Center for International Earth
Science Information Network
Earth Institute | Columbia University

Estimating and Mapping Populations in Off-Grid and Hard-to-Reach Places

Final Report

March 2020

Estimating and mapping populations in off-grid and hard-to-reach places

Lead organization: The Center for International Earth Science Information Network (CIESIN/Columbia University)

Collaborators: The Connectivity Lab at Facebook, Wetlands International Africa

Sustainable Development Goals

SDG 1 - No poverty

SDG 3 - Good health & well-being

SDG 4 - Quality education

SDG 13 - Climate action

SDG 17 - Partnership for the goals

Countries: Sierra Leone, Liberia, Guinea, Senegal, Cote d'Ivoire, Ghana

Data:
Administrative data
Commercially licensed data
Geospatial data
Official statistics
Open data
Satellite imagery
Survey data
Web or social media data

Project Objective: To increase the visibility of off-grid, hard-to-reach populations, living in mangrove areas of coastal West Africa, to various development and adaptation actors by providing information about the location and size of the small settlements/villages through an online interface, taking advantage of very high resolution satellite images, Volunteered Geographic Information (VGI), and modeling.

This report was prepared by Sylwia Trzaska and Greg Yetman, with inputs from Kytt MacManus, The Center for International Earth Science Information Network (CIESIN)

Acknowledgments

This project, *Estimating and mapping populations in off-grid and hard-to-reach places (Sierra Leone, Liberia)*, submitted in response to the 2017 call for proposals by the World Bank's Development Data Group (DECDG) and the Global Partnership for Sustainable Development Data (GPSDD), is supported by the World Bank's Trust Fund for Statistical Capacity Building (TFSCB) with financing from the United Kingdom's Department for International Development (DFID), the Government of Korea, and the Department of Foreign Affairs and Trade of Ireland.

Table of Content

Acronyms

AGOL - ArcGIS Online

CIESIN - The Center for International Earth Science Information Network

ESRI - Environmental Systems Research Institute

GIS - Geographic Information Systems

HRSL - High Resolution Settlement Layer

MAHRV - Mapping Hard to Reach Villages

SDG - Sustainable Development Goals

VGI - Volunteered Geographic Information

Executive Summary

Estimating and mapping populations in off-grid and hard-to-reach places contributed to filling the gaps in spatial coverage of population distribution and to increasing the visibility of off-grid, hard-to-reach populations to various development and adaptation actors. Such populations are often underrepresented and underserved in development projects, lack access to basic services including sanitation, health and education, and lag behind on achievement of the Sustainable Development Goals (SDGs). To achieve its objective the project built a publicly accessible portal -- *Mapping Hard to Reach Villages* (MAHRV) -- allowing anyone to contribute and access population data. The project focused on people living in mangrove areas of coastal West Africa.

The project's methodology innovatively combined the analysis of high resolution satellite imagery, Volunteered Geographical Information (VGI) and statistical modeling. Villages and their extents were identified from the High Resolution Settlement Layer (HRSL), taking advantage of machine learning approaches to remote sensing problems. Village extents and settlement-level population data, contributed through the VGI, were used in statistical modeling in conjunction with other covariates to further infer population counts in villages for which *in-situ* data were not available. Both processes were made publicly accessible through the MAHRV portal. Population estimates will be iteratively improved as more *in-situ* data become available through the VGI data collection process. Data validation and model updates run in the background.

ArcGIS Online (AGOL) was selected as the primary technology for the Online Services, Web Applications and Mobile Applications, as it is supportive of easy technology transfer. The framework of software and services allows for the update of the population model and public data product with a minimum of effort. The services can be maintained in the future without additional funding.

Overall, the new interface allows an increase in spatial resolution of village extents, thus the identification of very small villages, and provides population estimates at a granularity never before achieved at regional scale (most of West Africa mangroves). While this project focused on villages within the mangroves in West Africa, the approach and the technologies are easily transferable to other geographies.

1. Introduction

1.1 The background

Estimating and mapping populations in off-grid and hard-to-reach places is a 'Leave no-one behind' project supported by the World Bank Trust Fund for Statistical Capacity Building. 'Leave no one behind' implies reaching populations that are underrepresented and underserved in development projects. People living in hard-to-reach areas, such as remote riverine regions, natural protected areas and high elevation areas, deserve special attention as they are often under-reported by traditional data sources and still missed by cell phone data and night-time lights records. Providing information about the location, size and accessibility of very vulnerable, unaccounted populations to various agencies is a prerequisite to achieving sustainable development. Indeed, the populations living in hard-to-reach places often lack access to basic services including sanitation, health, and education. These populations lag behind on achievement of SDGs, yet are often excluded from interventions.

Attempts to map and estimate population size in these off-grid, hard-to-reach areas have often been ad-hoc or poorly documented (Shaghaghi *et al.*, 2011). To date there are no well-established methodologies to provide disaggregated information to help governmental and non-governmental organizations better target development and adaptation planning for hard-to-reach populations and achieve SDGs related to climate action, poverty reduction, and improving access to basic services such as clean water and sanitation, healthcare, and education.

This project aimed to increase the visibility of off-grid, hard-to-reach populations, living in mangrove areas of coastal West Africa, to various development and adaptation actors by providing information about the location and size of the small settlements/villages through an online interface, taking advantage of very high resolution satellite images, Volunteered Geographic Information (VGI), and modeling.

1.2 The approach

The project's main objective was to address gaps in spatial coverage of population distribution and to provide information about in hard-to-reach areas necessary to achieve selected SDGs. It targeted areas where other methods of locating and estimating population are not available thus contributing to advancing data production. By combining satellite data analysis, available and VGI-collected *in-situ* village data and statistical modeling it offers an alternative to systematic surveys specifically directed at small settlements in hard-to-reach areas that may be unsustainable in low-income countries with limited human and data resources and capacities.

To achieve its objective the project built a publicly accessible web application allowing a large number of end-users to have access to population estimates in the mangrove areas of West Africa at different scales. To produce the estimates the team innovatively combined analysis of high resolution satellite imagery, VGI approaches and statistical modeling. CIESIN has partnered

with the Connectivity Lab at Facebook to produce the High Resolution Settlement Layer (HRSL[1]) taking advantage of advances in the application of machine learning to remote sensing problems. Population counts were estimated through statistical modeling using settlement-level population data obtained through VGI and village extents derived from HRSL structures. The estimates will be iteratively improved as additional data are collected and stored through the VGI online capability. Estimated and *in-situ* recorded information about village population are publicly available through an on-line mapping interface and for download. While starting in mangrove areas of West Africa, the approach and tools developed can be easily transferred to other remote areas in need of filling the spatial population coverage gap.

