MULTI-DIMENSIONAL POVERTY PREDICTION AND MAPPING USING NOVEL DATA AND METHODS - COMBINING TRADITIONAL DATA WITH BIG DATA

-Neeti Pokhriyal, Dartmouth College

-Combining disparate data sources for improved poverty prediction and mapping. Pokhriyal and Jacques, Proc. Of National Academy of Sciences, 2017

-Virtual Networks and Poverty Analysis in Senegal, N. Pokhriyal et al. NetMob, MIT, 2015

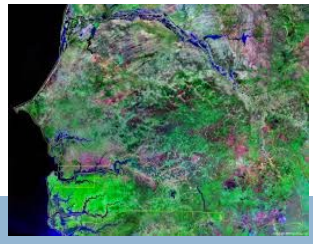
-Estimating and Forecasting Income Poverty and Inequality in Haiti Using Satellite Imagery and Mobile Phone Data, Neeti Pokhriyal, Omar Zambrano, Jennifer Linares, Hugo Hernandez, Working Paper, Inter-American Development Bank, 2020.

Talk Outline

- Idea
- Datasets used
- Challenges
- Method
- Results
- Conclusions

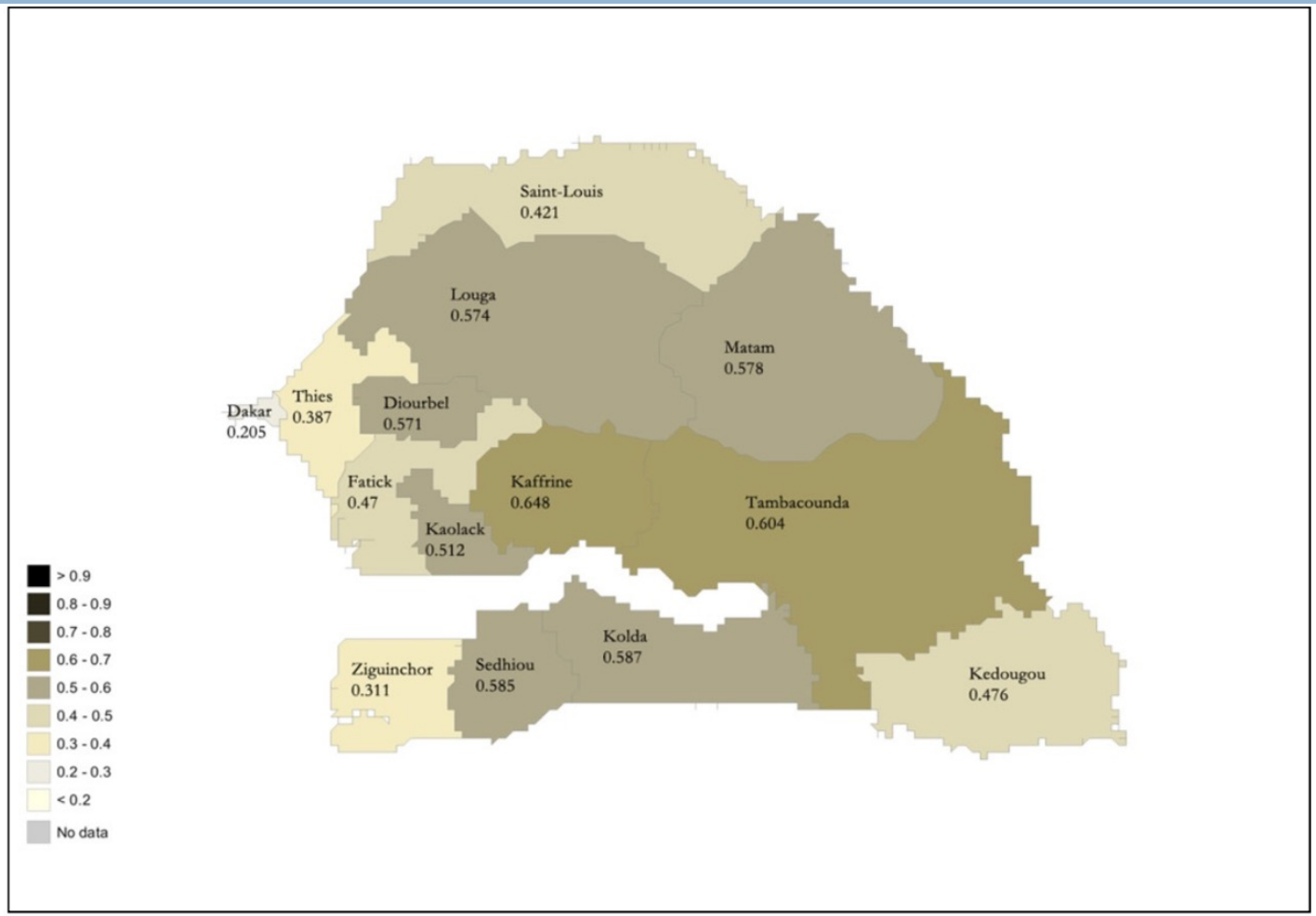
Idea

- Traditional ways to measure poverty
  - costly – time and money
  - timely updates of poverty – difficult
- To combine *rich* census and surveys with *auxiliary* data, like
  - mobile phone data
  - satellite and aerial imagery
  - weather stations
  - economic data
  - open street maps etc.
- **For**: an accurate intercensal estimates at policy planning microregions along the dimensions of MPI.



Detect signal of MPI

Poverty Map of Senegal



Oxford Poverty and Human Development Initiative, Country briefing: Senegal. Available at https://www.ophi.org.uk/wp-content/uploads/Senegal-2013.pdf.

