International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria

Email: moorthy@iiasa.ac.at **Department:** ESM

Working paper

Tracking poverty using satellite imagery and big data

Michiel van-Dijk, Inian Moorthy*, Binh Nguyen, Linda See, Steffen Fritz

*Corresponding author - moorthy@iiasa.ac.at

WP-19-014

Approved by:

Albert van Jaarsveld
Program: Director General and Chief Executive Officer
Date: 19 December 2019

Table of contents

ZVR 524808900

This research was funded by IIASA and its National Member Organizations in Africa, the Americas, Asia, and Europe.

Abstract

Despite recent improvements in the availability and quality of socioeconomic data from developing countries, there are still persistent data gaps that are preventing comprehensive monitoring and evaluation of targets and indicators of the Sustainable Development Goals. Approaches are hampered by the inconsistent spatial and temporal coverage of census data and Demographic and Health Surveys, which serve as the primary source for population-level statistics in most developing countries. Traditional censuses are too expensive to be implemented in remote areas where population density is low and road networks are poor. However, the recent and rapid diffusion of high-resolution satellite imagery offers a new wealth of relatively untapped information that can be used to gain in-depth information on groups that have historically been left out by traditional surveys. Moreover, there are new data streams such as call detail records from mobile phone data networks can help to derive behavioral indicators and improve tracking of expenditure and wealth. This exploratory research conflates various data streams to derive spatially explicit poverty indicators for Senegal with immense scaling potential to other regions.

About the authors

Michiel van-Dijk is a Research Scholar in the Ecosystems Services and Management Program (vandijkm@iiasa.ac.at)

Inian Moorthy is a Research Scholar in the Ecosystems Services and Management Program (moorthy@iiasa.ac.at)

Binh Nguyen is a Research Scholar and Software Engineer in the Air Quality and Greenhouse Gases Program (nguyenb@iiasa.ac.at)

Linda See is a Research Scholar in the Ecosystems Services and Management Program (see@iiasa.ac.at)

Steffen Fritz is the Acting Director of the Ecosystems Services and Management Program and Head of the Center for Earth Observation and Citizen Science (fritz@iiasa.ac.at)

Acknowledgments

The team would like to thank the organizers of the Data for Climate Action Challenge, namely UN Global Pulse, Western Digital, and the Skoll Global Threats Fund. Additionally, we are thankful to Orange and Planet for access to their datasets that made this exploratory work possible.

1. Introduction

The primary goal (SDG #1) of eradicating poverty in all its forms is considered the greatest global challenge and an indispensable requirement for sustainable development. Despite recent improvements in the availability and quality of economic data from developing countries, there are still persistent data gaps that are preventing comprehensive monitoring and evaluation of SDG #1's targets and indicators. Low and low-to-middle income countries are particularly susceptible to the effects of climate change (i.e. droughts, floods, extreme weather events). In addition, the rural poor, who are largely dependent on agriculture for their livelihoods have limited resources to protect themselves against the impacts of weather and climactic shocks. In order to formulate policies and implement strategies that help governments, NGOs and international donors protect the poor against the detrimental impact of climate change, it is essential to have detailed spatially explicit information on where they are located, preferably in the form of high-resolution poverty maps. Approaches are hampered by the inconsistent spatial and temporal coverage of census data and Demographic and Health Surveys (DHS), which serve as the primary source for population-level statistics in most developing countries. Unfortunately, most developing countries do not have the capacity to systematically collect data on household income and wealth that are needed to construct poverty indicators. Furthermore, if this data is available, it is often only presented at the national or subnational level, thus aggregating urban and rural areas, which are known to have very different wealth/income profiles. However, the rapid diffusion of high-resolution satellite imagery and resultant products (i.e. land cover) available at a global scale offers a new wealth of relatively untapped information that can be correlated with economic measures and models. Moreover, there are new data streams such as social media, mobile phone data (e.g. duration of calls, top-up rates and location data), information of financial flows and energy which can help to derive behavioral indicators and improve tracking of expenditure and wealth.

The rapid diffusion of satellite imagery and advances in remote sensing methods have led to the dynamic mapping of poverty and slums [1-3]. More recently, machine learning algorithms have been used to estimate consumption expenditure and asset wealth from high resolution satellite imagery [4]. Other modern and promising approaches integrate mobile phone data to estimate poverty [5-8] but pose additional questions about scalability across countries. In this study, we use a similar approach as presented by [8], conflating data/products from satellite data and mobile phone traffic to predict poverty using a Bayesian geostatistical model, with an initial focus of Senegal as a pioneer country. According to the World Bank, Senegal is ranked as a low-income developing country, and recent projections indicate that the progress in poverty reduction has been rather modest, and that Senegal continues to display high rates of monetary poverty. Statistics from the Word Data Lab also reveal that 32.4% of the population live in extreme poverty [9]. As such, creating a spatially explicit and scalable poverty mapping approach for Senegal is highly relevant. Such maps provide much needed data-driven evidence for policy development in climate mitigation by identifying areas with the most vulnerable populations for climate mitigation policy support.