The project can be summarized as follows:

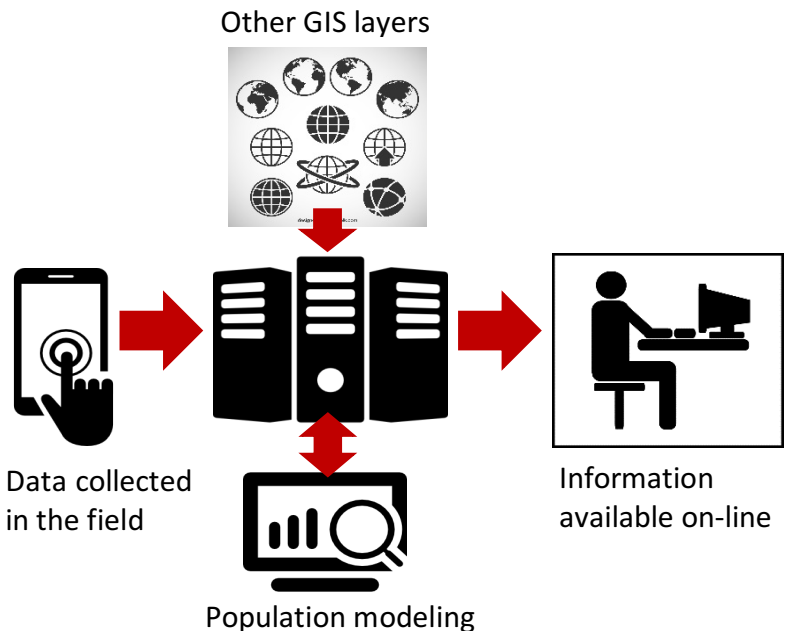

Figure 1: Schematic representation of the approach to generating on-line information about the populations living within the mangroves in West Africa.

The next three sections present the final interface, its main pillars - the VGI and population modeling- and the project implementation. The following two sections describe the challenges and provide final conclusions. The annexes include the full VGI survey the list of covariates used in the Random Forest model and Risk mitigation strategies and output indicators.

[1] See: https://data.humdata.org/dataset/highresolutionpopulationdensitymaps

2. Mapping Hard to reach Villages public interface - MAHRV

To increase the visibility of hard-to-reach villages through information gathering and display the project built a web-based interface or portal: Mapping hard to reach Villages or MAHRV.

https://mahrv.ciesin.columbia.edu/home

The portal is deliberately very simple to allow easy access in low band-width conditions and consists of three pages:

- Home page with brief description of the project and links to portal guide, project report (forthcoming) and to the other pages (fig.2)
- Information contribution page – for those who can collect and/or want to upload population data for villages in the mangroves of West Africa
- Information display and download page – for those who want to obtain information/data on village populations living in the mangroves of West Africa.

Figure 2: Screen capture of the home page of the public portal. The interface has been named: MApping Hard to Reach Villages – MAHRV.

2.1 Contribute Information

This page is designed to allow the upload of population data collected by various contributors who travel to mangrove villages. Whether they live in such a village or travel there for a project, or any other reason, they can contribute very valuable knowledge of the village – its population but also infrastructures etc. Volunteered Geographical Information (VGI) process, also known as 'crowdsourcing', allows to harvest such knowledge and contribute to the general understanding

of the vulnerabilities of the people living there and of the hurdles in achieving many of the SDGs.

To contribute information, a data collection device (cellphone or laptop) and a connection to internet (in the field or in the office) are needed. The process and technical characteristics of the data collection interface are described below.

2.1.1 Information contribution process

- Registration

Information contribution starts with a registration – a one-time request to submit data - as a contributor with minimum information such as the name, email address and institution. After being approved contributors can upload data with a user name that will be provided to them.

- Data contribution

There are several options for data contribution on- or off-line. The interface is currently available in English and in French and consists of:

- The AGOL interface for mobile and desktop data collection
- Survey software consisting of a pre-populated Survey123
- Off-line capabilities to allow the use of the application in remote areas with no internet connection and related syncing capability
- Capability to manually enter data collected in the field

The page also contains on-line Field Data Collection Guidebook, that explains the data collection and upload process step-by-step, also available for download as PDF. In all cases, we expect that the information is collected and contributed to the portal with the knowledge and consent of the populations themselves and that the data are collected through the Survey 123 and uploaded or manually input into the online office-based data collection interface.

2.2 Get Information

This page is designed to access information about villages located in the mangroves of West Africa. The information can be accessed in two formats:

- An interactive *map application* for viewing and querying data allowing to see all the villages within mangrove areas
- *Data download* as a comma-separated format file and Shapefile

2.2.1 The interactive mapping application

The mapping application allows to visualize the location and access the population estimates of the villages within the mangroves. It displays four layers: village location and extent, mangrove extents and district level boundaries. Only villages within or up to 100m from the mangroves are displayed. Village information can be accessed by clicking on the village. By clicking outside of the village information at the district level is displayed. The controls on the upper left side allow to zoom in and out and change the legend display.

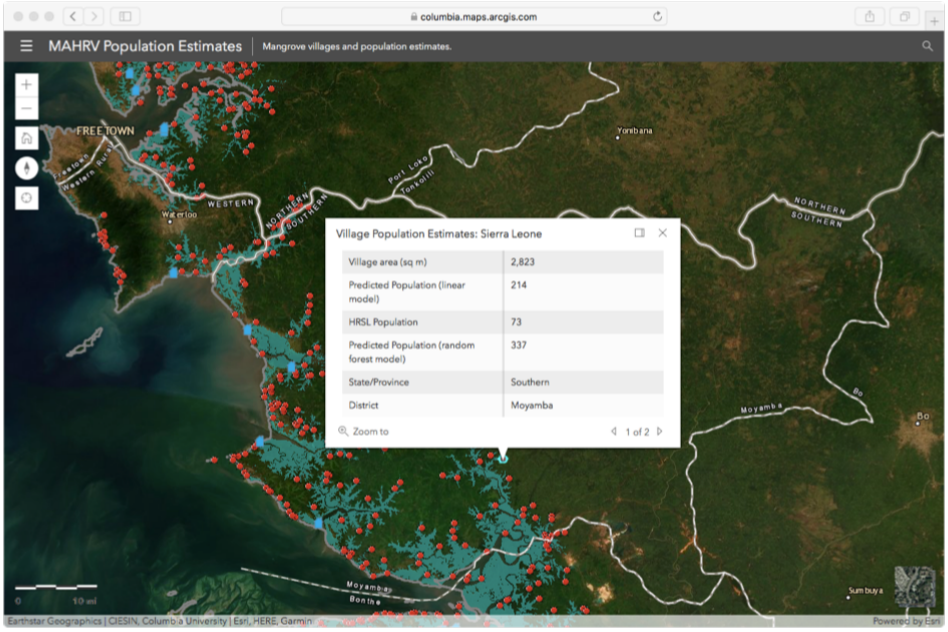

Figure 3: Screen capture of the Mapping interface displaying a pop-up information for a village.

2.2.2 Data download

The data can be downloaded as a zip archive which contains a file Geodatabase, shapefile, and Comma-Separated Values format for:

- Visited villages – data contributed through the VGI (point features)
- Village extents derived from the High Resolution Settlement Layer (polygon and point feature class), with population estimates from the HRSL, a linear population model, and a random forest population model.

Figure 4: Screen capture of the download page

3. The pillars supporting MAHRV

3.1 Volunteered Geographic Information

VGI is the main process allowing the collection of new data volunteered by various contributors. The collection and archiving of the data is supported through an on-line interface, a survey form allowing easy upload of the data and an off-line capability to account for inexistent or unreliable internet connection. Interface

The VGI interface is supported via the Arc GIS On Line (AGOL). It offers a number of possibilities for mobile and desktop data collection. In all cases the physical data resource is stored in a cloud-based Hosted Feature Layer supporting vector feature querying, visualization, and editing and having the capability to limit the possibility for data collection errors and maintain dataset integrity.

