%	17.20%	22.51%	11.73%	22.19%	10.65%		24.95%	16.05%
	Testing	25.75%	17.48%	29.29%	22.36%	25.76%	17.48%	13.44%	26.68%	18.85%
GCUDALL	Training	25.41%	16.61%	22.50%	11.70%	21.86%	9.86%		24.84%	15.88%
	Testing	25.69%	17.39%	29.30%	22.37%	25.61%	17.26%	12.94%	26.43%	18.50%
SQGCUD	Training	25.42%	16.64%	22.51%	11.73%	21.99%	10.20%		24.88%	15.94%
	Testing	25.61%	17.28%	29.29%	22.36%	25.40%	16.94%	12.96%	26.49%	18.59%
SQGC	Training	25.93%	17.43%	22.47%	11.58%	22.03%	10.26%		25.01%	16.14%
	Testing	25.78%	17.54%	27.36%	19.77%	24.87%	16.15%	12.64%	26.01%	17.87%
SQGCUDALL	Training	25.12%	16.17%	22.50%	11.70%	21.61%	9.28%		24.73%	15.70%
	Testing	25.57%	17.22%	29.30%	22.37%	25.37%	16.90%	12.49%	26.19%	18.16%

The lower the two statistics ($RMSE_t(model)$ and $RMSE_c(model)$) are for both the training set and testing set, the better the model. We can see that Random Forest usually delivers the lowest values on these statistics for all variable sets. So, we focus on the two columns of RMSE (random forest), among the six variable sets (combinations), SQGC and SQGCUDALL have lower cluster level RMSE for testing and training data respectively. Since training and testing data have around 65% and 35% of the sample clusters, we pooled the RMSE together with the calculation: $RMSE^2$(pooled) = $RMSE^2$(training)*0.65 + $RMSE^2$(testing)*0.35. SQGCUDALL has the lowest pooled RMSE. This should be our choice for predicting use of clean fuel at cluster level. However, the 10 SQ variables are hard to get and we are interested in testing whether we can use GCUDALL variables and obtain results statistically acceptable. Since GCUDALL is a subset of SQGCUDALL, we can conveniently use an F-test on comparing nested models.

$$F = \frac{\frac{SSE(SQGCUDALL) - SSE(GCUDALL)}{10}}{MSE(SQGCUDALL)} = 4.28,$$

Here 10 is the number of variables in SQ. F has first degree of freedom of 10, and second degree of freedom of 577(600-23). F-value for α=0.01 is 2.35. The model comparison F test indicates that SQGCUDALL is indeed a better model than GCUDALL.

Because of the difficulty of obtaining nationwide service data for every GIS location, even when SQGCUDALL produces the best prediction results, we might still use GCUDALL to predict the indicator value because GCUDALL variables are easily available. But for government agencies that have the service data already, adding SQ variables is desirable and doable.

A summary on all the indicators in this study are presented in Table 5. [8]

Table 5: Variable set comparison: best variable set vs. GCUDALL

Indicator	**Best variable Set**	**Pooled RMSE** (Random Forest)	**Pooled RMSE** (Random Forest) for GCUDALL	**F-value**	**Conclusion**
Poverty-ALL	SQ GCUDALL	13.07%	13.28%	1.975	GCUDALL is as good
AccessElectricity-ALL	SQ GCUDALL	15.95%	16.21%	1.859	GCUDALL is as good
SafeSanitation-ALL	SQ GC	17.53%	17.80%	2.627	SQGC is better
BankCard-ALL	SQ GCUDALL	13.52%	13.94%	3.667	SQGCUDALL is better
CleanFuel-ALL	SQ GCUDALL	12.49%	12.94%	4.283	SQGCUDALL is better
CleanWater-ALL	SQ GC	7.58%	7.85%	6.134	SQGC is better
SchoolAgeEducation-LowerSecondaryAge	SQ GCUDALL	14.54%	14.67%	1.057	GCUDALL is as good
SchoolAgeEducation-UpperSecondaryAge	GCUDALL	20.81%			GCUDALL is as good
SchoolAgeEducation-CollegeAge	SQ GCUDALL	16.79%	16.98%	1.262	GCUDALL is as good
AdultEducation-Older	SQ GCUDALL	8.52%	9.07%	7.755	SQGCUDALL is better
AdultEducation-Young	SQ GCUDALL	10.96%	11.36%	4.290	SQGCUDALL is better
CurrentlyWorking-WorkingAge	SQ GCUDALL	9.17%	9.22%	0.634	GCUDALL is as good
Stunt-Under5Age	SQ GCUDALL	16.03%	16.09%	0.429	GCUDALL is as good
Underweight-Under5Age	SQ GCUDALL	15.55%	15.64%	0.659	GCUDALL is as good

[8] Detailed model results are in supplement docs file1.

