DHS and Geo-covariates data integration

Case study on Bangladesh survey 2014

Yichun Wang (email: yichun.wang@gmail.com)

Data from DHS

Demographic and Health Surveys (DHS) (http://dhsprogram.com) are nationally-representative household survey data.

Information Available in DHS data sets

- wealth and household utilities
- education and work opportunities
- children's nutrition
- family planning
- maternal health
- women's empowerment

Normal covariates from DHS data sets

- urban-rural
- male-female
- rich-poor
- young-old
- different education levels

The case for Geospatial data

- The DHS Program routinely collects geographic coordinates of the primary sampling units (PSU, also known as cluster) in most surveyed countries.

- New resources such as publicly available geographic data as well as new and more accessible geographic information system (GIS) technologies and methods become accessible.

- The need from policy makers to estimate indicators to smaller administrative areas and areas not covered by the surveys.

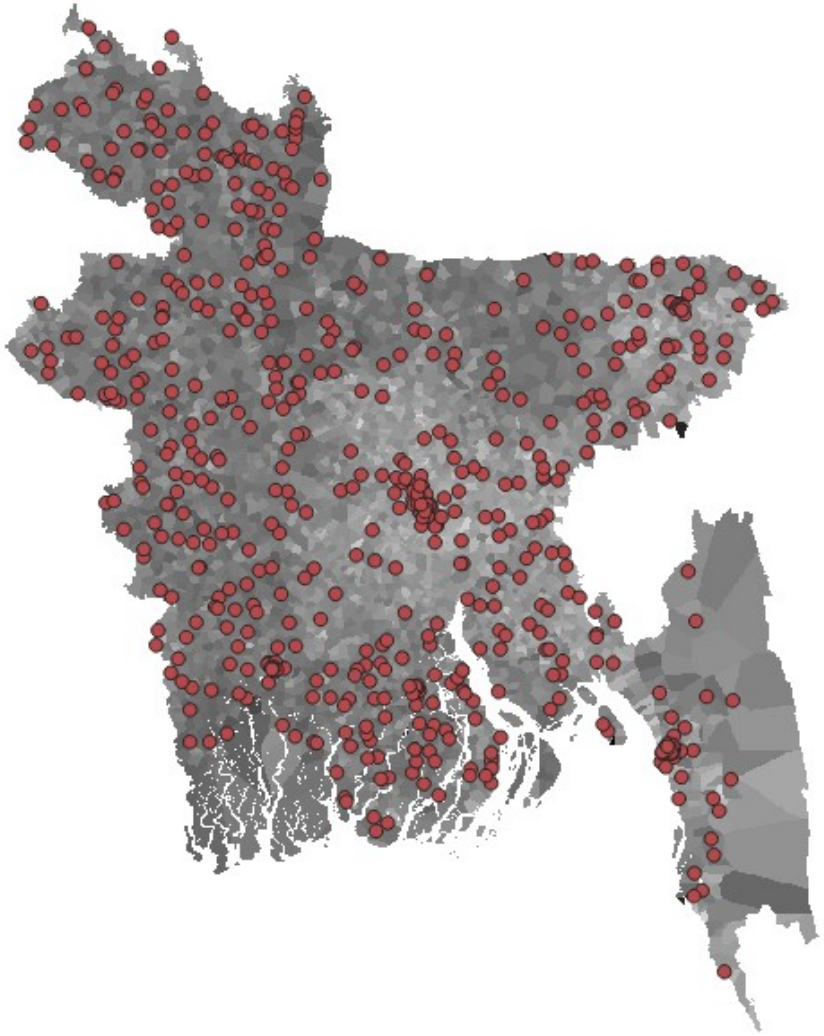
Geo-spatial Data (1)

- Traveling time: (in minutes) to the nearest city of more than 50,000 people.
- SMOD: “degree of urbanization” model using as input the population GRID cells.
- Build-Up: percentage of building footprint area in relation to the total cell area.
- Road friction: Calculated land-based travel speed for given Geo-coordinate position.
- Nightlight Composite: Average radiance composite from night time satellite image data from the Visible Infrared Imaging Radiometer Suite.

Geospatial Data (2)

- Vegetation Indices (AVI): vegetation reflected signal from measured spectral responses by combining two (or more) wavebands.
- Human Footprint: anthropogenic impacts on the environment created from nine global data layers covering human population pressure, human land use and infrastructure, and human access.
- Aridity : climate data related to evapotranspiration processes and rainfall deficit for potential vegetative growth.
- Drought Episode: 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.
- Population Density: number of inhabitants per cell (1km X 1km).
- Three income/poverty/wealth map generated by World Bank for Bangladesh (note, this data is specific to Bangladesh).