3.1.1 Survey software

The interface to collect data in handheld devices consists of a pre-populated Survey123 app because of its multi-language support and ability to support faceted responses where the response from one question impacts the possible responses of future questions. Figure 3 shows the interface.

Data collected in Survey123 forms can be uploaded directly from the field if internet is available (see 'Collect/Enter data from the field') or synced later from any other location where internet is available (see 'Enter field data from the office' and explanations below).

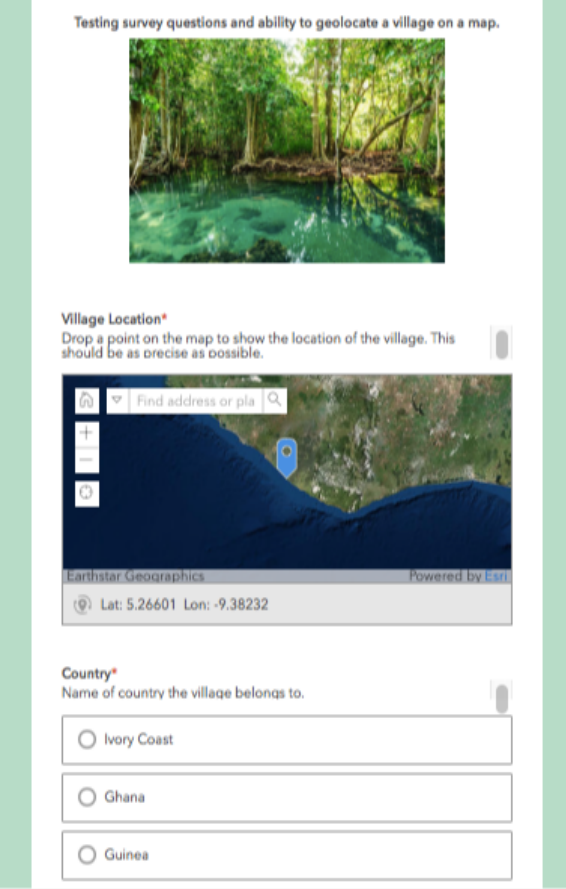

Figure 5: Screenshot of the VGI data collection interface.

3.1.2 Off-line capabilities

Because we anticipate that the application will often be used in areas with no access to internet and maps, we offer the possibility to identify the location of the village on a map through the Baseline map in off-line mode. However, the volume of very high resolution data precludes storage and use of the full dataset for West Africa on a simple mobile device. Thus, quarter degree tiles were preprocessed from the ArcGIS Imagery Basemap into tile packages for the coastal study area that can be downloaded individually as needed. These TPKs (~100GB) are hosted in the CIESIN repository and are accessible via a webmap on AGOL. Figure 4 shows and example of the tiles developed for the application.

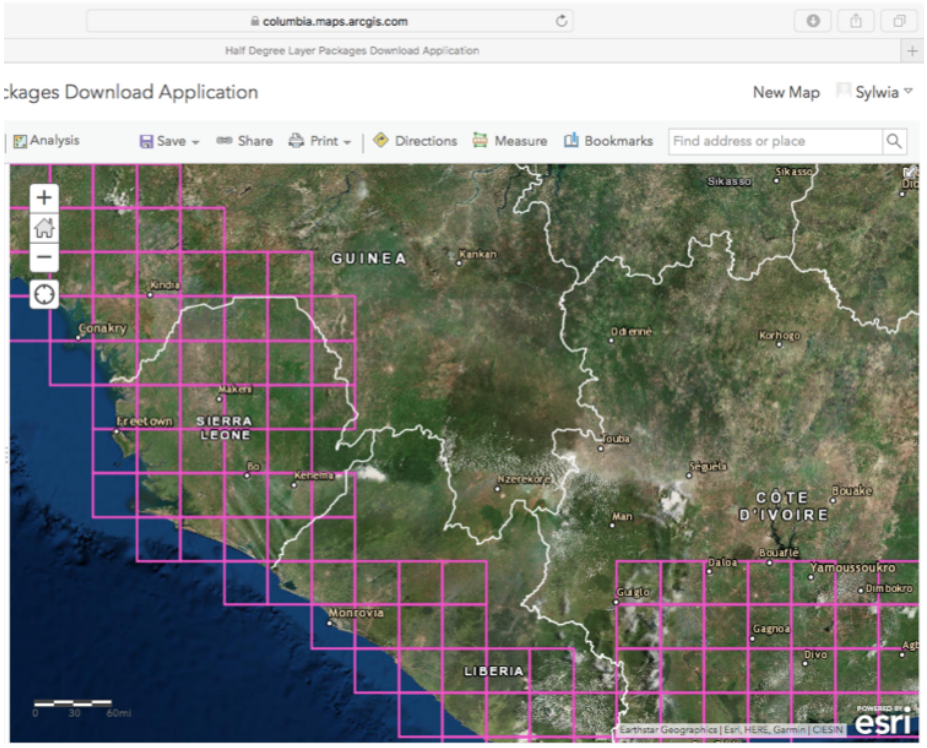

Figure 6. Map of the tiles developed for off-line use of the application

- Enter field data from the office

Data collected with the Off-line capabilities of Survey123 can be synced at any time when an internet connection becomes available. Additionally, a browser-based Office VGI Data Entry application is available so that users might enter manually transcribed field data if needed.

3.1.3 VGI data quality control

As new data are contributed through the VGI, their quality will be checked through comparison of population data to the estimates and to the *in-situ* recorded data from villages of similar extent.

3.2 Population models

3.2.1 Methodology

The process uses data collected and contributed through VGI to build population estimates in villages for which no data has been collected yet through statistical modeling. The approach capitalizes on advances in the application of machine learning to remote sensing problems by using the HRSL, jointly developed by CIESIN and the Connectivity Lab at Facebook. Population estimates for the villages with no *in-situ* data were obtained through three modeling approaches, all relating population counts to village size, and various covariates found in other models. Three methods were used to date:

- Direct HRSL estimate: average population density is derived from the total population and total villages extents within a given census unit; individual village population is obtained by using the average population density and the village extent.
- Linear model: based on a relationship between population recorded in villages visited and their extents from HRSL; this relationship is then used to estimate village population based on village extent.
- Machine learning (random forest) model: uses established relationships between population distribution and multiple biophysical characteristics; a random forest model was created using the available *in-situ* population data and standard biophysical covariates.

Codes and outputs for the linear model and the random forest model are publicly available at: https://github.com/gyetman/mahrv (check README file first).

3.2.2 Results

- Village identification and extents

High Resolution Settlement Layer is used to determine the location and extent of mangrove villages in West Africa. To build the HRSL The Connectivity Lab at Facebook applies state-of-the-art computer vision techniques to high resolution (0.5m) satellite imagery from DigitalGlobe (Tiecke *et al.*, 2017) to identify individual buildings. Buildings are then automatically clustered into settlements and village areas can be computed. For the current application only villages with boundaries within a 100 m distance from the mangroves were retained. Mangrove extents are based the Global Forests Mangrove Forests distribution data set (Giri *et al.,* 2005). Further examination of villages showed disjoined settlement patterns and additional clustering was applied, reducing the number of villages from 11,771 to 4,022.