Disparate datasets used

Table 1. Summary statistics and characteristics of the data used—CDRs, environment, census, and MPI

Summary statistics	CDRs	Environment data	Census	Poverty index
Timeline	January–December 2013	1960–2014	2013	2013
Number of total calls and text	11 billion	N/A	N/A	N/A
Number of unique individuals	9.54 M	N/A	1.4 M	N/A
Spatial granularity of available data	Antenna level (1666)	Vector data—100 $m^{-1} \cdot km$	Household level	Region level (14)
Cost incurred in data collection and preparation	Low/no cost (data exhaust)	Low/no cost (data exhaust)	US$29 million	Very high cost, and human expertise
Frequency of update of data	Real time	~1 y	3–5 y	3–5 y

Environmental Data

- Food security
  - Temperature; Precipitation
  - Elevation; Slope
  - Soil Type
- Economic Activity
  - Nighttime lights – intensity of urbanization
  - Land cover
- Accessibility to services
  - Proximity to urban centers, markets, main roads, schools/university, water tower, hospital
- Accessed via satellite and open street map products

Jacques D, et al. Genesis of millet prices in Senegal: The role of production, markets and their failures. *NetMob, MIT, 2015*
Min B, Gaba KM, Sarr OF, Agalassou A (2013) Detection of rural electrification in Africa using DMSP-OLS night lights imagery
Njuguna C, McSharry P (2017) Constructing spatiotemporal poverty indices from Big Data. *J Business Res*

Digital data - Call data records (CDRs)

- Population structure, socioeconomic ties, cultural interactions, and micro and macro patterns of human interaction are essential to understanding poverty.
- One way to study societal interactions is provided by the widespread use of digital technologies.
- Mobile phones are a prevalent technology, even with widespread poverty in Sub-Saharan Africa.
- CDRs capture how, when, where, and with whom individuals communicate in an anonymized manner.
- We use it in a privacy preserving manner and employ spatial and temporal aggregation at the level of microregions.

Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science*

Digital data

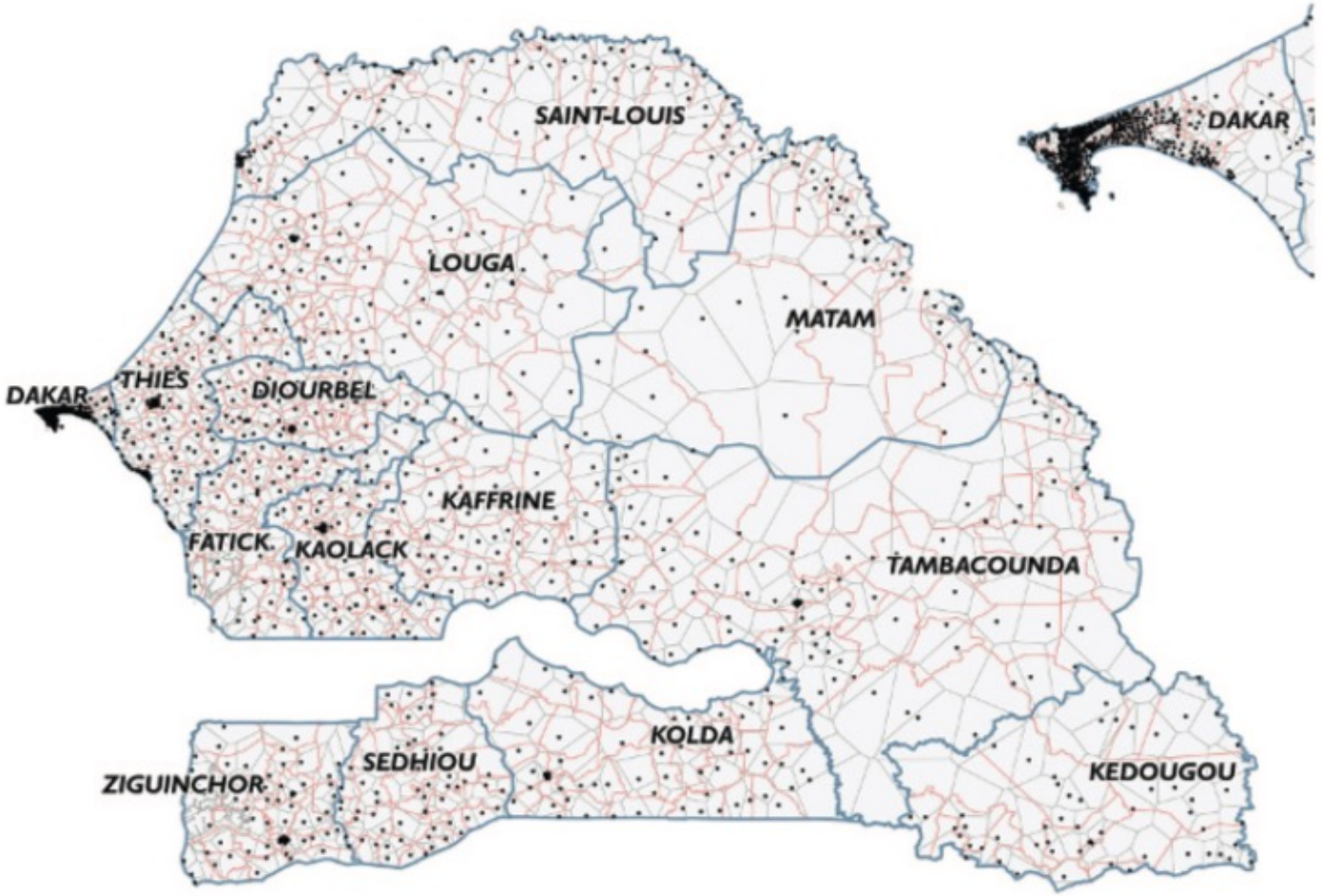


Figure: Black dots depict the location of mobile towers. The Voronoi tessellation formed by these towers is shown in gray. Commune boundaries are shown in red.