2. Methodology

We use a Bayesian approach to estimate the parameters of the posterior distribution, explicitly controlling for spatial random effects [10]. To estimate the geostatistical model, we apply the Integrated Nested Laplace Regression (INLA) modelling approach [11]. INLA is especially designed to implement latent Gaussian models, which cover a wide set of models, including generalized linear, mixed, spatial and spatio-temporal models [12-13]. The algorithm is a deterministic and computationally effective alternative to the Markov Chain Monte Carlo (MCMC) simulation methods that are commonly used for Bayesian inference. The latent Gausian model that we estimate can be summarized as follows. The observations y_i on locations $s_i, i = 1, \dots, n$ are assumed to belong to a distribution family that can be linked to a structured additive linear predictor eta_i by means of a link function $g(\cdot)$, such that $E(y_i) = g^{-1}(\eta_i)$. The linear predictor η_i is defined as follows:

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j z_{ij} + \sum_{l=1}^{L} f_l (u_{li})$$

where β_0 is the overall intercept, the coefficients β represent the linear effect of the fixed covariates z and f is a collection of functions of the coveriates u that represent the random effects. As the DHS values are normally distributed by approximation, we use a Gaussian link function, which assumes that $E(y_i) = \eta_i$. Currently, we use a simple function that only includes spatial random effects and a limited set of covariates, including remote sensing data and mobile phone traffic information.

Assuming conditional independence of observations y, The likelihood function is written as:

$$y \mid x, \theta_1 \sim \prod_{i=1}^{n} \pi (y_i \mid x_i, \theta_1)$$

where the x is a Gaussian Markov random field (GMRF) defined as joint distribution $x = (\eta, \beta_0, \beta, f)$, which is controlled by a set of hyperparameters θ_1. To model spatial processes, the GMRF is combined with a stochastic partial differential equation (SPDE) [14]:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau x(s)) = \mathcal{W}(s) \qquad s \in \Omega$$

where Δ is the Laplacian, α controls the smoothness, $\tau > 0$ is the spatial scale parameter, τ controls the variance, $\mathcal{W}(s)$ is the Gaussian spatial white noise process and Ω is the spatial domain. The model described above is fit using the R-INLA package [15-16].

Our approach resembles that of [8], who investigates the explanatory power of mobile phone data and remote sensing information on poverty in Bangladesh. An important difference is that they use Voronoi polygons that approximate mobile tower coverage as the unit of analysis, while averaging the underlying high-resolution remote sensing information [8]. However, the size of the polygons differs widely and can be quite large in rural areas, where mobile tower density is low. As a result, explanatory environmental variables such as aridity, precipitation and elevation are likely to substantially vary for these areas. Predicting poverty indicators using these values might therefore not be representative of the total area. To accommodate this issue, we rasterized the Voronoi polygon and implemented the model at the resolution of 1x1 km grid cells. As such, poverty projections can be made at the highest resolution, while maintaining tower level data at the level of Voronoi polygons.

3. Data resources and variables

3.1 Poverty indicator

We use the 2012 - 2013 Senegal DHS-based wealth index as our primary indicator of poverty. As part of the DHS, information is collected on various household assets, including ownership of land, livestock and transportation vehicles, and housing characteristics, such as number of rooms and roof material. Principal component analysis is used to determine the weights of the assets in order to calculate a composite wealth index. Only the first principal component, which explains the largest percentage of variance, is used to construct the index.[1] For the analysis we use the average wealth index for the enumeration areas which were sampled by the DHS. Wealth index scores range from -12.43 to 24.44, where higher scores imply a higher socioeconomic status. The DHS covers 200 geocoded enumeration areas (Figure 1). However, for confidentially reasons the GPS coordinates of all areas are presented with an offset, which varies from 2km for urban areas to 10km for rural areas. To control for this effect, we buffered the location of the enumeration areas accordingly and calculated the average of all covariates for the buffered area.