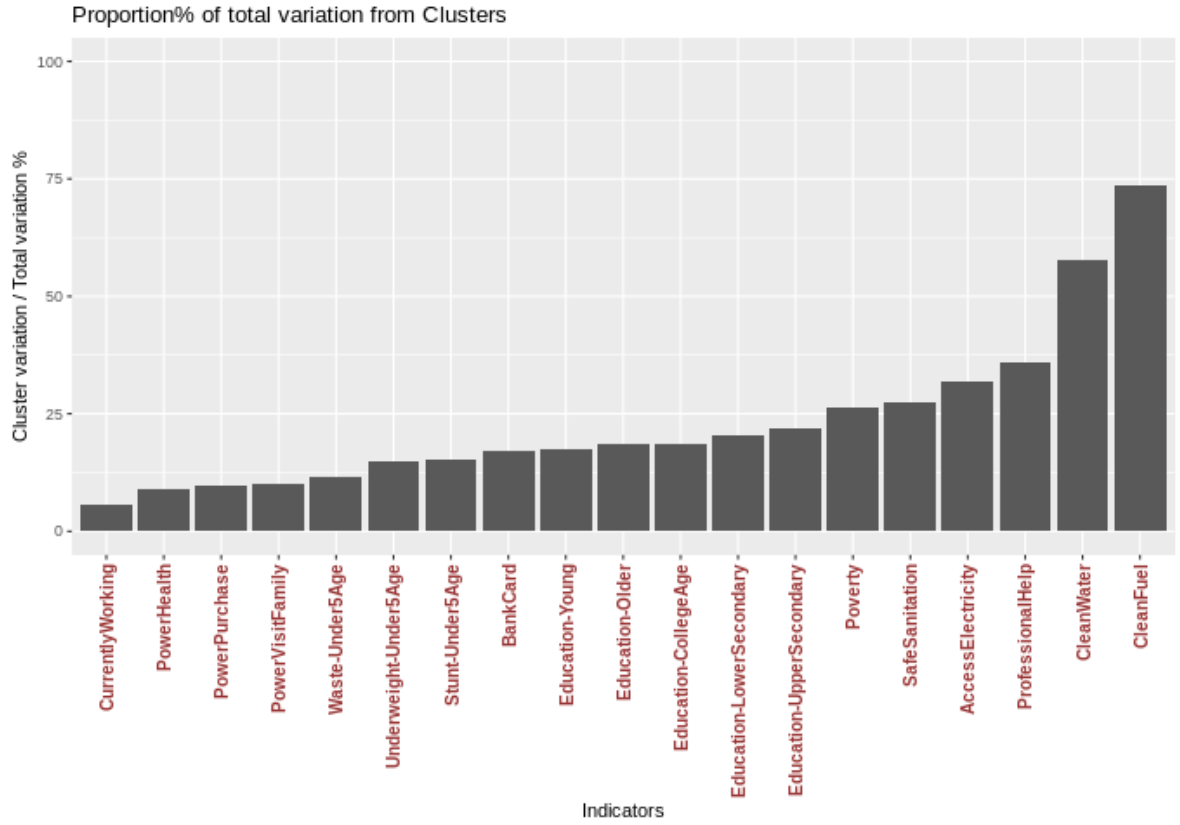
Indicator	Best variable Set	Pooled RMSE (Random Forest)	Pooled RMSE (Random Forest) for GCUDALL	F-value	Conclusion
Waste-Under5Age	SQ GCUDALL	10.74%	10.77%	0.388	GCUDALL is as good
PowerHealth-CurrenlyMarried	SQ GCUDALL	11.84%	12.01%	1.696	GCUDALL is as good
PowerPurchase-CurrenlyMarried	SQ GC	12.22%	12.34%	1.617	GCUDALL is as good
PowerVisitFamily-CurrenlyMarried	SQ GC	12.18%	12.32%	1.900	GCUDALL is as good
ProfessionalHelp-delivered	SQ GCUDALL	22.90%	23.12%	1.132	GCUDALL is as good

Now with chosen variable set (GCUDALL) and predictive model, we can attempt to answer questions one to three in our background section.

Table 6: Indicators variation pattern

Indicator	SD_t	SD_c	$\left(\frac{SD_c}{SD_t}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$
Poverty-ALL	38.08%	19.56%	26.39%	13.37%	53.91%
AccessElectricity-ALL	44.20%	25.02%	32.04%	17.86%	58.05%
SafeSanitation-ALL	45.95%	24.12%	27.55%	11.75%	45.50%
BankCard-ALL	46.53%	19.32%	17.24%	7.29%	47.93%
CleanFuel-ALL	36.06%	30.92%	73.55%	58.45%	82.48%
CleanWater-ALL	15.16%	11.54%	57.87%	29.91%	53.65%
SchoolAgeEducation-LowerSecondaryAge	37.57%	16.97%	20.41%	3.50%	25.29%
SchoolAgeEducation-UpperSecondaryAge	49.35%	23.15%	22.00%	2.30%	19.20%
SchoolAgeEducation-CollegeAge	44.59%	19.17%	18.48%	2.24%	21.55%
AdultEducation-Older	30.37%	13.04%	18.43%	9.59%	51.55%
AdultEducation-Young	36.62%	15.30%	17.47%	6.48%	44.88%
CurrentlyWorking-WorkingAge	49.39%	11.62%	5.54%	1.74%	36.99%
Stunt-Under5Age	48.14%	18.90%	15.42%	2.76%	27.57%
Underweight-Under5Age	46.84%	18.11%	14.95%	1.98%	25.44%
Waste-Under5Age	35.25%	11.96%	11.50%	0.17%	18.84%
PowerHealth-CurrenlyMarried	48.22%	14.52%	9.07%	2.07%	31.57%
PowerPurchase-CurrenlyMarried	49.04%	15.35%	9.80%	2.67%	35.41%
PowerVisitFamily-CurrenlyMarried	48.73%	15.36%	9.94%	2.78%	35.75%
ProfessionalHelp-delivered	49.68%	29.74%	35.83%	11.99%	39.55%

Here, the column $\left(\frac{SD_c}{SD_t}\right)^2$ represents the proportion of total variance that varies at cluster level. We see that it varies greatly from indicator to indicator. The indicator that varies the least with the geo-location is Currently Working (at 5.54%). Indicator reflecting women's power at home (whether they can decide to go visit family, doctors, shopping for themselves) also varies little with geo-location (under 10%) as well as "Nutritional status of children under 5 years old" (around 15%). The indicator that varies the most is Clean Fuel (at 73.55%), followed by clean water (at 57.78%), access to electricity, and maternal healthcare (getting professional help at birth delivery).