- HRSL-based population estimates

Once villages are identified and extents computed CIESIN used proportional allocation to distribute population from subnational census data. This approach assumes homogeneous population density per area unit and tends to underestimate population of the villages in the mangrove areas of West Africa (figure 9). This estimate is however displayed together with the results from the two other models, to highlight that the use of global products that are not necessarily calibrated locally may lead to large discrepancies between population estimates and observations.

- Linear model-based population estimates

A simple linear regression model was built relating population estimates of visited villages (n=16) and the area of those villages based on the HRSL layer.

Examination of the data show a skewed distribution for both population and area. The log of population is used for the y variable as it is better distributed, and the area is standardized using a standard scaler (transformed so that mean=0, with a standard deviation of 1). The resulting relationship is shown below. Confidence levels are plotted on the chart; however, a higher number of samples are required for the confidence interval to be dependable (generally, a minimum of 30 samples are required).

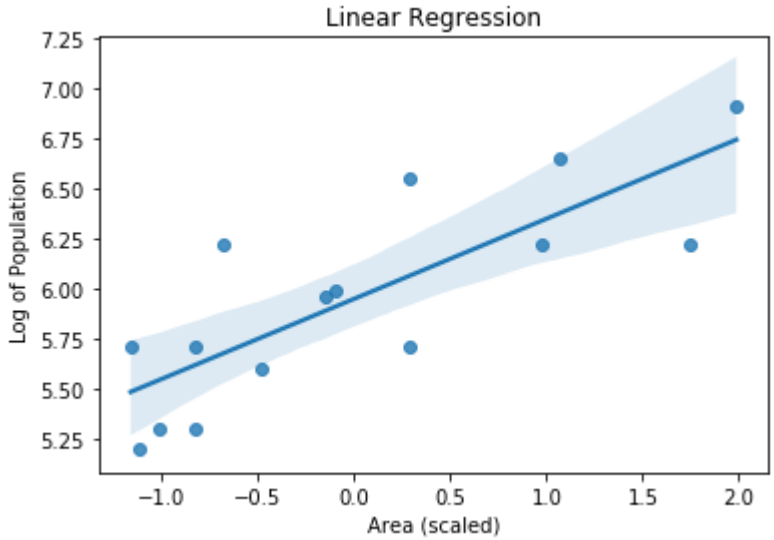

Figure 7: Linear model of the log of population and village area (scaled). The blue shaded area represents a 95% confidence interval, the input data for the model are plotted as blue points (n=16). Equation is y = 0.3997 x + 5.9466. R-value is 0.804, p-value is 0.0003

Population Prediction

The linear regression model is used to estimate population in villages with no *in-situ* data, using village extent from HRSL as predictor, after village extent consolidation.

The input data and Jupyter Notebook used to create the linear model are available on GitHub at https://github.com/gyetman/mahrv. The notebook includes additional plots (histograms of the variables) and descriptive statistics. To run the code in the Notebook requires Python, Jupyter notebook, and several open source Python modules (sklearn, scipy, matplotlib). Updated data collected in the field will be used to improve the model as it becomes available. The same notebook code can be used to quickly update the model and predictions with additional data points.

- Machine Learning Model – based population estimates

Past efforts to improve population prediction (Stevens *et al.,* 2015) use established relationships between population distribution and biophysical characteristics (Nieves *et al.,* 2019) to build the model. Using the spatial data on covariates developed for the WorldPop project (Lloyd *et al.,* 2019), with the addition of a measure of distance to mangrove edge, a random forest model was created using the same village population estimates that performed the basis of the linear model. A list of the spatial covariates used in the model and obtained from WorldPop global project[2] is shown in Table 1 Annex 2 together with their correlation with log of population from the sample. The covariates show both positive and negative relationships with population. Interestingly, unlike regional population models such as

[2] https://www.worldpop.org/project/categories?id=14

WorldPop, the distance to infrastructure (roads, buildings) and existing population centers exhibit negative relationships (Table1). Given that the sample is from small isolated villages this is not surprising.

Population Prediction

The Random Forest model is designed to weight covariates depending on the predictive power and reduction of error in the model through iterative leave-one-out testing, where a variable is left out of prediction and the increased error in prediction is measured. The model is fit using default parameters with the village population field data and the spatial covariates. The resulting importance for all predictors is included in the GitHub archive. In the model, the top three predictors are topography, village area, and distance to global human settlement layer and distance to urban footprint combined. These three factors provide the most predictive power in the model.

The model is next evaluated using cross-fold validation, where each input village population estimate is left out in turn and used to estimate error in the prediction. The result shows a consistently high accuracy rating (0.91 average across all folds), but the small sample size may inflate these numbers.

To improve the model, hyperparameter tuning is used to select the best model parameters. Using these parameters, the model was again run with k-folds validation, leaving one village sample out on each run. The overall (average) accuracy with these parameters improved from 0.91 to 0.97. These parameters were selected as the best for the final model, which is used to predict population for all HRSL-derived villages based on the same covariates.

More detailed information about the model and estimates is available at: https://github.com/gyetman/mahrv

4. Project Implementation

4.1 Setting the stage

4.1.1 Technology selection

Several technological solutions were considered, including open source and commercial solutions. Ultimately ArcGIS Online (AGOL) was selected as the primary technology utilized for the development of Online Services, Web applications, and Mobile Applications.

- Framework Selection Justification

The selection of ArcGIS Online Services and Client applications is supportive of technology transfer. Because the technologies are implemented using cloud technologies, the feasibility of transferring to another organization (with unknown institutional knowledge) is increased.

The ArcGIS client services can be customized by knowledgeable IT/GIS professionals, and hosted on local hardware if desired, or simply implemented in the cloud using well known templates. The data services can be downloaded from the cloud for local use, or managed

entirely through an AGOL organization account. The costs of both time and monetary resources to implement highly customized GIS field data collection solutions are often quite high, and the ability to transfer the technologies to others requires that the receiving party has significant expertise in the precise technologies that were used. AGOL offers a more democratic path, where the expertise of a receiving organization can easily be developed through modest training.

Web map templates, available from AGOL, were utilized to create an online user interface for data access and visualization, and for the input of VGI in the office.

- Security of services

AGOL provides tiers of sharing options for all pieces of content. We have created a project-specific group that requires account login for access and only shared content to members of that group (presently, only members of the research team). For the initial phase, users will be vetted and added to the group. Subsequently we may decide to share the services more broadly so that any person with a (free) AGOL account might contribute to the data collection or view the data collection results.

ArcGIS Online Services and client applications

1. Detailed data provided (imagery basemap, with 0.5-2m resolution data)
2. Ability to support multiple inputs
    a. Phone collection
    b. GPX file (GPS data log)
    c. Manual selection on map
    d. Upload of coordinates as text file (csv)
    e. Photos submission not enabled: privacy issue
    f. Offline (no signal) data collection

4.1.2 VGI protocol

- Questionnaire

As a first step to the data collection recruitment script, information text/informed consent and the questionnaire have been developed to ensure that the information is collected in appropriate way. The full informed consent text and questionnaire are provided in Annex 1.

- IRB

The VGI collection protocol, including the verbal informed consent and the questionnaire, as well as a French version, was then submitted to Columbia University Internal Review Board for Protection of Human Subjects in Research (IRB). Data collection process was qualified as ‘exempt’.