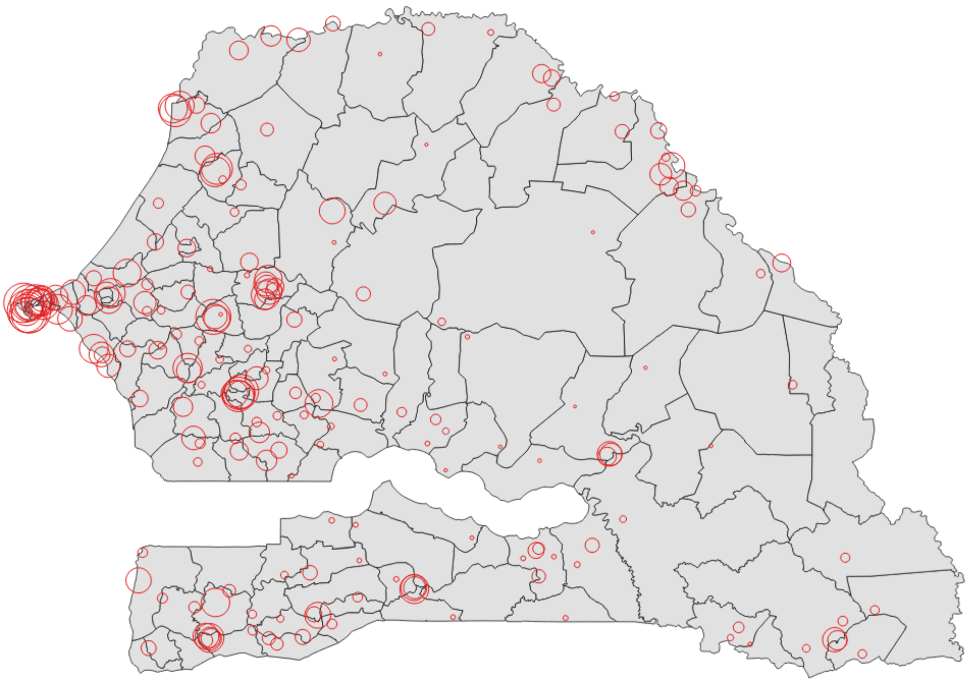

Figure 1: Distribution of the 200 enumeration areas where DHS were conducted in Senegal. Radius of circle is representative of the wealth index scores ranging from -12.34 to 24.44.

[1] The DHS presents both the wealth index and a five point scale, which is determined by assigning each household to a quintile in the distribution. We only use the continuous wealth index for this analysis.

3.2 Covariates

A wide number of factors are expected to be correlated with observed poverty levels. We broadly distinguish between two sets of covariates: a) Remote sensing information, and b) high frequency mobile phone traffic data. Some key covariates for describing socioeconomic conditions, acquired from satellite data, include the population distribution (Figure 2) and nighttime lights (Figure 3). Additional remote sensing information, including environmental variables such as elevation and precipitation are elaborated in Table 1. Furthermore, in this study we analyzed the mobile data acquired from the 1666 towers distributed across Senegal, provided by Orange (Figure 4). Emphasis was placed on key variables such as number of outgoing calls, duration of calls and entropy of calls, where entropy is a measure of the network variability of different towers contacted from a given tower (Table 2).

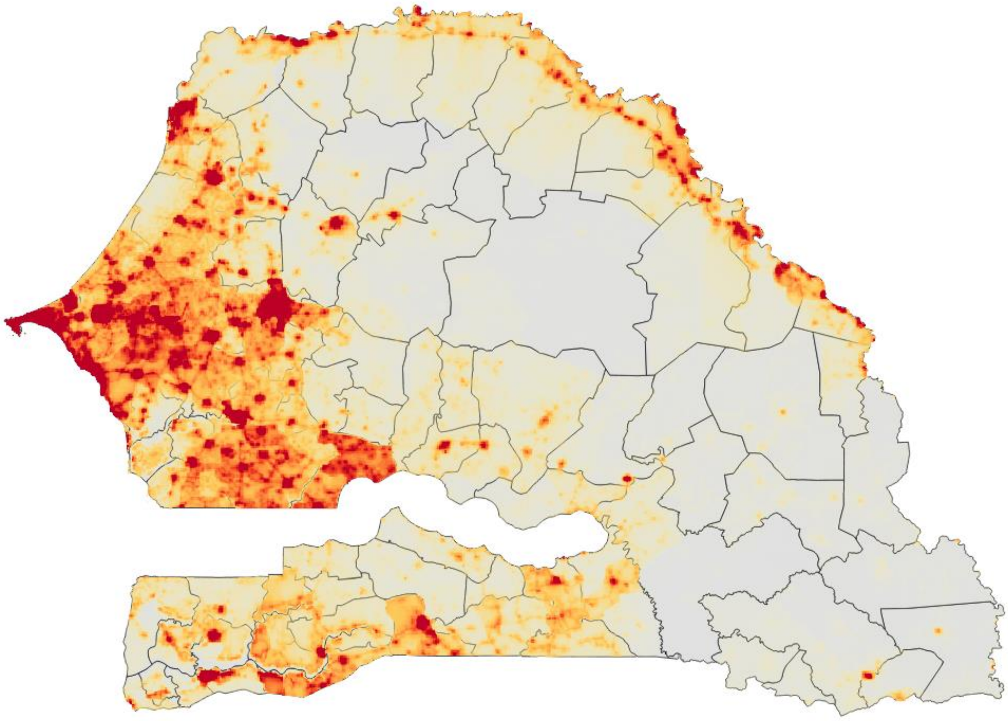

Figure 2: Population distribution of Senegal (from Worldpop)