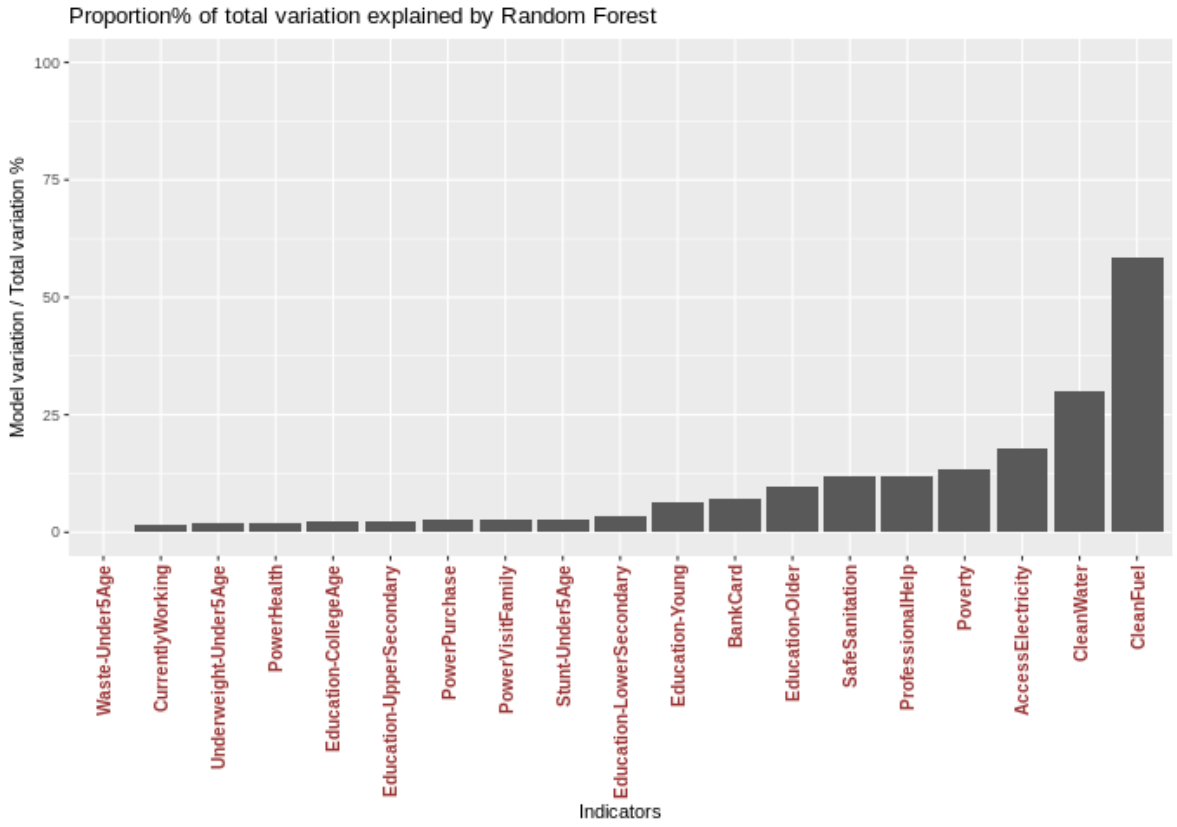


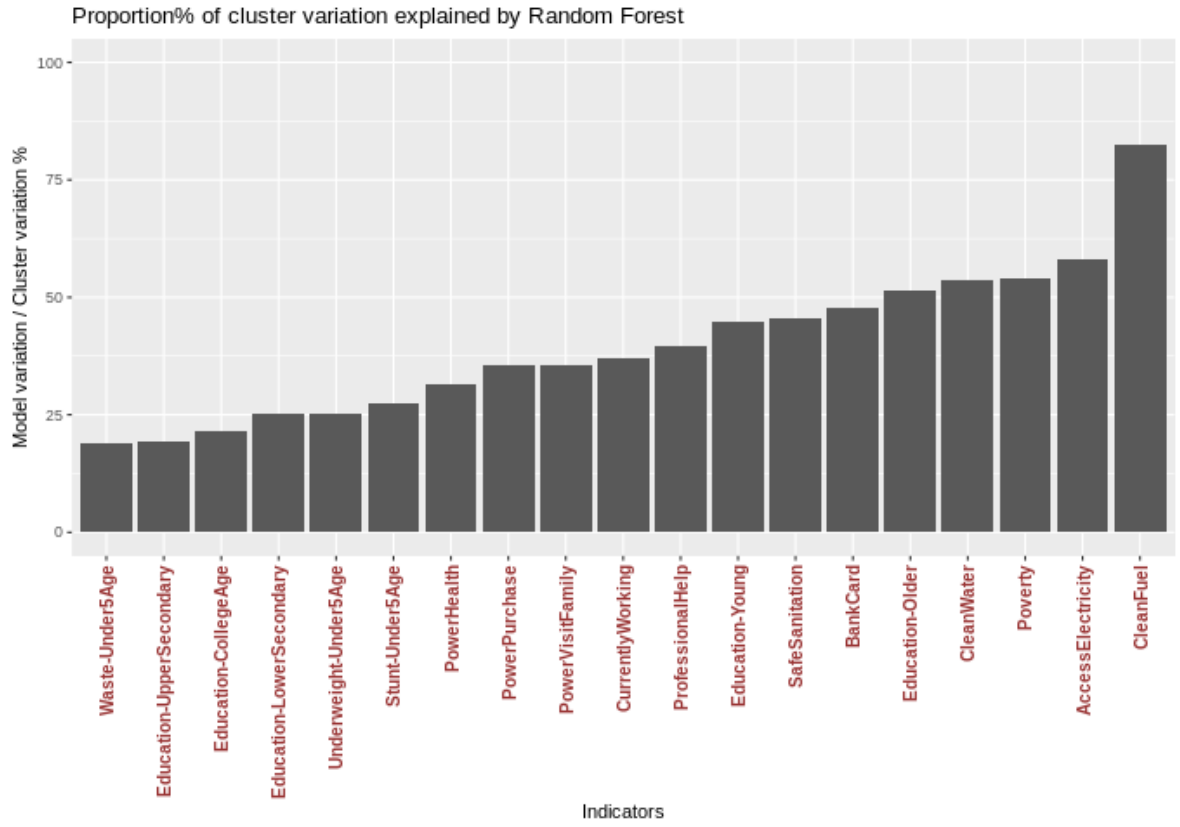
The Random Forest model prediction (with all geo-covariates at cluster level) can explain part of the cluster variation. And this leads to the model explaining part of the total variation, while the proportion of the total variation explained is limited by the proportion of total variation that is accounted by cluster variation, as we observed above, different indicators vary on this measure. A few observations:

1) So, it is natural and unsurprising that when the cluster level variation is low, the model explained very little of the total variation. Here three categories of indicators stand out as having very low variation explained by the model (around 2% or less): education for school aged children/youth, nutrition status of 5-year-old and younger, power to make certain decisions at home for married women. Currently working for working age population also has very low cluster level variation and the geo-covariate model explains very little. The low-level cluster variation, and the weak relation to geo-covariates might reveal some underlying social dynamics of Bangladesh.