CDR data features

- Regularity of call
- Diversity of call
- Active characteristics
- Basic phone usage
- Spatial characteristics

Figure 1: Sample Call Detail Records

Interaction	Direction	Correspondent ID	Date and Time	Call Duration	Antenna ID
Call	In	8f8ad28de134	2012-05-20 20:30:37	137	13084

Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA*
Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science*
Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E (2011) Prediction of socio-economic levels using cell phone records. *Proceedings of the 19th International Conference on User Modeling, Adaption and Personalization,* Springer
Sundsøy P (2016) Can mobile usage predict illiteracy in a developing country? arXiv:1607.01337.

Challenges

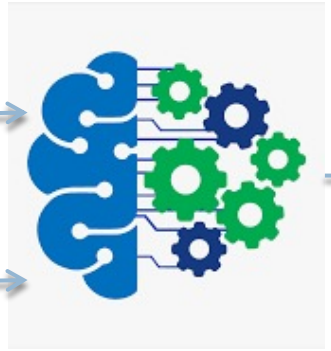
- Varying spatial granularity of different datasets
  - CDR data are available for each subscriber
  - Environmental data have mixed spatial resolution
  - Census and survey data are available for either individuals or households, depending on the response variable
- An aggregation mechanism to link them.
  - Extract the targets and inputs at policy planning microregions
- High dimensional features space characterizes data from different sources – overfitting with limited data points
- Data sharing among the ecosystem could be contentious owing to privacy and business concerns.
  - Method keeps each data private what is shared are the poverty estimates and uncertainty estimates that facilitates combining the different data sources.

Method details

Feature Extraction : *Turning raw data into quantitative metrics.*

Satellite and Mobile phone data for a microregion

Input



Gaussian Process-based regression model

MPI targets from census for a microregion

Output

Objective: Learn a relationship/mapping between inputs and outputs and validate it for out-of-sample generalization, which is done using spatial cross-validation procedure.

Model helps to:

1) Predict output given an input.
2) Insights as which input features are important for predicting MPI.
3) Provides uncertainty with its poverty estimates – measure of trust of our model.

Methodology

- For each data source, a Gaussian Process* (GP) Regression model is learnt of the following form:

Targets at a spatial location

$$y_i = \beta^\top \mathbf{x}_i + f(\mathbf{x}_i) + \epsilon$$

Models the nonlinear relationship

Features extracted from each data source

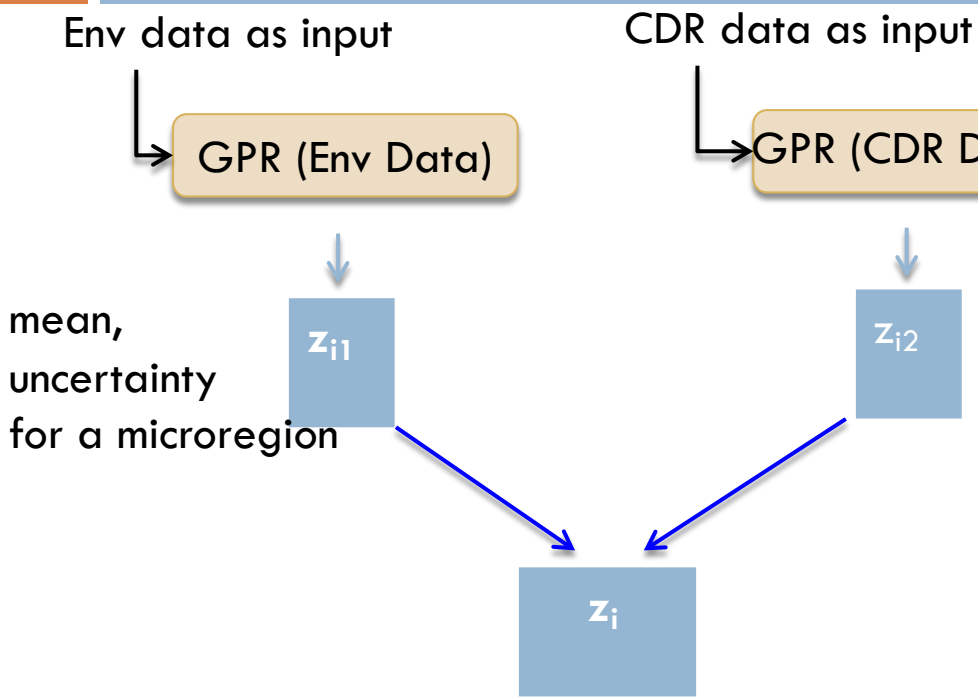
- GPs belong to the class of Bayesian non-parametric models, where no assumptions are made on the functional form of the relationships between covariates and targets, and thus, these methods are known to learn highly non-linear functions.
- A kernel function lies at the heart of GPs, which can encode different relationships between the covariates and the targets.
- Ability to incorporate prior knowledge along with observational data.

*Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning. (The MIT Press).

Methodology

- An independently trained Gaussian Process Regression (GPR) model for each data (source) – keeps data private
- Given a microregion's covariates as input, each GPR model outputs its own estimate of poverty and uncertainty
- The final estimate is a weighted mixture of each outputs' estimate, where weight captures the certainty of model in giving that estimate
  - Idea: give more weight to the estimates with higher certainty.
- To mitigate overfitting while learning from limited data points and high dimensional feature spaces, regularization is used
  - prevents learning from spurious features => feature selection