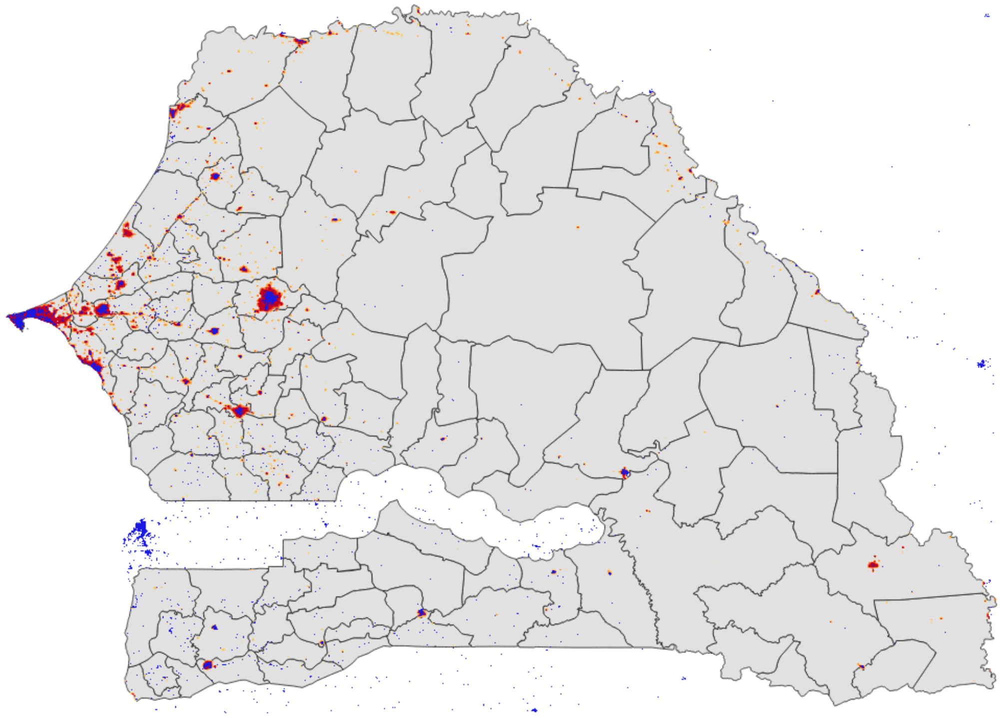

Figure 3: Nighttime lights from the VIIRS satellite indicated with yellow-red pixels. Blue pixels outline human settlements acquired from Global Urban Footprint (GUF) data from DLR Tandem-X.

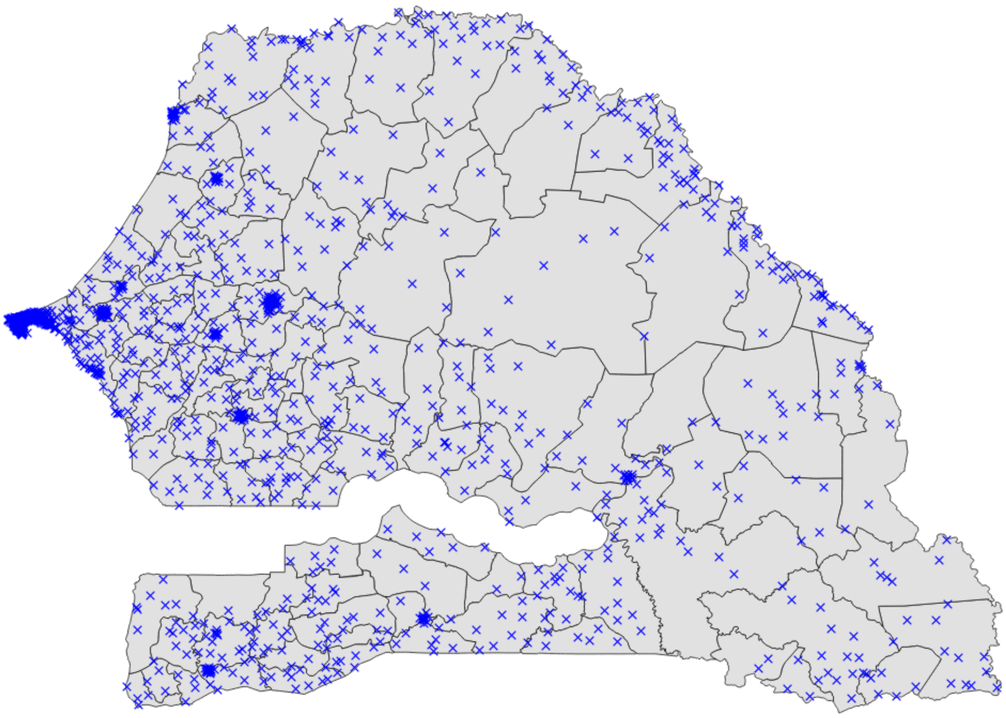

Figure 4 Distribution of the 1666 mobile towers in Senegal

Table 1: Remote sensing variables

Category	Description	Source	Resolution	year
Anthropogenic **(access)**	Accessibility to populated places with more than 50k	European Commission Joint Research Centre (http://forobs.jrc.ec.europa.eu/products/gam/)	0.0833 deg	2000
Demographic **(pop)**	Population count	WorldPop (http://www.worldpop.org.uk/)	0.0833 deg	2010
Topographic **(srtm)**	Elevation	CGIAR-CSI (http://srtm.csi.cgiar.org/)	0.0833 deg	2008
Climate **(pet)**	Average annual Potential Evapotranspiration [mm]	CGIAR-CSI (http://srtm.csi.cgiar.org/)	0.0833 deg	1950-2000
Climate **(ai)**	Mean Aridity Index	CGIAR-CSI (http://srtm.csi.cgiar.org/)	0.0833 deg	1950-2000
Nighttime lights **(viirs)**	VIIRS night-time lights [W cm-2 sr-1]	NASA Suomi NPP (https://viirsland.gsfc.nasa.gov/index.html)	1kmx1km	2015