2) Even though adult education (young and old), having bank card both have low cluster variation, the model explains half of the cluster variation. It indicates some influence of the geo-covariates.

3) Cluster level variation of household characteristics has good model explanation (around 50% or more). Clean Cooking Fuel stands out as having not only high cluster level variation, but also high model prediction power. 58% of household choice on clean cooking fuel can be explained by the geo-covariates model.



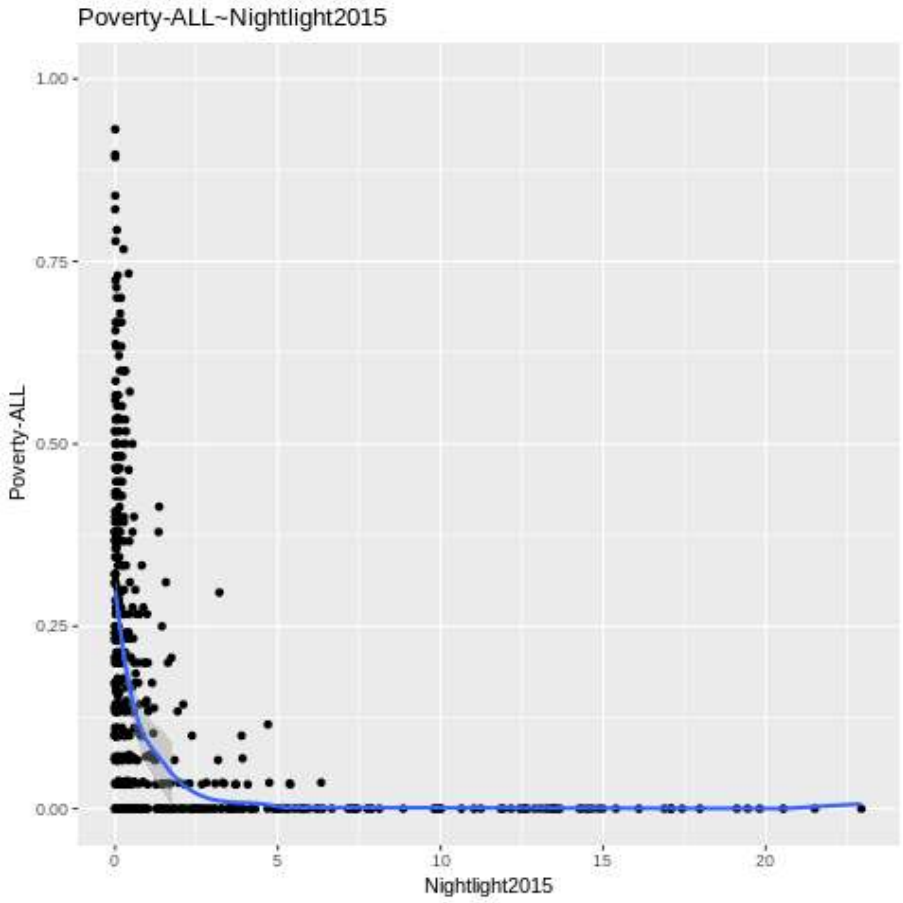


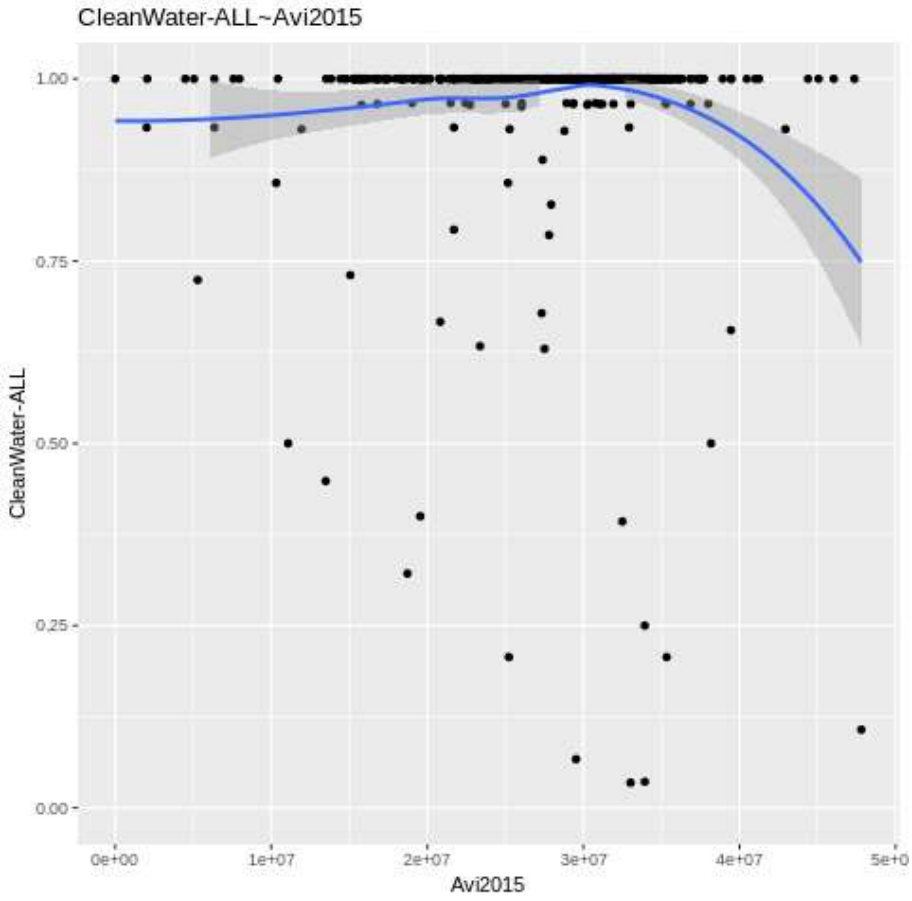
The Random Forest model allows us to see the importance of the geo-covariate, here we summarize the top three variables for all indicators for GCUDALL and GCUD, where “1 variable” has the most importance:

Table 7: Most important variables in random forest model for each indicator by two variable sets

Indicator	GCUDALL			GCUD		
	3rd variable	2nd variable	1st variable	3rd variable	2nd variable	1st variable
Poverty-ALL	Travel_Times2015	SMOD2015	aWealthIndex2011	Aridity2000	Avi2015	Nightlight2015
CleanWater-ALL	aPP2013	Friction2015	Avi2015	Friction2015	Aridity2000	Avi2015
SafeSanitation-ALL	Aridity2000	Travel_Times2015	aWealthIndex2011	Hfp2004	Aridity2000	Travel_Times2015
CleanFuel-ALL	Nightlight2015	Avi2015	aIncome2013	Density2015	Nightlight2015	Aridity2000
AccessElectricity-ALL	Aridity2000	Nightlight2015	aWealthIndex2011	Aridity2000	Travel_Times2015	Nightlight2015
BankCard-ALL	aPP2013	Travel_Times2015	aIncome2013	DroughtEpisode	SMOD2015	Travel_Times2015
SchoolAgeEducation-LowerSecondaryAge	Travel_Times2015	Aridity2000	Avi2015	Nightlight2015	Avi2015	Aridity2000
SchoolAgeEducation-UpperSecondaryAge	Avi2015	Aridity2000	Travel_Times2015	Travel_Times2015	Avi2015	Aridity2000
SchoolAgeEducation-CollegeAge	Density2015	Avi2015	aIncome2013	Density2015	Avi2015	Aridity2000
AdultEducation-Young	Avi2015	aIncome2013	Density2015	Aridity2000	Density2015	Hfp2004
AdultEducation-Older	Aridity2000	aIncome2013	Density2015	Nightlight2015	Avi2015	Density2015
CurrentlyWorking-WorkingAge	aIncome2013	Aridity2000	aPP2013	Avi2015	DroughtEpisode	Aridity2000
Stunt-Under5Age	aPP2013	Avi2015	Aridity2000	Travel_Times2015	SMOD2015	Aridity2000
Underweight-Under5Age	aWealthIndex2011	aPP2013	Aridity2000	Nightlight2015	Travel_Times2015	Aridity2000
Waste-Under5Age	Nightlight2015	aWealthIndex2011	aPP2013	Friction2015	Hfp2004	Buildup2015
ProfessionalHelp-delivered	Buildup2015	aWealthIndex2011	Aridity2000	Travel_Times2015	Nightlight2015	Aridity2000
PowerVisitFamily-CurrenlyMarried	SMOD2015	Hfp2004	Aridity2000	Friction2015	Hfp2004	Aridity2000
PowerPurchase-CurrenlyMarried	aPP2013	Hfp2004	Aridity2000	Avi2015	Hfp2004	Aridity2000
PowerHealth-CurrenlyMarried	aIncome2013	aPP2013	Aridity2000	Hfp2004	Avi2015	Aridity2000

Here are two scatter plots that shows how the most important geo-covariates relate to the indicator values at cluster level.





For each indicator, a scatter plot of indicator ~ most important geo-covariate is provided in a supplement doc file 2.

The reason we have two sets of results is that once we include the World Bank income/poverty estimates, they consistently show up as the most important variables. These variables are powerful but may not always be available for other countries. So, we want to look at the influence of other geo-covariates in the GCUD variable sets, where we see Aridity2000 consistently show up for almost all indicators.

Now we examine the individual/household characteristics in the same framework, by adding sex, Age, Education to the random forest model. Table 8 tells us how adding these variables (one at a time) help with the predictive power.

Table 8: Effect of individual characteristics such as sex, education and age in predicting indicator level

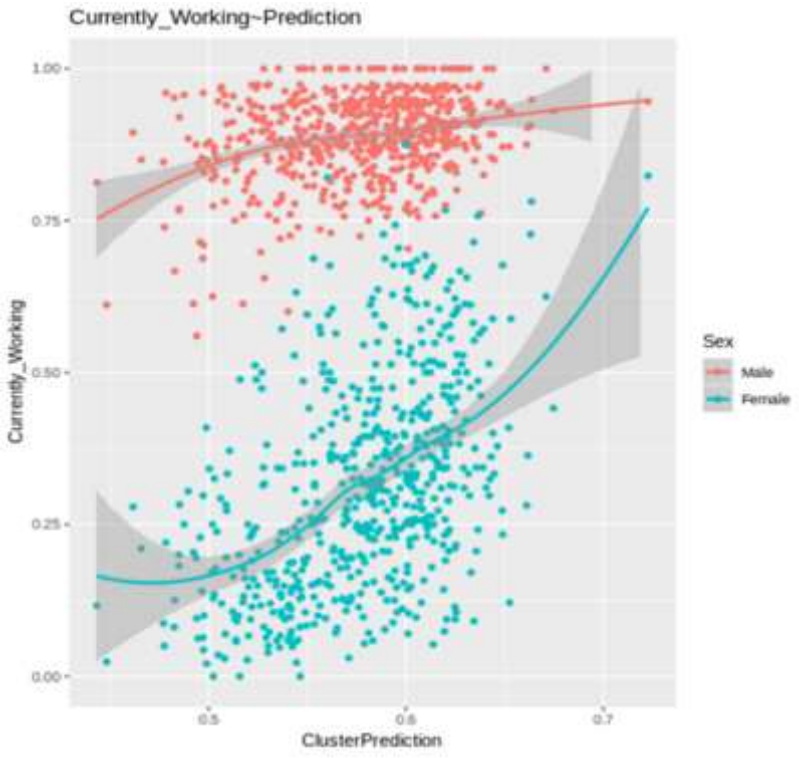
Indicator	GCUDALL		GCUDALL + Sex		GCUDALL + Education		GCUDALL + Age	
	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$
Poverty-ALL	13.66%	52.41%	13.55%	51.99%	15.21%	52.35%		
CleanWater-ALL	30.36%	54.61%	29.55%	53.22%	29.07%	52.46%		
SafeSanitation-ALL	11.85%	45.83%	11.76%	45.40%	14.28%	46.21%		
CleanFuel-ALL	58.41%	82.42%	58.26%	82.24%	59.23%	82.46%		
AccessElectricity-ALL	34.31%	63.43%	34.18%	63.18%	36.12%	63.73%		
BankCard-ALL	7.33%	48.16%	7.46%	47.45%	17.73%	61.66%		
SchoolAgeEducation-LowerSecondaryAge	3.46%	24.94%	4.80%	22.19%				
SchoolAgeEducation-UpperSecondaryAge	2.61%	21.37%	2.75%	18.23%				
SchoolAgeEducation-CollegeAge	2.47%	21.77%	3.93%	19.09%				
AdultEducation-Young	6.48%	44.44%	6.59%	43.62%				
AdultEducation-Older	9.51%	51.06%	11.62%	46.28%				
CurrentlyWorking-WorkingAge	1.72%	36.75%	33.09%	37.30%	2.13%	34.74%	5.54%	28.63%
Stunt-Under5Age	2.80%	27.40%	2.73%	26.69%	4.75%	28.67%		
Underweight-Under5Age	1.98%	25.33%	2.01%	24.96%	3.38%	25.54%		
Waste-Under5Age	0.19%	18.66%	0.18%	17.82%	0.17%	17.27%		
ProfessionalHelp-delivered	11.93%	39.29%			17.14%	43.38%	11.50%	37.52%
PowerHealth-CurrenlyMarried	2.13%	32.32%			2.12%	31.66%	3.71%	27.03%
PowerPurchase-CurrenlyMarried	2.62%	35.20%			2.74%	34.94%	5.54%	30.95%
PowerVisitFamily-CurrenlyMarried	2.72%	35.25%			2.81%	34.86%	5.43%	32.44%