Proposed GP Fusion – Illustration



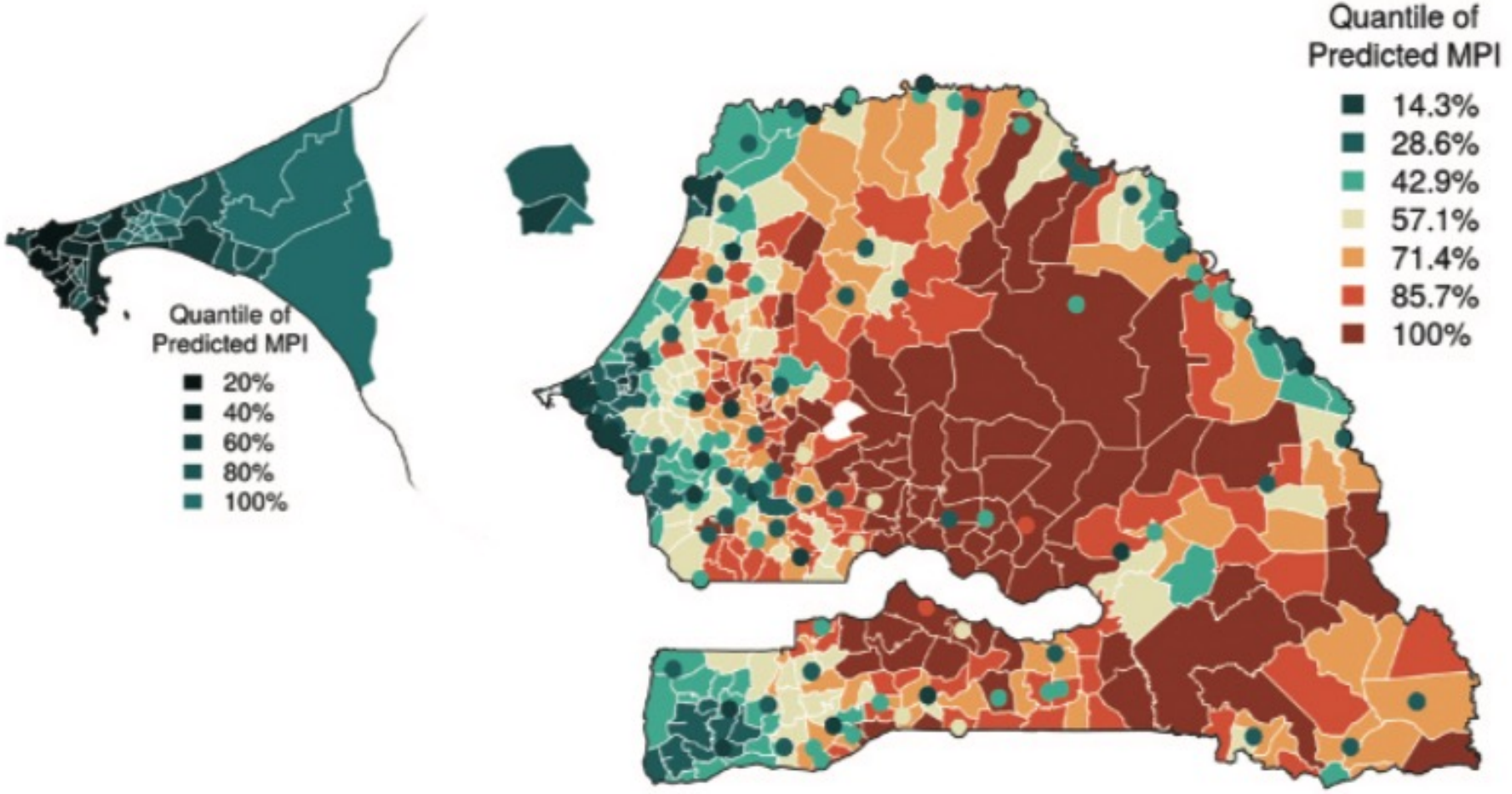
$$w_{1i} = \frac{\frac{1}{\sigma_{1i}^2}}{\frac{1}{\sigma_{1i}^2} + \frac{1}{\sigma_{2i}^2}}$$

$$w_{2i} = \frac{\frac{1}{\sigma_{2i}^2}}{\frac{1}{\sigma_{1i}^2} + \frac{1}{\sigma_{2i}^2}}$$

$$\bar{z}_i = w_{1i}\bar{z}_{1i} + w_{2i}\bar{z}_{2i}$$

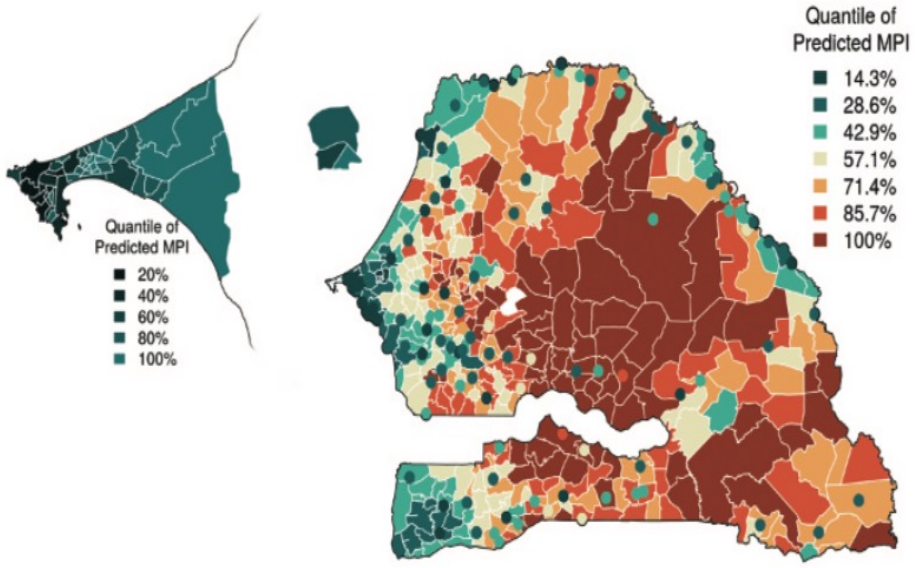
$$\sigma_i^2 = w_{1i}\sigma_{1i}^2 + w_{2i}\sigma_{2i}^2 + w_{1i}w_{2i}(\bar{z}_{1i} - \bar{z}_{2i})^2$$