4.2 Building the pillars

4.2.1 VGI

- Interface development

Once the technologies were selected, the questionnaire developed and IRB approval obtained, Survey123 app was prepopulated and made available through the AGOL interface. Tools for data download were developed and feedback of various staff in the field solicited.

- Feedbacks and technical adjustments

The main feedbacks were:

Limited use of smartphones in remote areas with handheld devices mostly available during specific field campaigns with often no possibility to add applications. This points to the need to provide the possibility to manually enter the data on a computer, which was added to the interface.

Limited capacity to navigate the AGOL interface and written instructions not sufficient to access and download the application and upload the data – field guidebook was developed with step-by-step instructions based on screenshots. It is available as PDF but also as drop-down menu on the webpage.

4.2.2 Village extents

- HRSL

At the start of the project HRSL layer was only available for 3 countries (Ghana, Senegal Guinea). During the project, The Connectivity Lab @Facebook has completed the HRSL for additional countries: Sierra Leone, Liberia and Cote d'Ivoire.

- Village Clustering

The villages derived from HRSL sometimes had disparate polygons due to the 30m grid cell size and some missed buildings or disjoint settlement patterns. An example is shown in figure 8.

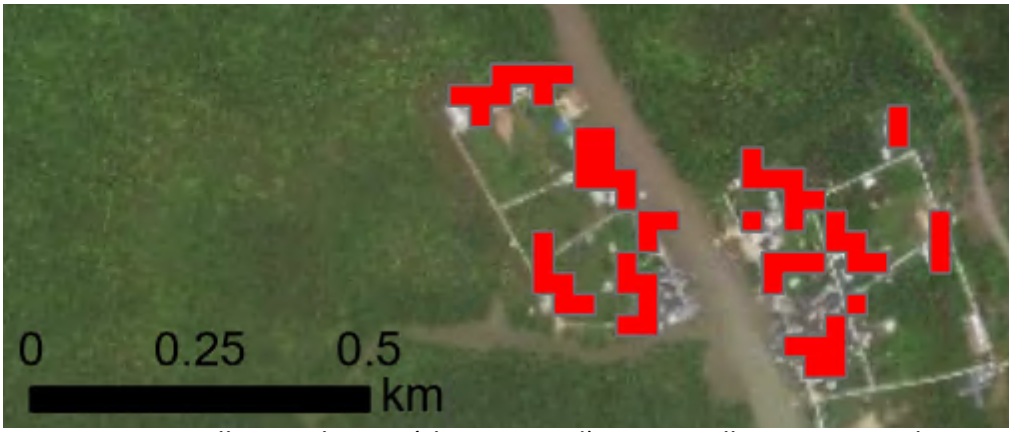

Figure 8: HRSL village polygons (shown in red) over satellite imagery. The disjointed nature of the village extents can be seen.

The villages were clustered based on spatial proximity using defined distance (DBSCAN), with a search radius of 250m. This reduced the count of HRSL-derived settlement polygons from 11,771 to 4,022 villages.

4.2.3 Model development and prediction

In an initial step population estimates from the HRSL were compared to population counts recorded *in-situ* in 16 villages in Sierra Leone. While the sample size is too small to conduct any statistical testing, this initial analysis confirmed the observations from the field that population density in mangrove villages is higher than in neighboring farming villages and that the village population estimates based on an average population density computed over a census unit will underestimate population counts in mangrove villages, thus total population living in the mangrove areas at district, national or regional level. Simple linear model based on village extents and population in the sample available showed a better skill in estimating village population in cross-validation mode (fig. 9, with the caveat of small sample).

In subsequent steps, a linear model and machine learning models are developed using the *in-situ* data and various covariates and used to estimate population in villages where no *in-situ* data are available yet. The models have been constructed using Python code that can be re-used to re-create the outputs when new village population estimates are added. Therefore, as VGI data are obtained, the model predictions can be improved with a minimum of effort, and the web mapping application and data download updated to reflect the new data.

- Caveats

Small observed sample size

Due to delays in the start of the project (see the section: Challenges and Lessons Learned) additional data collection, through VGI or project-specific field work could not take place within the duration of the project. Therefore, the models are based on a very limited sample of data and extensive assessment of confidence levels and error could not be performed. However, the system is in place and as soon as sufficient data is collected these assessments will be performed.

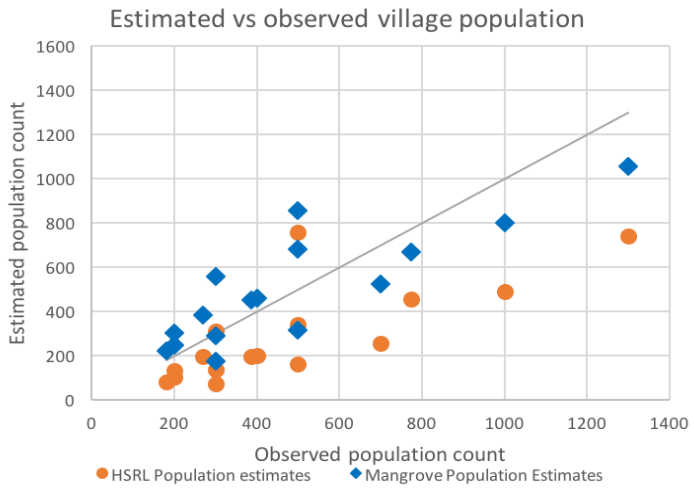

Figure 9: Population estimates vs observed population size using HRSL approach (based on settlement size and census data) and using village data and extent in a linear model.

The interpretation of estimates

Users need to bear in mind that the information in non-visited villages is only an estimate with numerous sources of 'error' such as the village extent, mangrove extent that determines whether a village is selected or not and the distance to the mangrove (one of the covariates used in the Random Forest model), and the imperfections of the models. To highlight the uncertainty around the estimates for non-visited villages, the values of estimates generated with all three models are displayed separately and not aggregated into one central estimate with an uncertainty range (e.g. mean/middle value and range [min, max]). In our experience the latter is often conducive to considering only the central value and we hope that displaying all the estimates highlights the uncertainty surrounding the estimates. As more *in-situ* data is provided, the uncertainty range will be reduced for a larger number of villages as well as for spatially aggregated information.

4.3 Data visualization and access

4.3.1 The mapping interface

The mapping interface displays the village data used in development of the population models (visited villages) along with the estimated population for the villages with 100m of the mangroves derived from the HRSL. The application supports query (through clicking on village features and the use of Structure Query Language of the feature attributes) and panning and zooming to see the villages in context over high resolution imagery. The mangrove data are displayed (with transparency) over the imagery but behind the village data. Administrative units at the second level (district) are overlain on the image with only outlines drawn; these can be queried by clicking to show the percentage of population from the HRSL and each population model compared to the total population for the district in 2015 from census data sources. Detailed information on the census data sources is available by country on the Gridded Population of the World web site, from which the population totals are derived.[3]

The services are all available via Esri's ArcGIS Online, and can be added directly to GIS software such as QGIS or ArcGIS Desktop. The data are been published in a web application template that allows a legend to be shown, layers to be turned on and off, and for simple queries to be made. The application, like the services, are available without the need for logging in to Esri's ArcGIS Online (public).

