Child Marriage and Geo-covariates data integration

Data from DHS

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

Information Available in DHS data sets

- Women Interview data: 15-49 year old, “Eligible” for interview
- Age at first marriage
- Child marriage – married before age 15, married before age 18
- Individual characteristics of respondents: age, household wealth index, education
- GIS coordinates of the Primary Sampling Units (village) the respondent is from.

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.

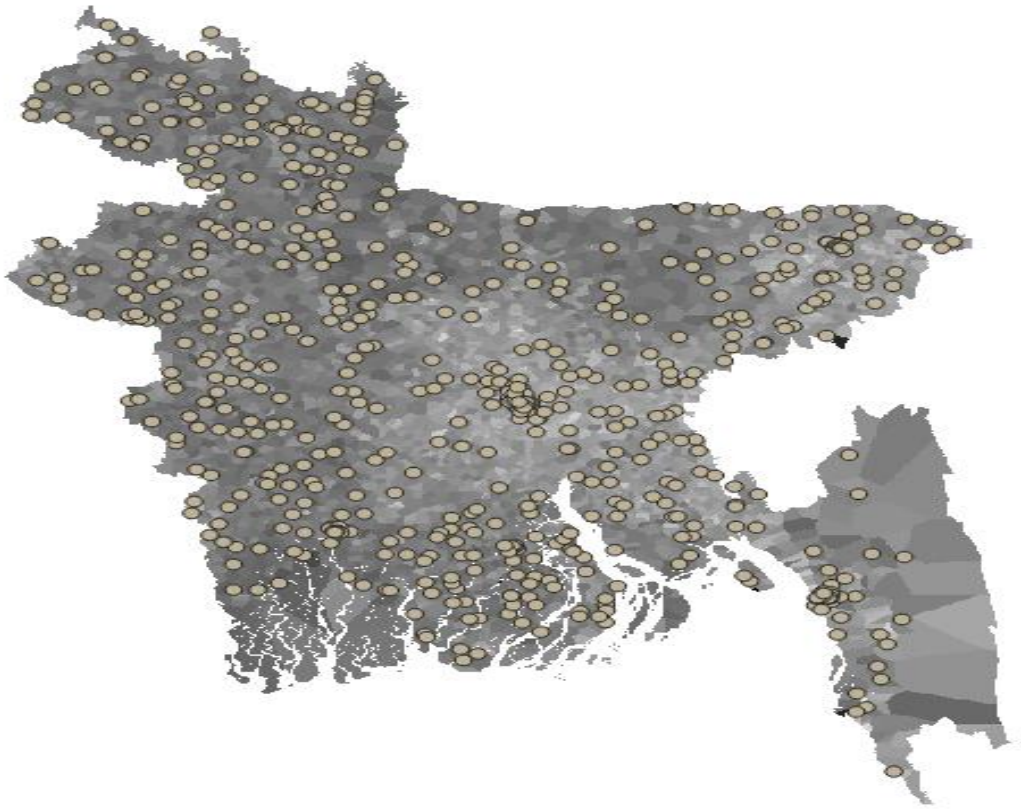
Geo-spatial Data (1)

- 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 .