Results

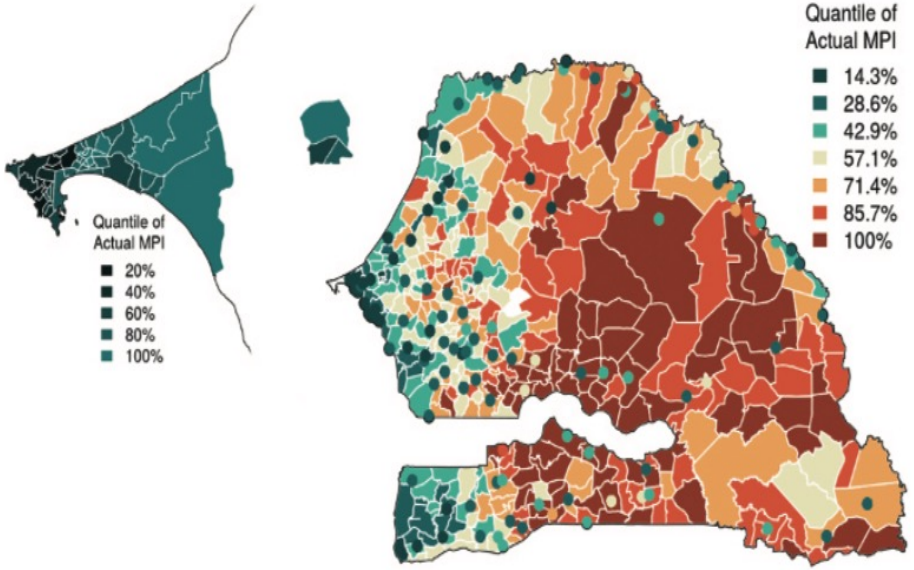


Dots on the map:121 urban centers. Rest are 431 rural communes

Estimated Poverty Map

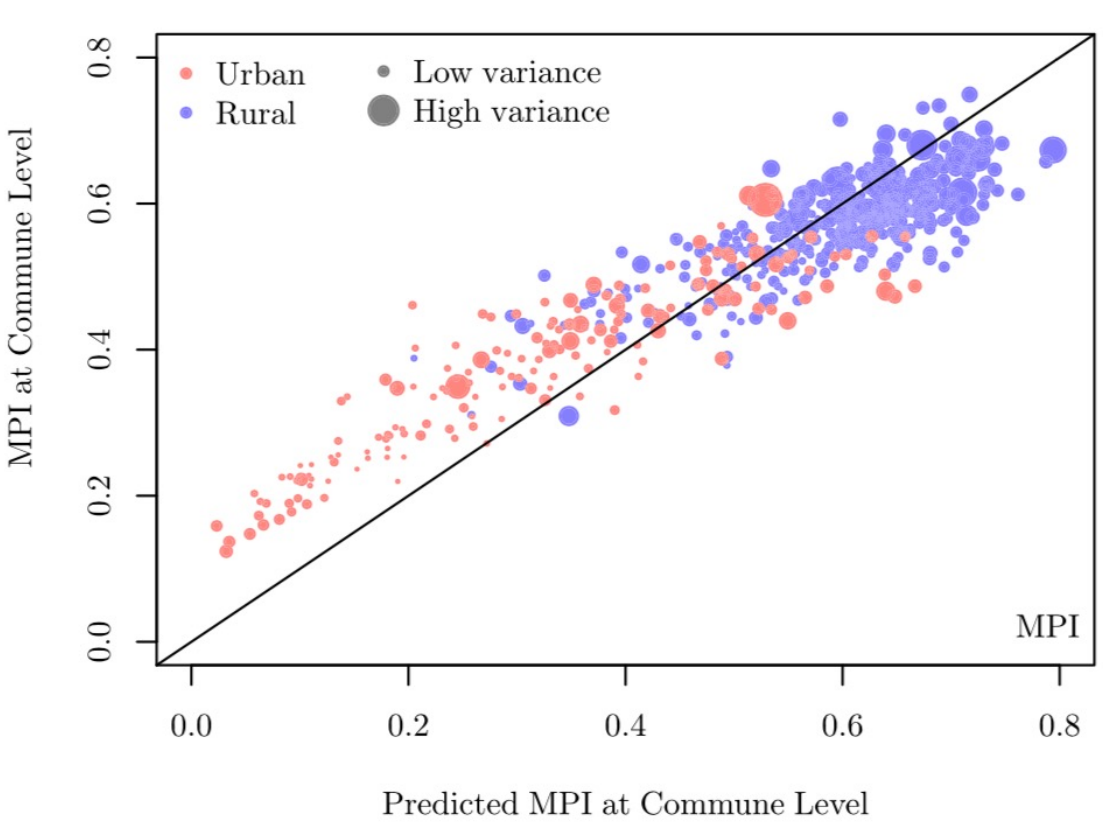


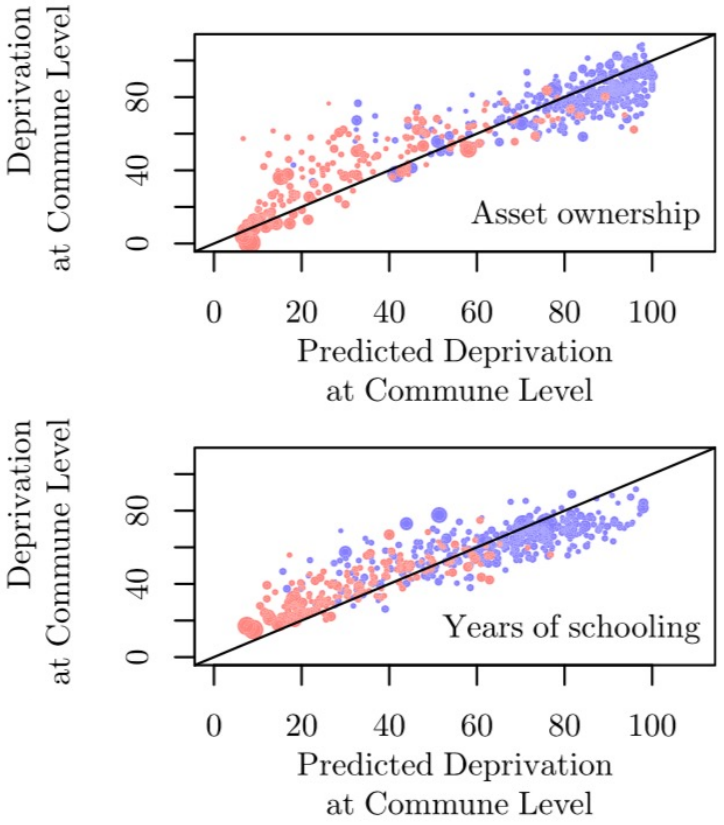
Predicted using our model



Estimated from the census

Validation against ground truth (Census)