4.3.2 Data download

The data download page is also deployed on ArcGIS Online, as a simple mechanism to provide access to the files in the services in three formats: Shapefile, File Geodatabase, and as Comma-separated values files (CSV). The page is also publicly available and can be accessed without having to log in to download the zip archive containing all of the files. The download page is linked directly from the project web site.

[3] See: https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11/docs

4.4 Making the results public and sustainable

The final step to achieve project's objectives was to make the various interfaces and the code used to generate them. While these were the last bricks of the project they were designed with future contributions and updates in mind. The code can be re-used easily with updated inputs (new village location, size, and population estimates) to create updated models and outputs. Data updates added to the ArcGIS Online services (and download page) are simple to perform—the new data with the same structure is simply published again to the existing services, and will be immediately be available in the map viewer and data services.

4.4.1 MAHRV portal

https://mahrv.ciesin.columbia.edu/home

Data contribution and Data access capability were grouped under the same portal, a one-stop-shop for population data on hard-to-reach villages.

The web site portal for the MAHRV site is built on open-source Drupal 7 and presents general information and a detailed guide to the Portal. The portal was intentionally built as a very simple interface to collect and disseminate data, to allow easy access in low band-width and unreliable internet. We use webform capabilities to facilitate interactions with users; for those interested in contributing information, requests are evaluated and passed on for manual registration with the ArcGIS project that makes our data-gathering tools available. Contributed data will be archived in the Columbia University Commons, as well as persisted on the ArcGIS Online site as part of the Columbia University subscription. Web-based mapping and download capabilities are also hosted within ArcGIS project and are dynamically linked to the portal. Google Analytics capability tracks website users.

4.4.2 GitHub repository

https://github.com/gyetman/mahrv

Code for population estimates and all the necessary data to run the code (data inputs) are available GitHub. The GitHub repository includes documentation that briefly describes the data. There are two Jupyter Notebooks that can be downloaded and run using Python (or viewed on the GitHub site in a browser without the need for Python) linked in the GitHub readme file. These include comments for each section of code and plots of the data inputs and outputs.

4.4.3 Sustainability

The services and data are hosted on the Esri cloud as part of the Columbia University site license with Esri, an ongoing subscription (over 20 years to date) that is extremely likely to continue, as both CIESIN and the University use Esri software extensively for research and teaching. Continued maintenance of the ArcGIS Online subscription and contents is also funded by internal University funds, and maintenance of the cloud infrastructure, including security patches, software updates, and migration of existing services, is all provided by Esri as part of the subscription service. The primary investigator can also submit the data, code and report to the University Library Academic Commons for long-term archiving; the Libraries have committed to maintain archives of submitted and approved data for a minimum of 50 years.

As new organizations and volunteers submit data to the site, the update of the model and publishing of updated services and data can be completed with a minimum of effort. The scripts to update the model, including data preparation scripts and processes to publish the data are available internally. The GitHub repository (and local copies) include all of the logic required to update the models and produce new population output estimates.

5. Challenges and lessons learned

The main objective of the project was to help fill the gaps in spatial coverage of population distribution and to increase the visibility of off-grid, hard-to-reach populations, living in mangrove areas of coastal West Africa. To achieve this objective the project built a system that allows both, the collection of population data through a volunteered geographic information (VGI) system and a display/download of population data (recorded or estimated), through a user-friendly on-line interface. Throughout its implementation the project planned to liaise with stakeholders and potential contributors to gather feedback on various technological solutions, test them and collect data, but achieved it in fewer instances than anticipated.

The interface and the underlying technical pillars were successfully built and the hypotheses that a specific population system is needed for some geographies such as mangrove areas, was validated. The interface can be easily adapted to other geographies and can be maintained at CIESIN at no additional cost.

During its implementation, the project faced a major delay in the contract process and project start - approximately 30% of the planned duration - and was confronted with challenges in staffing and timing. The structure of project was relatively simple with only few components and the initial workflow was sequential with designing and building the blocks, testing them and disseminating them and numerous opportunities for feedbacks from the users. This approach, however, had many dependencies between the different building blocks. While both, the VGI and the models were developed in parallel, further refinement of the models relied on contributions from the field using the VGI, also meant to test in real time the overall process and interfaces. Final portal development was also contingent on completion and testing of the pillars. Despite the delay in the start, the project decided to focus on the technical parts underpinning the final interface and speed the process without modifying its structure in depth. The team held regular meetings for all the staff involved in building the different components to follow the progress and address problems, such as communication between different parts of the system and feedbacks. This approach, however, revealed to be semi-successful: it allowed to build the final product but reduced staffing meant that the process was slow and the technical part took longer than anticipated. Thus, the testing of the full system and VGI data contribution could not be fully rolled out and models could not be updated because the field activities the project intended to interact with were completed by the time the VGI was ready. Revising in depth the very sequential initial workflow and removing certain dependencies would have been a better option. For example, additional data collection necessary to inform the models ultimately did not need to use VGI and could have happened earlier. Similarly, the public-facing portal could have been built from the beginning, independently of the completion of the pillars. This would have made testing of the VGI and of the mapping interface easier, as our colleagues in the field found navigating the AGOL system

directly cumbersome. Lack of one-stop–shop public interface from the start also delayed the submission of the intermediate technical deliverables until the portal was developed. Through those problems the project better understood the technical challenges in accessing the pillars and made sure that detailed, step-by-step visual explanations of the structure of the portal and the VGI process are easily accessible on the final portal.

Reduced staffing also meant no built-in redundancies, leading to limited leadership when the PI became sick hampering the steady progress of the project and the originally planned data collection, and to limited outreach to end-users during the project. Narrowing the focus on the technical component of the project, however, allowed to have a fully functional and publicly accessible interface – the engine of the data collection and access – and it will be further disseminated through CIESIN's networks after the end of the project. Those networks will also be mobilized to contribute new data through the VGI, following the intention of the system.

CIESIN is open to expanding the geographical scope and to adapting the interface to other geographies as needs arise.

6. Conclusions

In order to increase the visibility of off-grid hard-to-reach populations, the project successfully developed an interface to collect, visualize and download population counts for villages located within the mangroves of West Africa. The approach adopted innovatively combined the analysis of high resolution satellite imagery, Volunteered Geographic Information (VGI) and statistical modeling. The use of the High Resolution Settlement Layer (HRSL) improved the spatial resolution/granularity of village extents (moving from points to 30 m resolution village extents), thus the identification of very small villages. By using the HRSL as a base for population models, the project provided population estimates at a granularity not before completed at the scale achieved here (most of West Africa mangroves). The information is publicly accessible for visualization and download through the Mapping Hard to reach Villages (MAHRV) portal.

The project also created a framework of software and services that allows updating the population model and public data product with a minimum of effort. These services can be maintained for many years without additional funding and the updates are possible under existing resources.

The project contributed to the production of population estimates in a specific geography -- the mangroves in West Africa -- and demonstrated that the standard covariates used in most population models may not be appropriate for remote hard-to-reach places.

References

Giri, C., E. Ochieng, L.L. Tieszen, Z. Zhu, A. Singh, T. Loveland, J. Masek, and N. Duke. 2005: Global Mangrove Forests Distribution, 2000. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4J67DW8. Accessed August, 2019.