Connecting PSU to Grid Data



Grey scale map: 2013 estimates of income in USD per grid square .

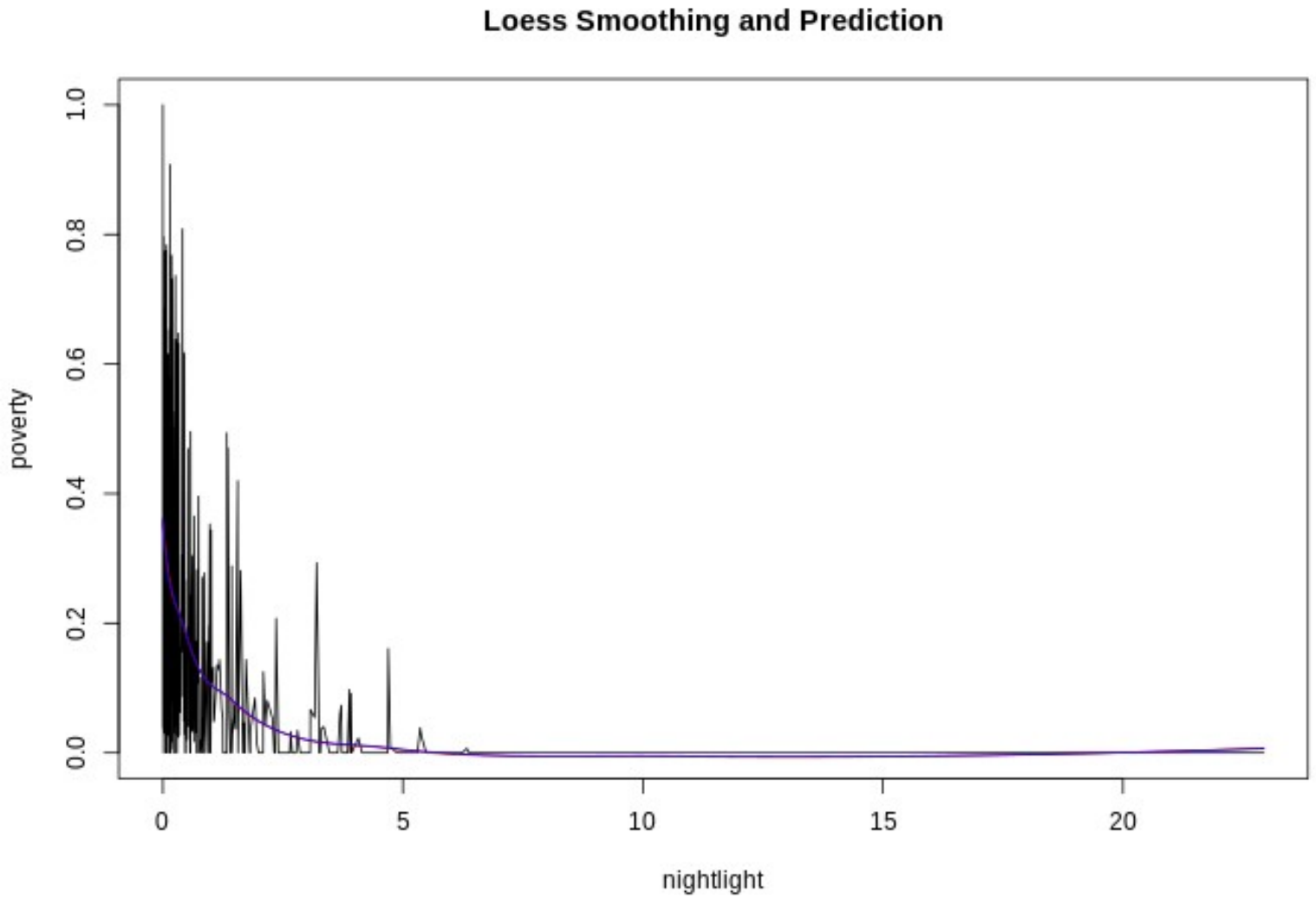
Red Dots: PSUs from 2014 DHS survey.