Here we see that Sex plays a dominating role in people's working opportunities. This indicates a very different pattern of working opportunity for men and women. So, we disaggregate the data by sex and rerun the model separately.

Table 9: Different models for currently working for male and female

	GCUDALL		GCUDALL + Education		GCUDALL + Age	
Indicator	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$	$\left(\frac{SD_t(RF)}{SD_t}\right)^2$	$\left(\frac{SD_c(RF)}{SD_c}\right)^2$
CurrentlyWorking-WorkingAge-Female	4.12%	32.15%	4.51%	29.92%	5.36%	26.86%
CurrentlyWorking-WorkingAge-Male	0.58%	25.59%	7.32%	30.65%	14.43%	20.37%

Here we see different patterns for men and women. Women's working opportunity depends on geo-covariates more than men. But Education and Age have little effects, while for men, education and age make sizable difference. Below is scatter plot where for each cluster, female and male cohort working % are disaggregated. X-axis is working % predicted at cluster level (using GC variables), Y-axis is actual male (red dots) and female (blue dots) for the cluster. Male and female from the same cluster share the same x-values, and we can see that their y-values are so different that they have very little overlap. For every cluster, male working % is substantially higher than female working %. This is the image when R-square of GCUDALL+Sex (33%) is much higher than R-square of GCUDALL alone (1.7%).



1) Key observations

Here, we point out some of the key observations of our study:

- Adding individual level variable to cluster level variables normally does not improve the explained variation at cluster level, $\left(\frac{SD_c(RF)}{SD_c}\right)^2$, but could improve the explained variation at total level, $\left(\frac{SD_t(RF)}{SD_t}\right)^2$. This is especially true when the individual characteristic relatively correlates little with the geo-covariates, like, for example, sex, age.

- Highest education level in household plays some limited roles in poverty, sanitation, access to electricity, and a more sizable role in using bank card.

- Mother's education plays limited roles in the nutrition of the children, as we can see that the R-squared has slight increase for underweight and stunting.

- The women's education plays a sizable role in maternal health.

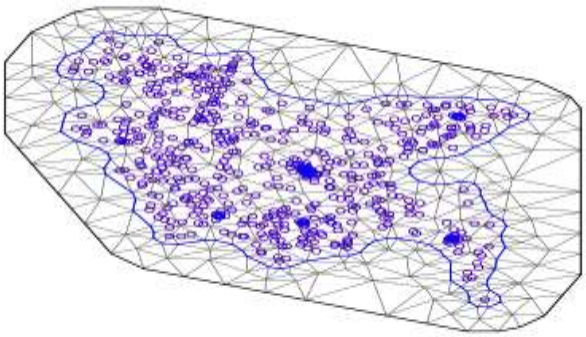
- Age plays some role in working opportunity, and in women's power to make decision at home. Notably, women's education has little effect on her power to make decisions at home. (Age is a much more important factor in social norm when treating women.)

- Sex plays roles in education and working opportunities. While the gain in education is small, between 1-2%, the gain in working opportunity is huge, at 30%. This indicates a very different pattern of working opportunity for men and women.

2) Bayesian Geo-statistical model

(Geo-Statistical model is very commonly used in this area when geo-covariates are major inputs of the model. The main difference between Geo-statistical model and predictive model like random forest models is that the geo-statistical model assumes that in addition to the influence of the Geo-covariates, neighboring clusters tend to have association with each other because of the proximity. It is natural to assume that the association depends on the distance between the clusters and it gets weakened as the distance grows. Here, we test if this association really exists in the DHS sample data.

GeoSpatial model is estimated using integrated nested Laplace approximations (INLA) method, combined with stochastic partial differential equation (SPDE). The SPDE approach approximates the Gaussian field using a Gaussian Markov random field. For geostatistical data, the spatial field is described using a weighted sum of piece-wise linear basis function that is usually defined from a triangularization of the study area, called the mesh. R-INLA package provides functions to do that.



Each blue point on the mesh is a cluster.
We ran the INLA model with alpha=2 in three different setting:

1) INLA model without geo-covariates, to see the spatial correlations.

2) INLA model with predictions from Random Forest based on geo-covariates, without spatial correlations.