Nieves, J.J., F.R. Stevens, A. E. Gaughan, C. Linard, A. Sorichetta, G. Hornby, N.N. Patel and A. J. Tatem, 2019: Examining the correlates and drivers of human population distributions across low- and middle-income countries. Journal of The Royal Society Interface,Volume 14 Issue 137. https://doi.org/10.1098/rsif.2017.0401

Lloyd, C.T., H. Chamberlain, D. Kerr, G. Yetman, L. Pistolesi, F. R. Stevens, A. E. Gaughan, J. J. Nieves, G. Hornby, K. MacManus, P. Sinha, M. Bondarenko, A. Sorichetta & A. J. Tatem, 2019: Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets, Big Earth Data, 3:2, 108-139, DOI: 10.1080/20964471.2019.1625151

Shaghaghi et al., 2011: Approaches to recruiting 'hard-to-reach' populations into research: a review of the literature. Health promotion perspectives, 1(2), 86.

Stevens F.R, A.E. Gaughan, C. Linard, A.J. Tatem, 2015: Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10 (2): e0107042. doi:10.1371/journal.pone.0107042

Tiecke T.G , Xi. Liu, A. Zhang, A. Gros, N. Li, G. Yetman, T. Kilic, S. Murray, B. Blankespoor, E. B. Prydz, H.-A. H. Dang, 2017: Mapping the world population one building at a time, arXiv:1712.05839v1. https://arxiv.org/abs/1712.05839

Annex 1 – The verbal consent script and the questionnaire included in the VGI interface

A1.1 The verbal consent/Information sheet

[Interviewer, please read the following to the respondent.]

Good morning/ afternoon. My name is [interviewer's name] and I am here to ask you some questions about your village. As a knowledgeable person in the community, I am reaching out to see if you are willing to answer some general questions about this village.

Your participation is voluntary

Answering questions is voluntary, you can skip any question at any time.

Purpose of this research

We are interested in learning more about how many people live in this village, whether this human settlement is seasonal or permanent, and the various services available (for example, is there a health clinic in this village? or is there a primary school?). With this research we will understand better the configuration and characteristics of human settlements within mangrove forests.

How long does it take?

The survey takes around 15 minutes to complete.

Risks and possible discomforts

If you choose to participate, we anticipate minimal risks and only the minor discomfort that might accompany answering a questionnaire.

Potential benefits

We cannot promise any benefits to you or others from taking part in this research. However, possible benefits include contributing to the general knowledge about communities living near or within mangrove forests.

What happens to the information collected?

I will store all your answers in this mobile device [show smartphone/ tablet]. After we finish, I will save all the information here, and once I get access to the internet, I will send the data to a place where answers from other respondents, from other villages, are grouped together. All the information that is collected is stored and used securely.

Privacy protection

I am not collecting any identifiable information from you. I will neither take your name or contact information. My interest of interviewing you is only to collect information about the village only.

Contact information

The research for this study is performed by the Center for International Earth Science Information Network (CIESIN) Research Team, at Columbia University. Their mailing address is: 61 Route 9W Palisades NY 10964, USA. Their telephone number is: +1-845-365-8988. Contact them if you think the research has harmed you; if you have questions, concerns, or complaints about this research; or if you wish to withdraw your answers from this study.

This research has been reviewed by the Columbia University IRB. Their mailing address is: Columbia University Medical Center, 154 Haven Avenue, 1st Floor, New York, NY 10032. Their telephone number is: +1-212-305-5883. Contact them if you are unable to contact the research team; if the research team does not respond to your contact; if you want to talk to someone besides the research team; or if you have questions about your rights as a research participant.

Do you agree to participate in this study?

Yes → [continue with the questionnaire]

No → Are there any questions I could answer?

→ [if still no] Thank you for your time!

A1.2 The questionnaire

1. Name of individual entering data (automated, based on controlled access).
2. Date -start (automated)
3. Date -submitted (automated)
4. IMEI (automated, if using mobile device)
5. UUID (automated, for each submission).
6. Location (latitude, longitude, precision if collected in the field using mobile device)
7. Administrative hierarchy (nested, pre-populated domains, as applicable)
   a. Country
   b. Admin level 1 (e.g. state)
   c. Admin level 2 (e.g. county)
   d. Admin level 3 (e.g. township)
   e. Admin level 4 (e.g. village)
   f. (varies per country)
8. Traditional hierarchy (nested, pre-populated domains, as applicable)
   . Country
   a. Level 1 (e.g. zone)
   b. Level 2 (e.g. area)
   c. (varies per country)
9. Name of village or human settlement
10. Alternative name of village or human settlement
11. Estimated total population within the village
12. Year of estimated total population within the village

13. Estimated number of houses (dwellings) within the village
14. Year of estimated number of houses (dwellings) within the village
15. Is this a seasonal or permanent settlement?
16. Is there a {…} in the village? Please select all that apply

a. Health facility
b. Health post
c. Nursery/ Pre-K school
d. Primary school
e. Secondary school
f. Vocational/ technical school
g. Market
h. Church
i. Mosque
j. Other place of worship

17. If not how far is the {…} using the best transportation (estimate in distance or time)

a. Health facility
b. Health post
c. Nursery/ Pre-K school
d. Primary school
e. Secondary school
f. Vocational/ technical school
g. Market
h. Church
i. Mosque
j. Other place of worship

18. If not what is the best transportation to get to

a. Health facility
b. Health post
c. Nursery/ Pre-K school
d. Primary school
e. Secondary school
f. Vocational/ technical school
g. Market
h. Church
i. Mosque
j. Other place of worship

19. Source of information, please select:

a. Personal knowledge
b. Village authorities

c. Village resident
d. Local official
e. Regional official
f. Census worker
g. NGO worker
h. Other

20. Notes

Annex 2 – Random Forest Model Covariates

This annex lists the covariates used in the Random Forest Model and presents the correlation colormap. The covariates can be found at:
https://www.worldpop.org/project/categories?id=14

The code for the model is available on GitHub (https://github.com/gyetman/mahrv), including the input and output data (in comma-separated values format, .csv), and in the form of a Jupyter notebook:
https://github.com/gyetman/mahrv/blob/master/Random%20Forest%20SLE.ipynb
The notebook includes the Python code, comments, and plots, and can be viewed directly with only a web browser. The archive can also be downloaded and run locally using an installation of Python and several open source and free Python modules, which are also listed in the Jupyter notebook. A notebook showing the linear model is also available:
https://github.com/gyetman/mahrv/blob/master/Linear%20Model%20v2.ipynb

Table 1. Covariates used in training the random forest model together with variable names as used in the code available on GitHub.