Table 2: Mobile phone variables

Category	Description	Source	Resolution	year
Basic phone usage **(outgoing_calls)**	Number of outgoing calls	Orange	tower	2013
Basic phone usage **(dur_out_per_call)**	Average duration of outgoing calls	Orange	tower	2013
Network **(outgoing_calls_entropy)**	Outgoing calls entropy	Orange	tower	2013

4. Results

Prior to implementing our model, we conducted an exploratory analysis of the relationship between the DHS wealth scores and the explanatory variables. Figure 5 compares the wealth index with each of the explanatory variables used in this study. The blue line represents the result of a loess regression, which is a non-parametric approach where least squares regression is performed in a local neighbourhood. The grey area is the 95% confidence region. To account for the skewed distribution, we log transform the viirs, population, duration of calls and number of outgoing calls indicators. The figure shows a positive relationship between the wealth scores and a number of the indicators, in particular population (0.84), viirs (0.77) and number of outgoing calls (0.73). This is confirmed by Figure 6, which presents the correlation across all variables.

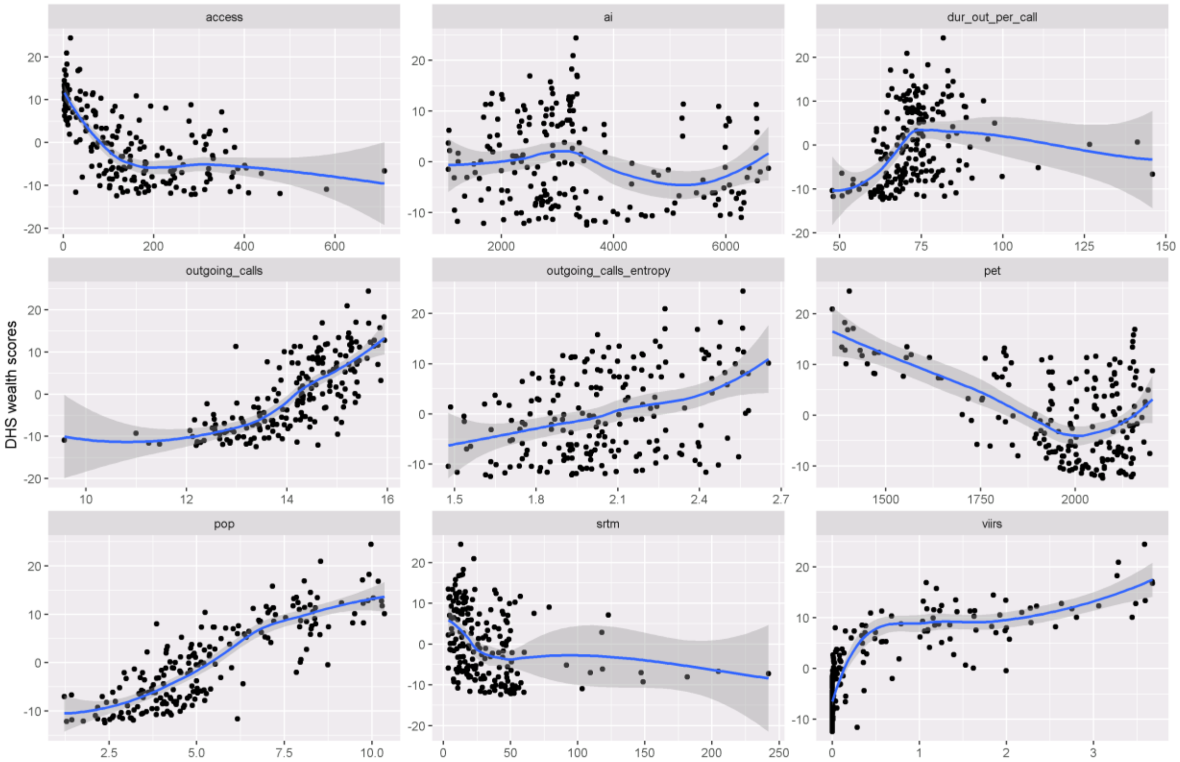

Figure 5: Relationship between DHS wealth scores and explanatory variables derived from remote sensing and mobile phone data