Take average of income within 2km

Red Dots: a PSU

Yellow circle: 2km neighbourhood of the PSU

Append Geo-Covariates to 599 PSUs



Y: poverty estimated from DHS data for each PSU

X: Night Light composite average 2km around the PSU

Blue Line: average poverty rates as a function of nightlight composite.

Note that poverty rate lowers as nightlight compoiste increases. Also the variation of poverty rate increases as nightlight composite decreases.

Using Geo-Covariates to estimate indicator values

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

- Y represents the indicator value for j-th household (person) in i-th PSU. Si represents the location of the PSU, ei() is the function that represents the effect of the location.
- This model is assuming that the value of the household is decided by its PSU location (which decides values of the geo-covariates) and households/personal choices.

Using Geo-Covariates to estimate Indicator values

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

- 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

Predictive model

- Several predictive models, including GLM, Decision Trees, Random Forest, and XGboosting have been explored. Random Forest gives the best results.
- Several explanatory variable sets, including geo-covariates provided by DHS, geo-covariates obtained from public sources, service accessibility data provided by DHS.

Explanatory variables

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

Variable set selection (selected indicators)

Indicator	Best variable set	Best Pooled RMSE	Pooled RMSE (GCUDALL)	F-test	Conclusion
Poverty-ALL	SQGCUDALL	13.07%	13.28%	1.975	GCUDALL is as good
CleanFuel-ALL	SQGCUDALL	12.49%	12.94%	4.283	SQGCUDALL is better
SchoolAgeEducation-CollegeAge	SQGCUDALL	16.79%	16.98%	1.262	GCUDALL is as good
AdultEducation-Young	SQGCUDALL	10.96%	11.36%	4.290	SQGCUDALL is better
ProfessionalAssistance-birth delivery	SQGCUDALL	22.90%	23.12%	1.132	GCUDALL is as good

Statistically, for some indicators, information on service accessibility improves model predictive power.

Model Results (selected Indicators)

Indicator	Total standard deviation SD_t	Cluster Standard deviation SD_c	Proportion cluster variation / total (SD_c^2/SD_t^2)	Proportion model variation / total $(SD_t(RF)^2/SD_t^2)$	Proportion model variation / cluster $(SD_c(RF)^2/SD_c^2)$
Poverty-ALL	38.08%	19.56%	26.39%	13.37%	53.91%
AccessElectricity-ALL	44.20%	25.02%	32.04%	17.86%	58.05%
CleanFuel-ALL	36.06%	30.92%	73.55%	58.45%	82.48%
CleanWater-ALL	15.16%	11.54%	57.87%	29.91%	53.65%
ImprovedSanitation-ALL	45.95%	24.12%	27.55%	11.75%	45.50%

Sanitation and electricity have the biggest total standard deviation

Clean fuel has the biggest cluster standard deviation, highest proportion of cluster variation / total

Geo-covariate model explains cluster level variation

Model Results (selected Indicators)

Indicator	**SD_t**	**SD_c**	(SD_c^2/SD_t^2)	$(SD_t(RF)^2/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$
SchoolAgeEducation-UpperSecondaryAge	49.35%	23.15%	22.00%	2.30%	19.20%
CurrentlyWorking-WorkingAge	49.39%	11.62%	5.54%	1.74%	36.99%
Stunt-Under5Age	48.14%	18.90%	15.42%	2.76%	27.57%
PowerHealth-CurrenlyMarried	48.22%	14.52%	9.07%	2.07%	31.57%
ProfessionalAssistance-birth delivery	49.68%	29.74%	35.83%	11.99%	39.55%