3) INLA model with predictions from Random Forest based on geo-covariates, and spatial correlations.

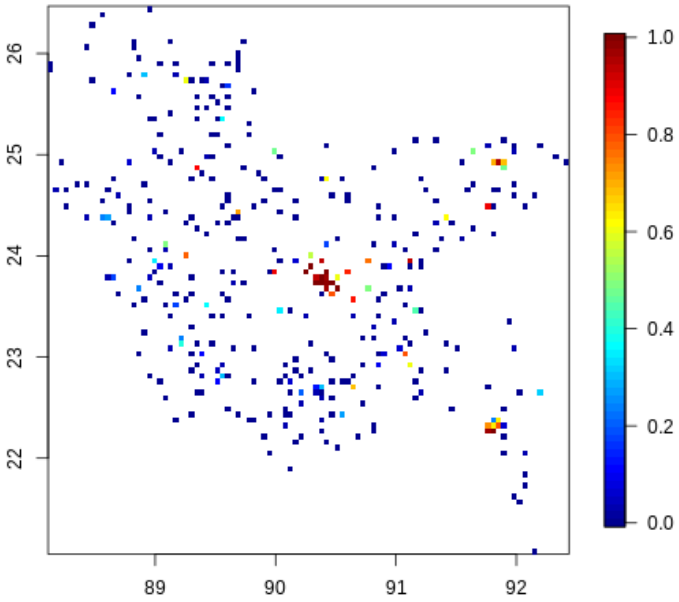
To check whether this association really exists in the DHS sample data, we compare the SE term. If SE changes little after geo-spatial term is included in the model, we can conclude that there is no/weak spatial correlation. In other words, if there is spatial relation, we will see standard error in (1) smaller than original standard error without the modeling. We should also see that standard error in (3) smaller than that of (2).

Table 10: Standard Error for Bayesian Geo-statistical models

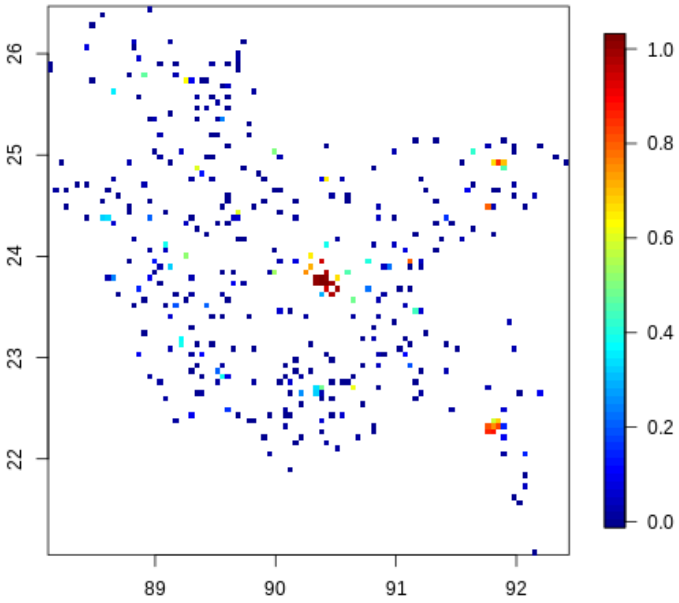
Indicator	SD	SD with Spatial Modeling	SD with GC prediction	SD with GC prediction and Spatial Modeling
Poverty-ALL	19.5%	19.5%	13.4%	13.4%
CleanWater-ALL	11.6%	11.5%	7.5%	7.5%
SafeSanitation-ALL	24.1%	24.1%	17.6%	17.6%
BankCard-ALL	19.4%	19.3%	13.8%	13.8%
CleanFuel-ALL	30.9%	30.9%	12.9%	12.9%
AccessElectricity-ALL	25.0%	25.0%	16.4%	16.4%
SchoolAgeEducation-LowerSecondaryAge	16.9%	16.9%	13.6%	13.6%
SchoolAgeEducation-UpperSecondaryAge	23.1%	23.1%	19.9%	20.0%
SchoolAgeEducation-CollegeAge	19.2%	19.1%	15.6%	15.6%
AdultEducation-Young	15.3%	15.3%	10.6%	10.6%
AdultEducation-Older	13.1%	13.0%	8.6%	8.6%
CurrentlyWorking-WorkingAge	11.6%	11.6%	9.1%	9.1%
Stunt-Under5Age	18.9%	18.9%	15.3%	15.3%
Underweight-Under5Age	18.1%	18.1%	14.9%	14.9%
Waste-Under5Age	11.9%	11.9%	9.7%	9.7%
Poverty-WorkingAge	18.5%	18.4%	12.6%	12.6%
ProfessionalHelp-delivered	29.8%	29.7%	22.2%	22.2%
PowerVisitFamily-CurrenlyMarried	15.4%	15.3%	11.9%	11.9%
PowerPurchase-CurrenlyMarried	15.4%	15.4%	12.0%	12.0%
PowerHealth-CurrenlyMarried	14.6%	14.5%	11.4%	11.4%
Clean Fuel Urban	39.1%	38.9%	18.3%	18.3%
Clean Fuel Rural	14.3%	14.3%	8.7%	8.7%

From this table, we conclude that there is no spatial correlation for clusters for any of the indicator in this study. To illustrate the results, images for the variable Clean Fuel are presented below.