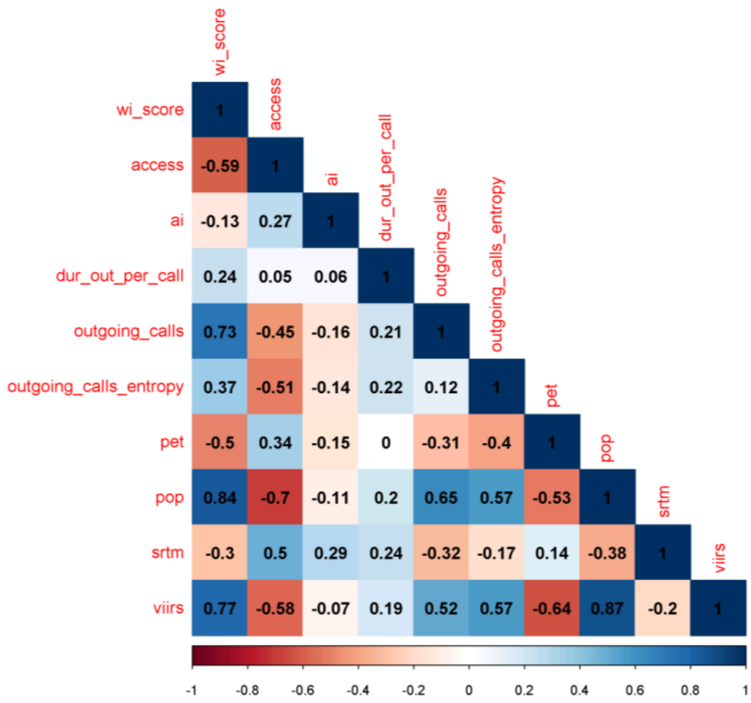

Figure 6: Correlation across the tested explanatory variables

Posterior estimates for the mean and the 95% credibility interval for our model are presented in Table 3. In particular the *viirs*, *pop*, *number of outgoing calls*, and a*verage duration per outgoing call* have a positive correlation with the wealth index scores, while *outgoing calls entropy* has a negative association with wealth index. The other variables are not significant from zero. Figure 7 depicts the predicted versus observed wealth scores.

Table 3: Model parameters

	mean	0.025quant	0.975quant
intercept	-24.94	-40.47	-9.39
access	-0.01	-0.01	0.00
ai	0.00	0.00	0.00
viirs	1.76	0.34	3.17
srtm	0.01	-0.02	0.03
pet	0.00	-0.01	0.00
pop	1.81	1.11	2.52
outgoing_calls	2.00	1.25	2.76
dur_out_per_call	0.06	0.01	0.11
outgoing_calls_entropy	-4.55	-7.55	-1.56

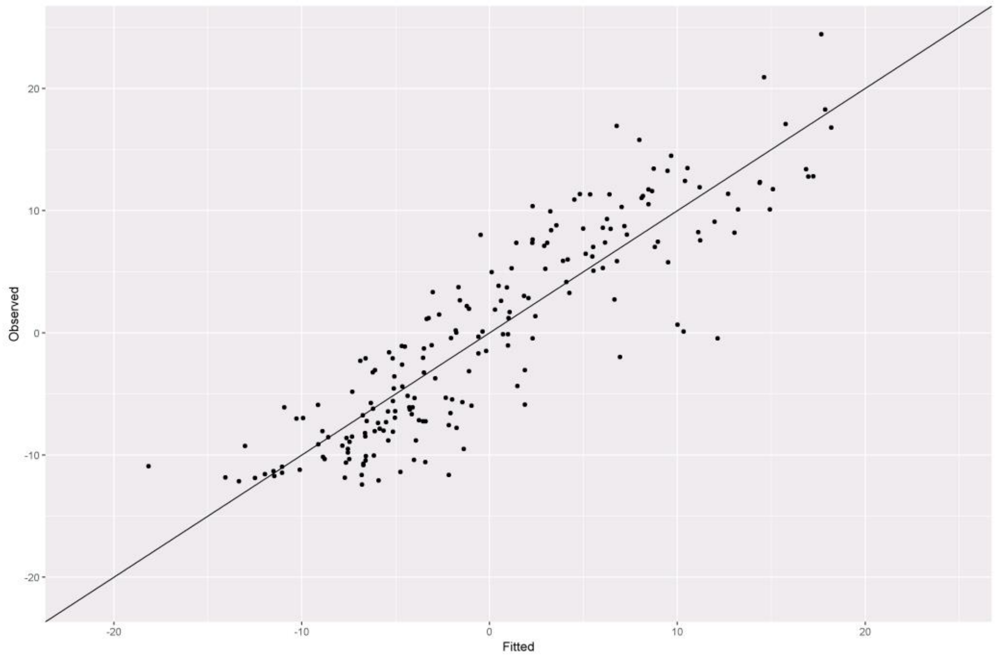

Figure 7: Model predicted versus observed wealth scores

Combining the spatial data layers for the remote sensing data and the rasterized voronoi diagram for the mobile phone data, we can create a poverty map for all of Senegal. The results are shown in Figure 8, while Figure 9 presents the standard deviations.