Low cluster level variation observed

Adding individual/household variables

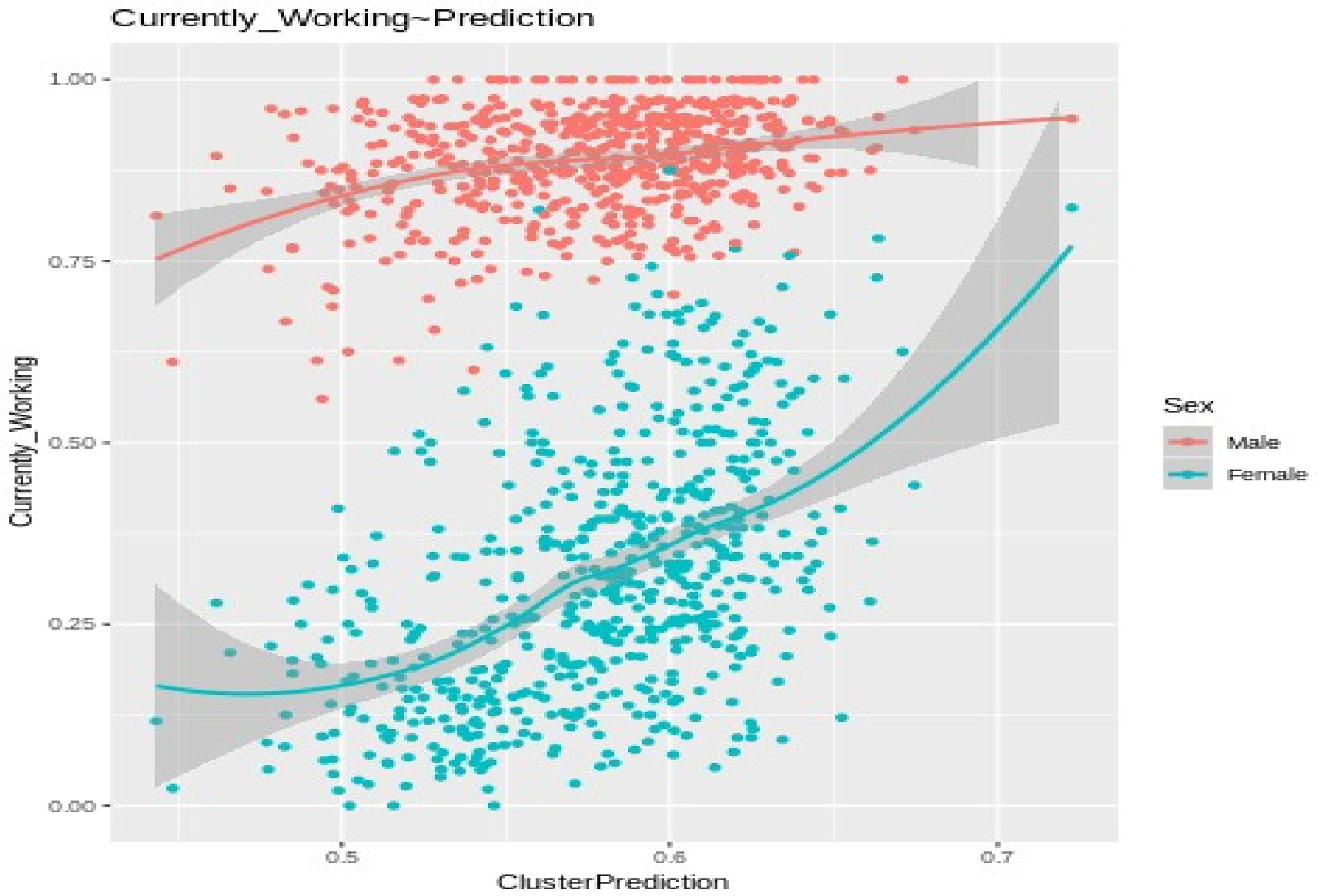
	GCUDALL		GCUDALL + Sex	
Indicator	$(SD_t^2(RF)/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$	$(SD_t^2(RF)/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$
Currently Working	1.72%	36.75%	33.09%	37.30%

Adding individual variable (Sex) to the model, does not improve prediction on cluster level, but improve prediction on total level.

Indicator	**GCUDALL**		**GCUDALL + Education**		**GCUDALL + Age**	
Currently Working	$(SD_t^2(RF)/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$	$(SD_t^2(RF)/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$	$(SD_t^2(RF)/SD_t^2)$	$(SD_c(RF)^2/SD_c^2)$
Female	4.12%	32.15%	4.51%	29.92%	5.36%	26.86%
Male	0.58%	25.59%	7.32%	30.65%	14.43%	20.37%

Male and female have different model patterns. Geo-covariates effect female more than male, while education and age affect male more than female.

Currently Working



Conclusion

- We also did Bayesian Geo-statistical models, but the model has little contribution to improving the predictive power.
- We learned how to integrate DHS data with publically available geospatial data. For some of the indicators, these geo-covariates can be a reliable predictor.
- From published research literature, we know that the power of geo-covariates also varies from country to country.

Conclusion

- We found for some indicators, knowing the availability of services can add to power of prediction. These data are not publically available, but most governments have the data.
- Some of the geo-spatial data is outdated, some are not defined well for our region. Updated data is needed.