First, the image plot of the cluster value:



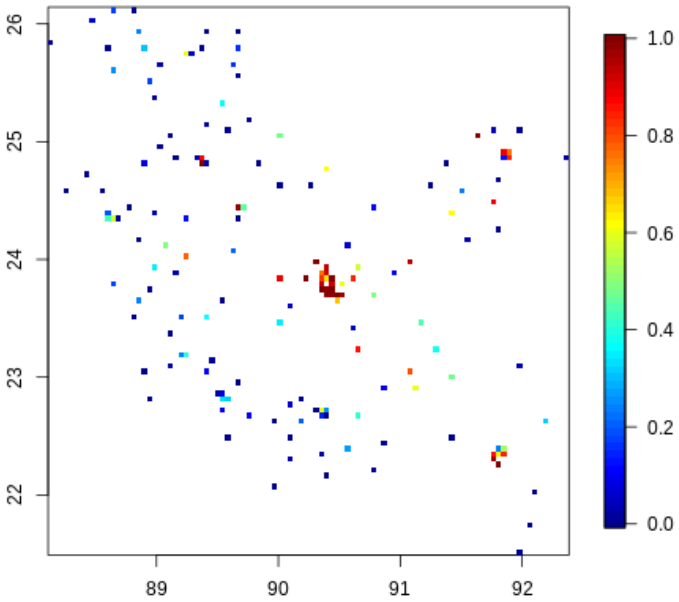
Second, the image of predicted cluster value:



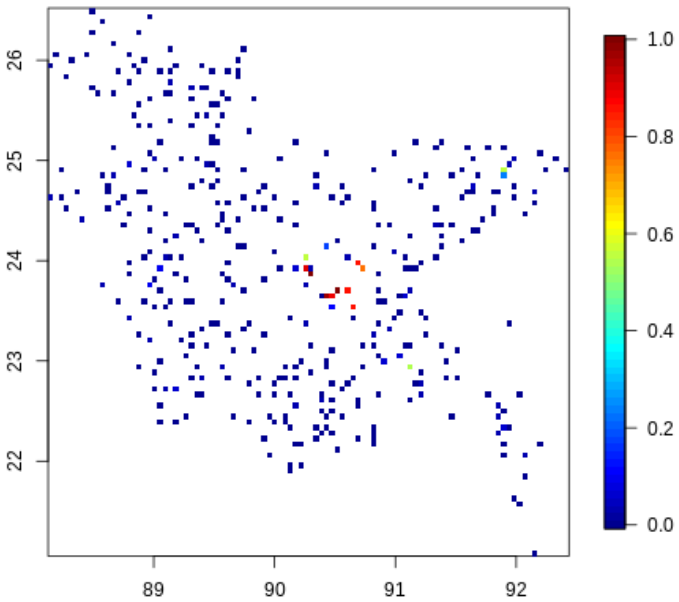
We can see that predictive model based on geo-covariates captures the pattern of usage of clean cooking fuel well.

We also tried the spatial modeling for urban clusters and rural clusters separately, but there is still no spatial correlation. The graphs are as below.

Urban:


26
25
24
23
22
89
90
91
92
1.0
0.8
0.6
0.4
0.2
0.0

Rural:


26
25
24
23
22
89
90
91
92
1.0
0.8
0.6
0.4
0.2
0.0

E. Summary and further research

We used Bangladesh as an example to illustrate that

1) we can integrate the most up to date geo-information created by a body of data providers and research communities to DHS (or any survey data) with GIS information to enhance our understanding about how geo-location and related information impacts on indicators interested.

2) By looking at the proportion of variations from cluster level, and the proportion that can be explained by the random forest method with geo-covariate input, we can see indicators are not influenced by the same degrees by the geo-covariates. And in fact, other research paper shows that this variation structure depends on countries (ref 10). Bangladesh has good distribution of educational and health facilities across different geographic areas, and with NGOs like the Grameen Banks in almost every village providing business assistance to women, households access to bank card is widely spread in villages. Children's nutrition needs is still very challenging, but at least we know that this phenomenon is not highly dependent on geo-locations (very small proportion of variation on children's nutrition can be explained by geo-covariate model). Feeding and cooking habits in individual households seems to be more responsible. While African countries in other research papers show much higher correlations with geo-location (ref 10).

3) Some of the geo-covariates seemed very out of date. Aridity is a statistics that changes rapidly in current climate change environment, while Bangladesh is particularly hit. An update will be great appreciated.

4) Travel time to the city is defined as travel time to a city center of 50,000 habitats or more. In Asia, with its huge population everywhere, this does not mean a city with promising economic or cultural opportunity. New calculations too much larger city center might have potential to explain more of geo-diversity.

Reference

1- Moreover, the DHS programme is now routinely providing modelled surfaces with each new country survey produced through spatial interpolation (http://spatialdata.dhsprogram.com/modeled-surfaces/) ---- Bangladesh 2014 survey was used and indicators on children stunting, women contraception, water and sanitation are included

2- DHS Spatial Analysis Reports No. 9 Spatial Interpolation with Demographic and Health Survey Data: Key Considerations (https://dhsprogram.com/pubs/pdf/SAR9/SAR9.pdf)

3- DHS Spatial Analysis Reports No. 14 Guidance for Use of The DHS Program Modeled Map Surfaces (https://dhsprogram.com/pubs/pdf/SAR14/SAR14.pdf)

4- DHS Spatial Analysis Reports No. 11, Creating Spatial Interpolation Surfaces with DHS Data (https://dhsprogram.com/pubs/pdf/SAR11/SAR11.pdf)

5- DHS Spatial Analysis Reports No. 16 A Primer on The Demographic and Health Surveys Program Spatial Covariate Data and Their Applications (https://dhsprogram.com/pubs/pdf/SAR16/SAR16.pdf)

6- High resolution age-structured mapping of childhood vaccination coverage in low- and middle-income countries (GLM) https://eprints.soton.ac.uk/418237/1/1_s2.0_S0264410X18301944_main.pdf

7- Development of High-Resolution Gridded Poverty Surfaces Jan 10th 2014 (BGM) (http://www.worldpop.org.uk/resources/docs/Poverty-mapping-report.pdf)

8- Examining the correlates and drivers of human population distributions across low- and middle-income countries. (RF) Published 13 December 2017 (http://dx.doi.org/10.1098/rsif.2017.0401) ~ this work forms part of the WorldPop Project (www.worldpop.org.uk) and Flowminder Foundation (www.flowminder.org)

9- Mapping poverty using mobile phone and satellite data. (BGM) J. R. Soc. Interface 14: 20160690. Steele JE et al. 2017 (http://dx.doi.org/10.1098/rsif.2016.0690)

10- Bosco C et al. 2017 Exploring the high-resolution mapping of gender-disaggregated development indicators. J. R. Soc. Interface 14: 20160825. http://dx.doi.org/10.1098/rsif.2016.0825