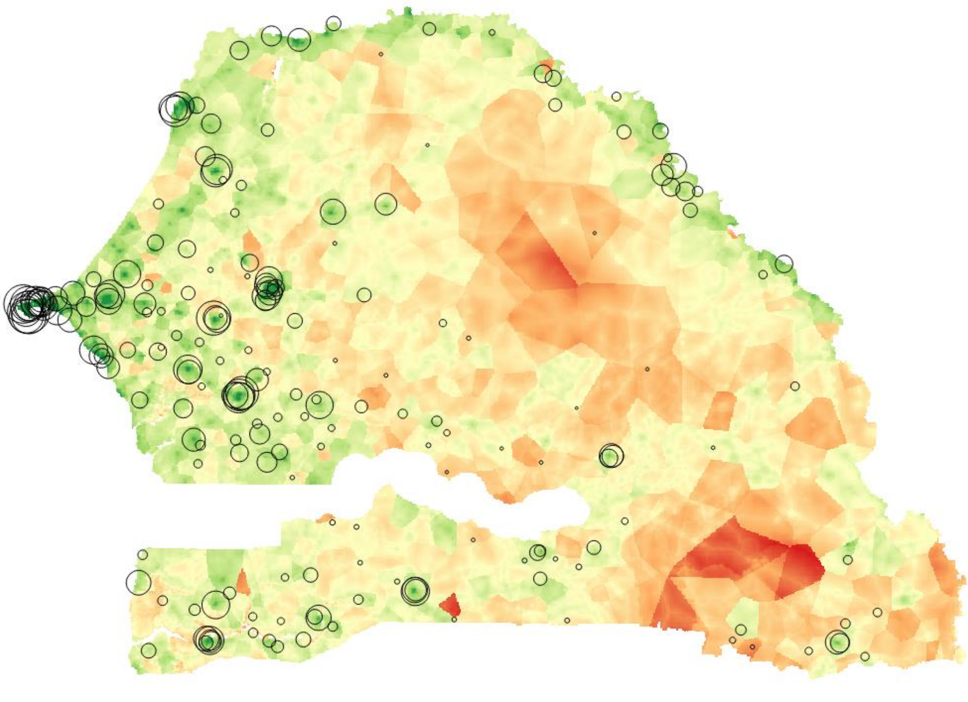

Figure 8: Fine-scale (1x1km) poverty map for Senegal, where predicted wealth scores range from -24.9 (red; low wealth) to 19.5 (green; high wealth). Superimposed on map are the observed DHS wealth scores (circles)

where circle radius is proportional to wealth

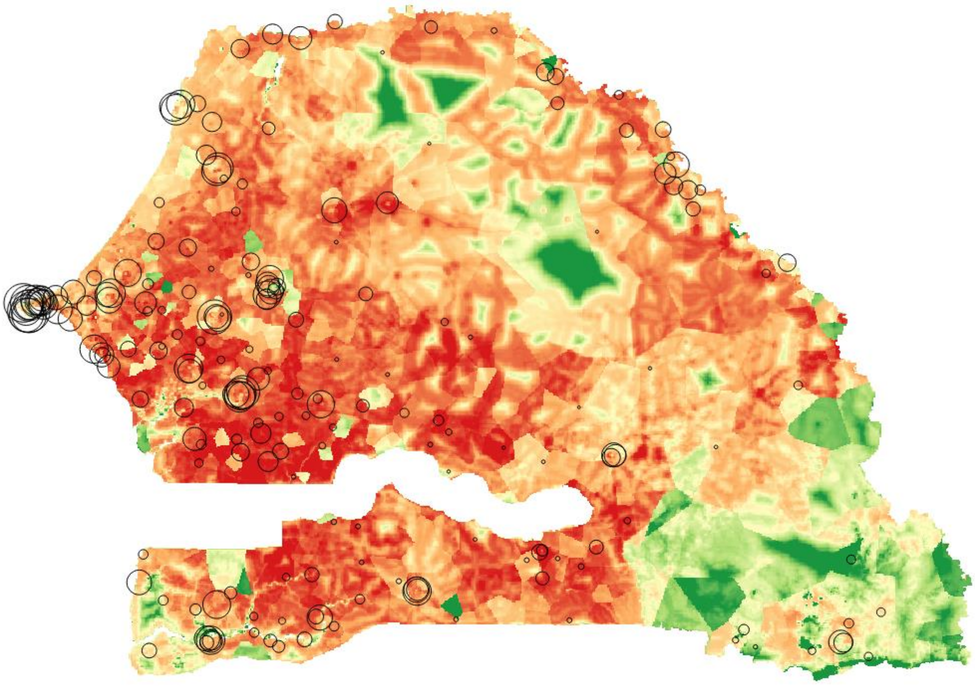

Figure 9: Standard deviation map of predicted fine-scale (1x1km) poverty map for Senegal, where values range from 0.54 (red) to 1.9 (green). Superimposed on map are the observed DHS wealth scores (circles)

5. Discussion

The Bayesian geostatistical model shows that remote sensing-based and mobile phone data-based variables have explanatory power. The derived maps are the initial steps to identify the poorest households in Senegal that are highly susceptible to climate change and other shocks. It is important to point out the limitations and sensitivities of this study. The DHS survey for Senegal only covers 200 enumeration areas, which is a relatively low number of data points for estimating a geostatistical spatial model.[2] Not surprisingly, we only find that spatial random correlation is not significant in our model. Another problem is the spatial offset for the DHS enumeration area locations. Although we controlled for this by buffering the sample locations, it potentially creates a bias in the analysis.