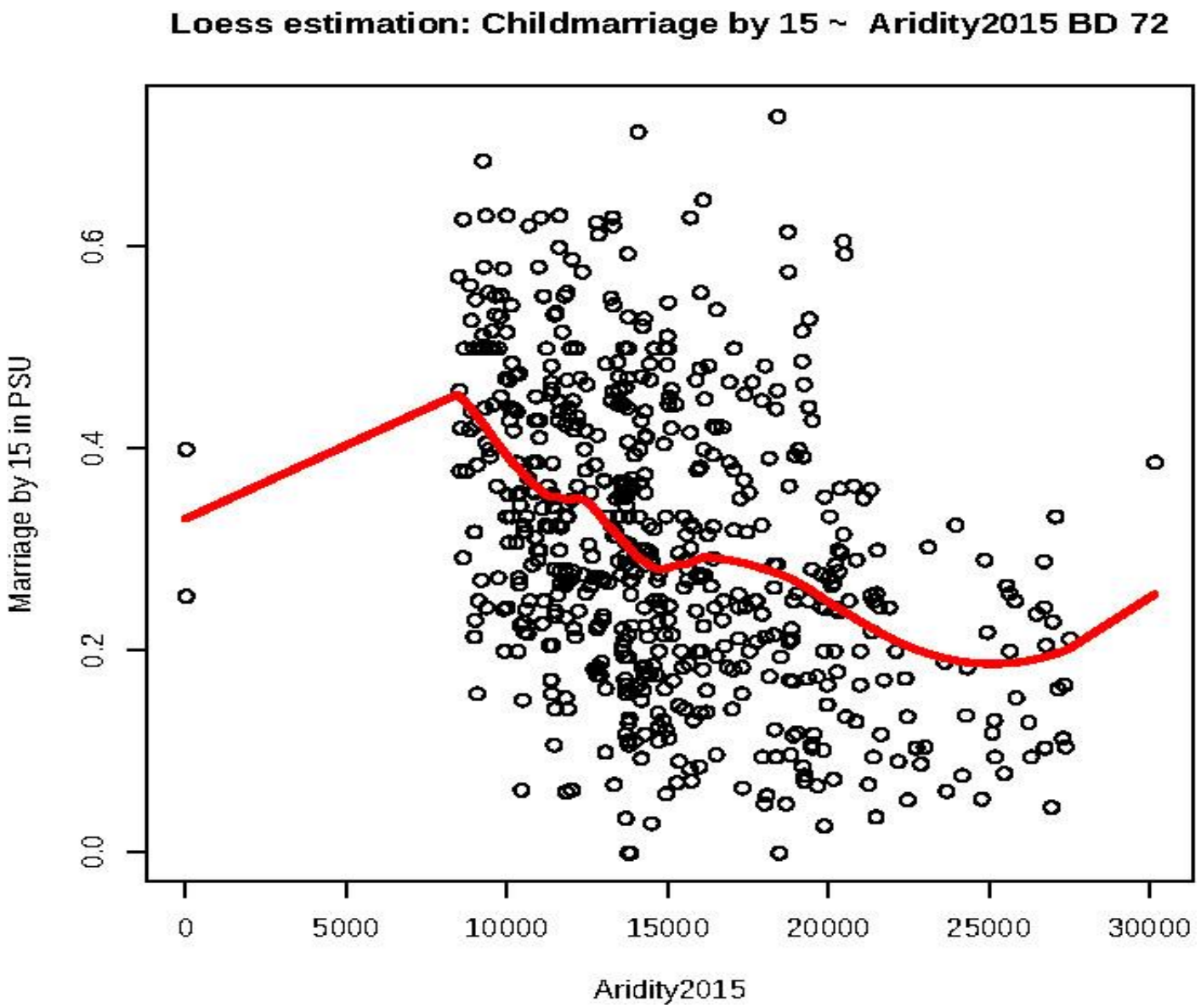
Red Dots: PSUs from 2014 DHS survey.

Take average of income within 2km

Red Dots: a PSU

Yellow circle: 2km neighbourhood of the PSU

Append Geo-Covariates to 599 PSUs



Y: % of women married before 15 yo from DHS data for each PSU

X: Aridity Index average 2km around the PSU

Red Line: average child marriage rates as a function of Aridity Index.

Note that child marriage rate lowers as Aridity Index increases, which means humidy and rain fall increases.

Using Geo-Covariates to estimate indicator values

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

- Y represents the indicator value for j-th respondent in i-th PSU (e.g. Y=1 if she married before 15 year old, Y=0 if not). Si represents the location of the PSU, ei() is the function that represents the effect of the location.
- This model is assuming whether the respondent married by age of 15 is decided by her village (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, family wealth, education, etc

Child Marriage Stats

	Bangledesh 2014	Bangladesh Rural	Bangladesh Urban
Married before 15 (15-49) – sample calculation	30.36%	32.55%	24.95%
Married before 18 (18-49) – sample calculation	68.99%	72.66%	59.99%