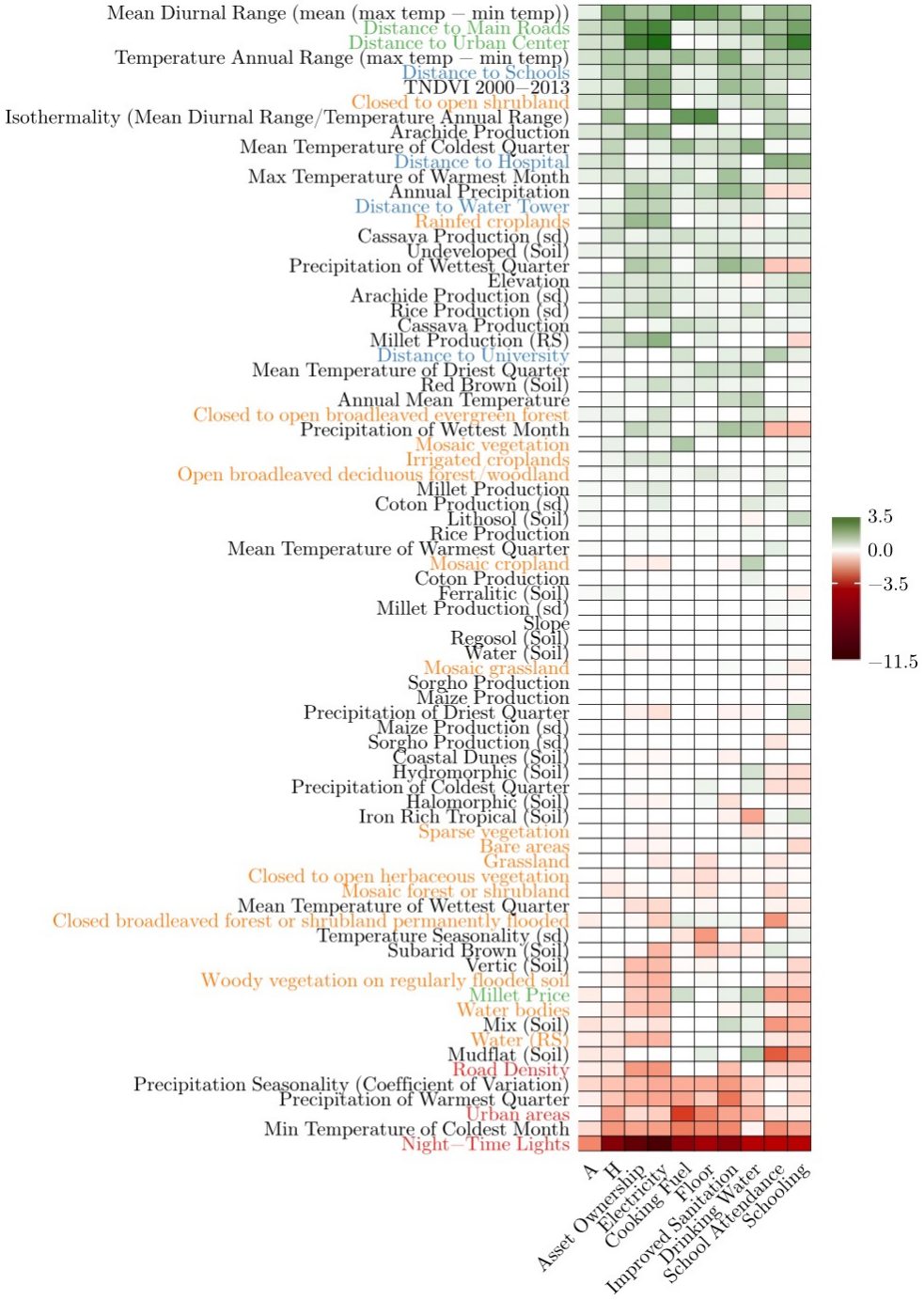




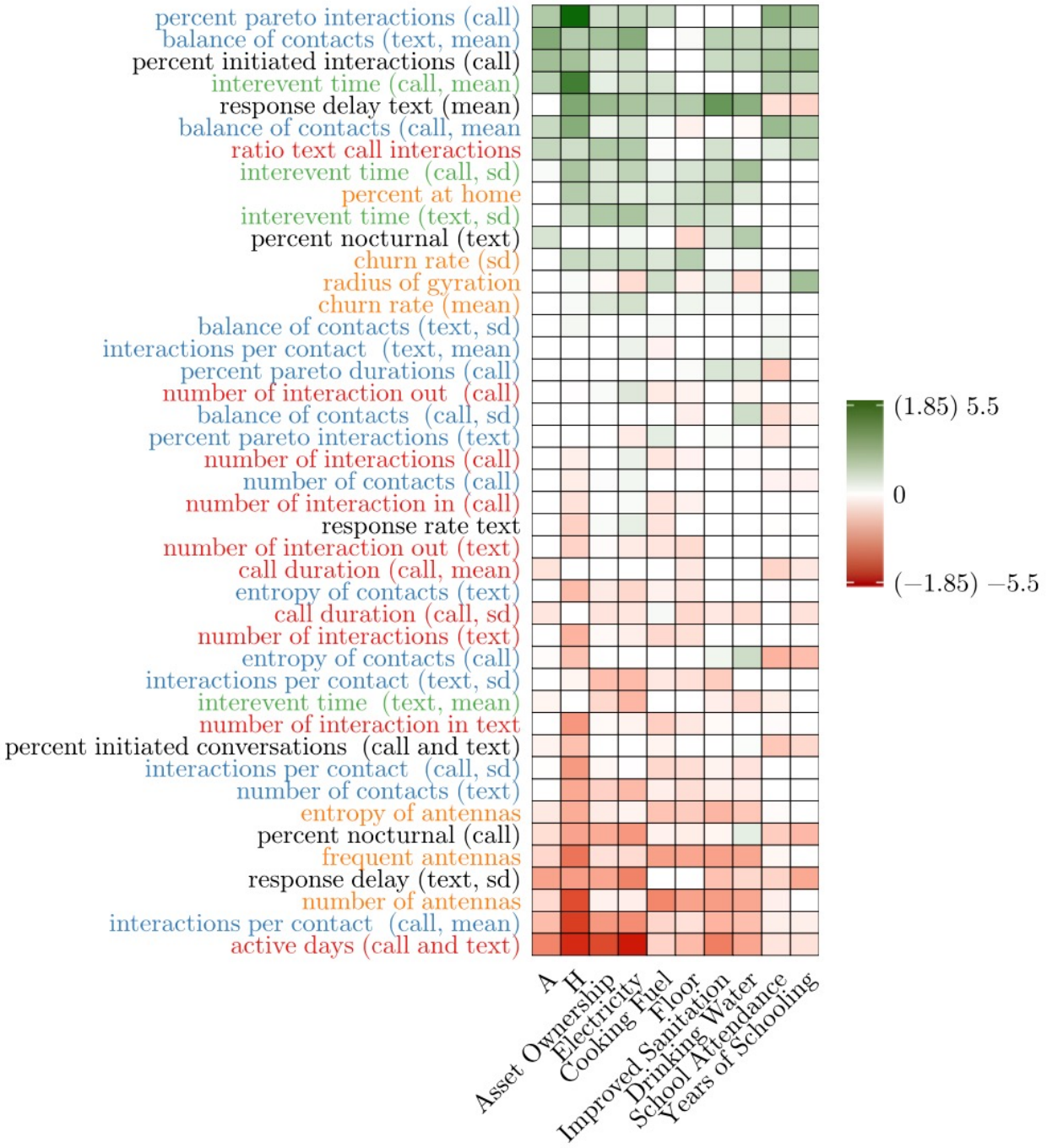
Quantitative Results

Poverty indicator	Multisource data		
	Corr.	Rank corr.	RMSE
MPI	0.91 (0.06)	0.88 (0.06)	0.08 (0.01)
H	0.91 (0.07)	0.85 (0.08)	10.79 (3.96)
A	0.86 (0.05)	0.85 (0.07)	04.71 (0.96)
Individual indicators of poverty			
Education			
Years of schooling	0.85 (0.04)	0.85 (0.04)	12.00 (1.21)
School attendance	0.86 (0.05)	0.83 (0.06)	11.68 (1.83)
Health			
Child mortality	0.45 (0.15)	0.46 (0.16)	10.91 (0.58)
Nutrition	0.52 (0.15)	0.53 (0.15)	14.61 (3.65)
Standard of living			
Cooking fuel	0.86 (0.14)	0.70 (0.18)	13.82 (8.76)
Sanitation	0.79 (0.17)	0.70 (0.18)	16.99 (3.42)
Water	0.75 (0.14)	0.72 (0.14)	14.60 (3.22)
Electricity	0.88 (0.04)	0.84 (0.07)	15.09 (0.98)
Floor	0.78 (0.15)	0.68 (0.14)	15.79 (5.79)
Asset ownership	0.89 (0.04)	0.86 (0.05)	12.61 (1.33)

Visualization of selected features – environment data



Visualization of selected features – CDR data



Concerns

- Selection bias arising from mobile phone ownership and using data from only one provider.
- Some demographic subgroups like children and the ultra poor are left out by the analysis
- Data sharing among the ecosystem could be contentious owing to privacy and business concerns.

- Results may be biased toward urban regions, rather than rural regions, because of factors like lack of electricity in rural areas.
- Our model doesn't do well for nutrition and child mortality dimensions.
- Validation for intercensal periods or when no ground truth data is available.