[2] The 2010-11 DHS Survey for Senegal included 400 enumeration areas. These data can unfortunately not be combined with the 2012-13 survey, and we currently have no information on why the number of areas has been halved.

6. Conclusion

Organizing household surveys to measure poverty is a very costly and tedious undertaking, in particular if large parts of a country, are relatively inaccessible. This study shows that it is possible to create a poverty map by conflating different data sources that can be acquired at relative low costs. Remote sensing information, specifically environment and climate related data, is increasingly provided at high resolution by national and international research institutions at no cost. Mobile phone data is not easily available yet but, as illustrated by this Data for Climate Action Challenge, mobile phone operators, such as Orange are more and more willing to contribute data for scientific purposes. Nonetheless, nationally representative poverty indicators drawn from household surveys remain necessary to estimate the structural model that forms the basis for the spatial socioeconomic monitoring, visualization and forecasting. Leveraging the capacities of satellite imagery and big data resources, such as mobile phone records, this study delivers a preliminary spatially explicit, fine-scale (1x1km) poverty map of Senegal.

As a result of our participation in this challenge, the team has also identified a series of future research activities with vast potential to transform the field of fine-scale socioeconomic monitoring. Firstly, we aim to continue this work to derive additional explanatory variables from satellite data with spatial resolution of 3-5m with high repeat frequency (i.e. Planet). There are ongoing efforts to derive poverty proxies from very high-resolution satellite data (i.e. < 1m) to serve as calibration and validation data for the 3-5m imagery. These variables will be integrated into our Bayesian geostatistical model to develop more robust maps with scaling potential. Furthermore, the team has tested an innovative microtasking tool that can be crowdsourced to volunteers to help visually assess satellite imagery and generate validated poverty maps.

References

1. Veljanovski et al. (2012) Object-based image analysis of VHR satellite imagery for population estimation in informal settlement Kibera-Nairobi, Kenya. *Remote Sensing – Applications* Escalante, B. Ed.; InTech; Rijeka, Croatia; pp 407-434
2. Taubenbock H., Kraff N.J. (2014) The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data. J. Hous. Built Environ. 29, 15-38.
3. Tatem A.J., Gething PW, Pezzulo C., Weiss D., Bhatt S (2014) Development of high-resolution gridded poverty surfaces. See http://www.worldpop.org.uk/reosources/docs/Poverty-mapping-report.pdf
4. Jean N et al. (2016) Combining satellite imagery and machine learning to predict poverty. Science 353, 790-794.
5. Blumenstock J., Cadamuro, G. (2015) On R. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073-1076
6. Smith-Clarke C., Mashhadi A., Capra L. (2014) Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks. Proc. Of the SIGHI Conf. on Human Factors in Computing Systems, Toronto, Ontario, Canada, pp 511-520.
7. Pokhriyal N., Dong W., Govindaraju V. (2015) Virtual networks and poverty analysis in Senegal. In Data for Development Senegal Challenge.
8. Steele, JE et al. (2017) Mapping poverty using mobile phone and satellite data. J. R. Soc. Interface 14: 20160690 http://dx.doi.org/10.1098/rsif.2016.0690
9. World Data Lab (2017) http://worldpoverty.io/index.html
10. Diggle PJ., Ribeiro PJ. (2007) Model-based Geostatistics (Springer Series in Statistics). Vol. 1. New York: Springer. doi:10.1111/1467-9876.00113.
11. Rue H., Martino S., Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society. Series B: Statistical Methodology 71 (2): 319–92. doi:10.1111/j.1467-9868.2008.00700.x.
12. Martins TG., Simpson D., Lindgren F., Rue H. (2013) Bayesian computing with INLA: New features. Computational Statistics & Data Analysis 67 (November): 68–83. doi:10.1016/j.csda.2013.04.014.
13. Rue H et al. (2017) Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application* 4 (1). Annual Reviews: 395–421. doi:10.1146/annurev-statistics-060116-054045.
14. Lindgren F., Rue H., Lindström J. (2011) An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society. Series B: Statistical Methodology 73 (4): 423–98. doi:10.1111/j.1467-9868.2011.00777.x.
15. Blangiardo, M., Cameletti, M. (2015) Spatial and Spatio-temporal Bayesian Models with R-INLA. Vol. 7. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9781118950203.
16. Lindgren F., Rue H. (2015) Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software 63 (19): 1–26. doi:10.18637/jss.v063.i19.