Correlation Coefficients

	Bangladesh		Rural		urban	
	Before 15	Before 18	Before 15	Before 18	Before 15	Before 18
Travel_Times2015	0.217	0.272	0.077	0.0956	0.289	0.231
SMOD2015	-0.290	-0.373	-0.109	-0.1361	-0.340	-0.330
Buildup2015	-0.222	-0.318	-0.039	-0.0989	-0.226	-0.255
Friction2015	0.174	0.196	0.052	0.0221	0.253	0.222
Nightlight2015	-0.254	-0.366	-0.074	-0.1394	-0.264	-0.310
Avi2015	0.092	0.112	0.031	0.0683	0.127	0.075
Hfp2004	-0.242	-0.336	-0.027	-0.0355	-0.279	-0.299
Aridity2015	-0.392	-0.340	-0.440	-0.4150	-0.321	-0.275
DroughtEpisode	0.283	0.272	0.301	0.2872	0.246	0.267
Density2015	-0.259	-0.365	-0.048	-0.1007	-0.278	-0.311
aWealthIndex2011	-0.346	-0.448	-0.250	-0.2966	-0.309	-0.336
aIncome2013	-0.414	-0.472	-0.355	-0.3734	-0.363	-0.360
aPP2013	0.440	0.415	0.394	0.3533	0.412	0.344
FloodRisk25	0.067	0.083	0.000	-0.0169	0.088	0.092
FloodRisk50	0.070	0.083	0.001	-0.0063	0.093	0.060

Using Random Forest modeling to predict PSU level Child Marriage

- Radom Forest is one of the most popular machine learning methods in prediction.
- It is a Decision Tree based Ensemble method.
- It boost-straps random sub-samples, and build trees for each subsample, then take the average of the results trees.
- It produces good prediction power.

Model Results

	SDt	SDc		R-sq
Child Marriage	Total Variation	Cluster Variation	Cluster/Total Variation	Model Variation/Cluster
before 15	46%	13%	8%	48.83%
before 15 rural	0.47	0.15	10%	54.79%
before 15 urban	0.44	0.14	10%	58.94%
before 18	0.466	0.141	9%	55.56%
before 18 rural	0.45	0.14	10%	55.70%
before 18 urban	0.49	0.17	13%	62.14%

Variable Importance

	Before 15 – model I		Before 15 – model II		Before 18 – model I		Before 18 – model II	
	%IncMSE	NodePurity	%IncMSE	IncNodePurity	%IncMSE	IncNodePurity	%IncMSE	IncNodePurity
Travel_Times2015	7.33	12.73	3.85	6.00	6.02	20.78	5.04	14.98
SMOD2015	5.42	8.68	5.69	5.93	7.13	17.80	5.09	11.52
Buildup2015	6.46	11.88	6.33	6.67	5.91	15.92	5.79	8.91
Friction2015	6.66	7.65	5.11	4.50	4.82	6.62	4.87	5.11
Nightlight2015	4.10	5.72	4.34	3.94	5.62	21.63	5.17	13.38
Avi2015	10.41	8.04	7.70	7.31	15.53	20.09	13.95	18.38
Hfp2004	7.26	8.80	6.65	6.66	4.58	11.10	4.95	5.39
Aridity2015	13.89	31.36	9.67	22.86	12.40	13.93	7.85	7.35
DroughtEpisode	8.79	17.73	5.03	9.49	4.10	4.71	3.59	1.78
Density2015	3.25	5.86	3.34	3.73	4.79	9.20	5.39	6.83
FloodRisk25	3.37	0.82	3.15	1.14	3.36	1.04	3.00	0.70
FloodRisk50	5.47	1.85	4.22	1.61	5.86	3.60	5.47	1.85
aWealthIndex2011			5.52	6.16			6.36	18.73
aIncome2013			4.94	9.02			5.75	10.00
aPP2013			11.87	38.08			12.98	27.61

Conclusion

- We used two sets of variables in model 1 and 2. The difference is the three economic variables – average wealth index, average probability of poverty, average personal income. Model 1 excluded these three variables, because these information, unfortunately, is not always available for many countries.
- We presented R-sq for model 2 in our previous slide. Model 1 does have 3-6% lower R-sq compared to model 1. Indicating that the three economic variables are powerful.
- Aridity Index, which measure the humidity and rain fall, are also powerful in predicting child marriage. The aridity index used in this exercises is based on climate data from 1950-2000. New updated version has just been released earlier this year.
- Overall, we can say, rural area suffer higher child marriage rate, poorer villages also. For marriage before 15, drought and aridity index play important roles. For marriage before 18, night light index, travel time to the city, and vegetation index play important roles. Predicted Flood risk, however, has weak association with child marriage rate.