Conclusions

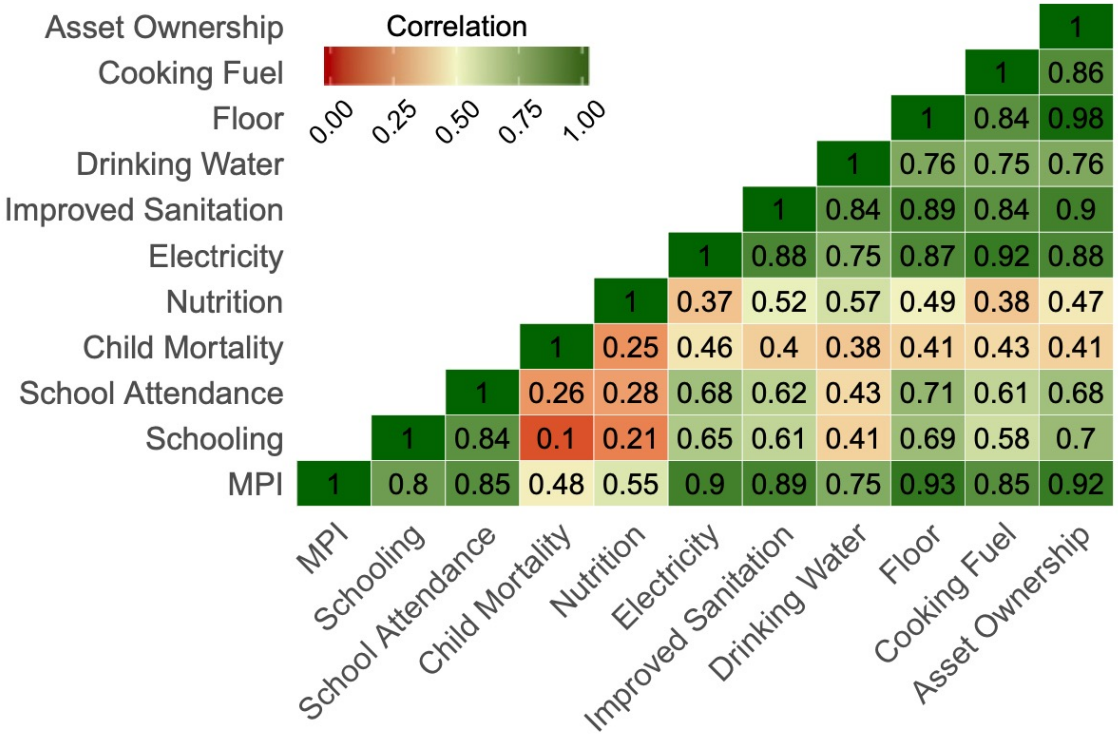
- Our method can combine disparate data to provide accurate and frequent MPI maps at policy planning locations – works well for small datasets and provides interpretability
  - supplement surveying tasks.
- Uncertainties aid policy makers in providing a measure of trust in the model's estimates.
- CDR data seems to have better predictive power.
- Mitigated concerns of intercensal validation by exploiting the correlations in the targets of deprivations – *to provide the evolution of energy poverty for intercensal periods in Senegal.*
- Interesting to see how more local, fine granularity and diverse data help in understanding the deprivations of MPI.

Backup

Conversion of census data to MPI using OPHI's methodology

Poverty indicators	Deprivation standards of a household used by OPHI for MPI calculation	RGPHAE census questionnaire response used by our methodology for MPI calculation
Health		
Child mortality	At least one child has died	About living and deceased children in the household
Nutrition	Any member is undernourished	About going hunger at night for the past few months
Education		
School attendance	Any school-aged child is not attending school up to grade 8	About school-aged currently not in school
Years of schooling	No member who has completed at least 5 y of education	About higher schooling of any member
Standard of living		
Cooking fuel	Uses solid fuels for cooking	Household does not use electricity or natural gas for cooking
Electricity	No access to electricity	No electricity or generator
Sanitation	No access to adequate sanitation or if it is shared	Household has no sewer connection or pit
Drinking water	No access to safe drinking water	No water tap in household
Flooring	Has dirt/earth/dung floor	Household has dirt/earth/dung floor
Assets	Has only one small asset (radio, TV, refrigerator, phone, bicycle, motorbike) and it has no car	Household has one asset (radio, TV, refrigerator, phone, bicycle, motorbike) and it has no car

Features (Number of statistics)	Description
	Regularity
Interevent time (4)	The interevent time between two records of the user.
	Diversity
Number of contacts (2)	The number of contacts the user interacted with (call and text handled separately).
Entropy of contacts (2)	The entropy of the user's contacts, both for call and text.
Balance of contacts (4)	The balance of interactions per contact, This feature is calculated - each for text and call. For every contact, the balance is the number of outgoing interactions divided by the total number of interactions (in+out)
Interactions per contact (4)	The number of interactions a user had with each of its contacts.
Percent pareto interactions (2)	The percentage of user's contacts that account for 80% of its interactions.
Percent pareto durations (1)	The percentage of user's contacts that account for 80% of its total time spend on the phone.
	Active Behavior
Percent nocturnal (2)	The percentage of interactions the user had at night (call and text).
Percent initiated conversations (1)	The percentage of conversations that have been initiated by the user both for call and text.
Percent initiated interactions (1)	The percentage of calls initiated by the user.
Response delay (2)	The response delay of the user within a conversation (in seconds). This is calculated for text (standard deviation and mean of the response delay).
Response rate (1)	The response rate of the user (between 0 and 1).
	Basic Phone Use
Active days (1)	The number of days during which the user was active.
Call duration (2)	The standard deviation and the mean of the duration of user's calls.
Number of interactions (6)	The number of interactions.
Ratio of text & call interactions (1)	This computes the ratio of the text and call interactions.
	Spatial Behavior
Number of antennae (1)	The number of unique places visited.
Entropy of antennas (1)	The entropy of visited antennas.
Percent at home (1)	The percentage of interactions the user had while he was at home.
Radius of gyration (1)	Returns the radius of gyration, the equivalent distance of the mass from the center of gravity, for all visited places.
Frequent antennas (1)	The number of location that account for 80% of the locations where the user was.
Churn rate (2)	The standard deviation and mean of the frequency spent at every antenna each week.
Total	43


Correlation
0.00
0.25
0.50
0.75
1.00
Asset Ownership
Cooking Fuel
Floor
Drinking Water
Improved Sanitation
Electricity
Nutrition
Child Mortality
School Attendance
Schooling
MPI
1
1 0.86
1 0.84 0.98
1 0.76 0.75 0.76
1 0.84 0.89 0.84 0.9
1 0.88 0.75 0.87 0.92 0.88
1 0.37 0.52 0.57 0.49 0.38 0.47
1 0.25 0.46 0.4 0.38 0.41 0.43 0.41
1 0.26 0.28 0.68 0.62 0.43 0.71 0.61 0.68
1 0.84 0.1 0.21 0.65 0.61 0.41 0.69 0.58 0.7
1 0.8 0.85 0.48 0.55 0.9 0.89 0.75 0.93 0.85 0.92
MPI
Schooling
School Attendance
Child Mortality
Nutrition
Electricity
Improved Sanitation
Drinking Water
Floor
Cooking Fuel
Asset Ownership