Variable Name	**Description**	**Correlation with Population**
area_sqm	Village area in square meters	0.450307
bsgm_wpgp_2015_dst	Distance to built up (2015)	-0.443961
guf_ghsl_dst_2014	Distance to combined Global Human Settlement Layer and Global Urban Footprint	-0.443961
ccilc_dst011_2015	Distance to ESA-CCI-LC cultivated area edges (2015)	-0.126465
ccilc_dst040_2015	Distance to ESA-CCI-LC woody-tree area edges 2015	-0.097729
ccilc_dst130_2015	Distance to ESA-CCI-LC shrub area edges 2015	-0.060866
ccilc_dst140_2015	Distance to ESA-CCI-LC herbaceous area edges 2015	-0.05005
ccilc_dst150_2015	Distance to ESA-CCI-LC sparse vegetation area edges 2015	-0.035995
ccilc_dst160_2015	Distance to ESA-CCI-LC aquatic vegetation area edges 2015	0.092098
ccilc_dst190_2015	Distance to ESA-CCI-LC artificial surface edges 2015	0.021814
ccilc_dst200_2015	Distance to ESA-CCI-LC bare area edges 2015	-0.067607

ghsl_dst_2014	Distance to Global Human Settlement Layer	-0.423925
guf_dst_2012	Distance to Global Urban Footprint	-0.549707
cciwat_dst	Distance to edge of water bodies	0.244927
cciwat_dst_negative	Distance from edge of water body (internal, negative)	0.244927
gpw4coast_dst	Distance to coastline	0.249973
osmint_dst	Distance to road intersections	-0.173572
osmriv_dst	Distance to rivers	0.187002
osmroa_dst	distance to roads	-0.196094
slope	Slope (degrees)	0.224786
topo	Elevation	0.472291
tt50k_2000	Accessibility to cities	-0.452345
wclim_prec	Precipitation (30-year norm)	-0.284977
wclim_temp	Temperature (30-year norm)	-0.217391
urbanaccessibility_2015	Urban accessibility	-0.29769
viirs_2016	Lights at night, annual luminosity	0.178633
wdpa_cat1_dst_2017	Distance to protected area (category 1), World Database on Protected Areas	0.05247

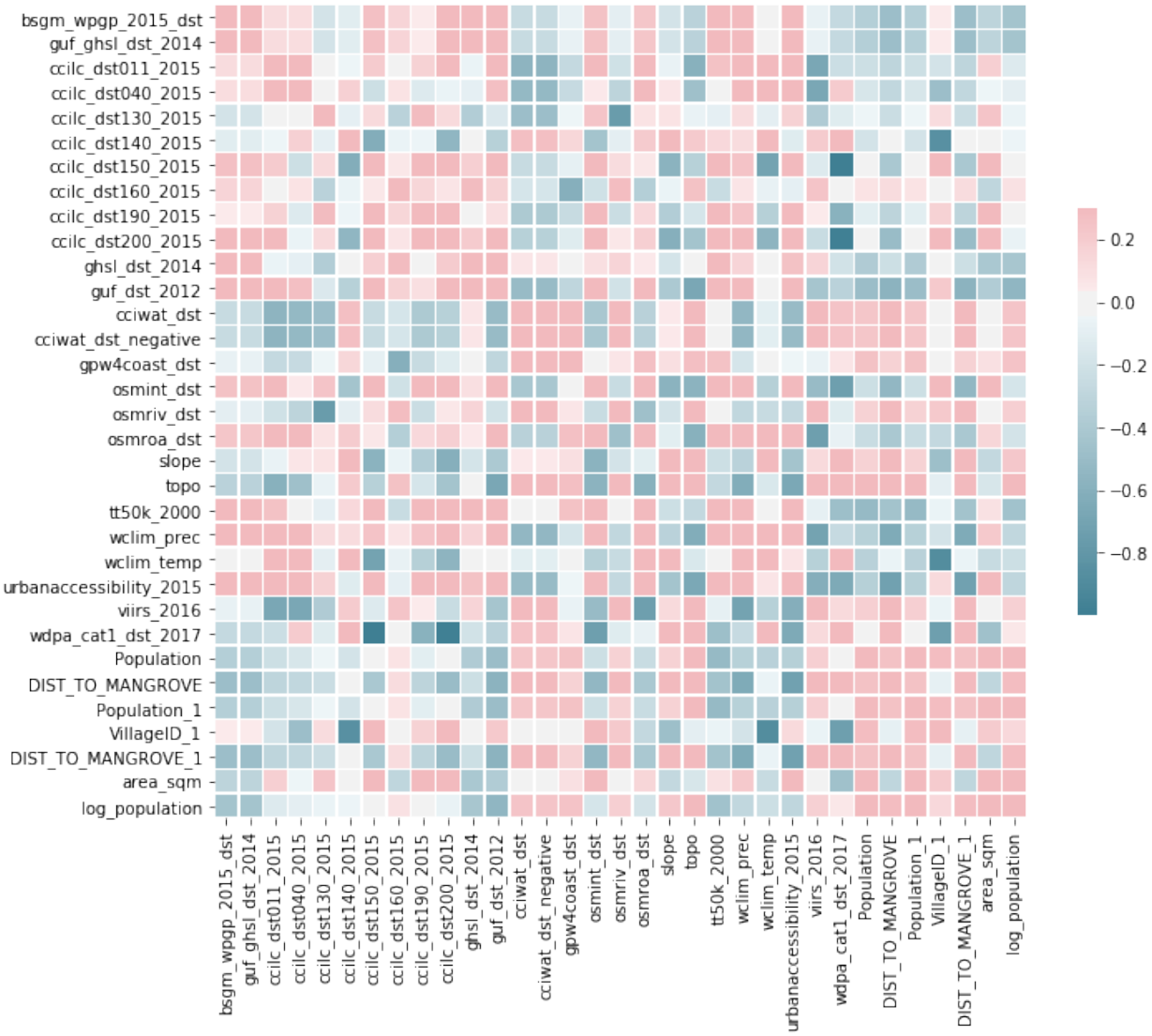

Figure 10: Colormap of correlations between covariates used in the Random Forest model; the colormap includes the log of the population.

Annex 3 – Risk mitigation strategies and Output Indicators

- Risk mitigation

Did you need to address any risks? If yes, what were the risks and how were they managed or mitigated?

Yes

The project faced a major challenge at its inception with a six months delay in its start. While a risk of a small delay in project start was planned for, such large delay was not anticipated. The major consequence of the delay was the reduction in personnel participating in the project, with an additional consequence being end of field work that was supposed to support field-testing of the VGI interface and the first collection of additional data. The challenge was mitigated by focusing the work on the part of the project that was critical to the overall concept underlying the project – the interface allowing future data contribution and continuous improvement of the visibility of hard-to-reach villages in the mangroves of West Africa, leaving additional data collection and outreach for a later date. Despite drastically reduced staff the interface was successfully built and is ready for data collection and/or village information access. CIESIN will ensure the dissemination of the interface through its network of collaborators after the end of the project so that data can be contributed and the information about hard-to-reach villages within the mangroves in West Africa accessed by relevant stakeholders.

- Output Indicators

Have there been any final results or outcomes in which data or methods have allowed data to be produced: faster; more cheaply; at a higher resolution or granularity, or where there was no data before? If yes, please describe.

Yes

The use of the HRSL product has allowed an increase in spatial resolution / granularity of village extents (moving from points on the map to 30m resolution village extents); more importantly, it permits the identification of very small villages. Even though the HRSL is aggregated to 30m, it is based on 50cm optical satellite imagery, which is detailed enough to locate small villages that are clusters of only a few buildings. Through using the HRSL as a base for population models, the project provides population estimates at a granularity not before completed at the scale achieved here (most of West Africa mangroves). The project created a framework of software and services that allows the update of the population model and public data product with a minimum of effort; the services can be maintained for a number of years without additional funding, and the updates are possible under existing center resources.

Has the project contributed to the production and/or use of data disaggregated by a) sex b) disability c) age, d) geography (or other)? If yes, please summarize the of types of disaggregations and the context.

Yes

The project contributed to the production of population estimates in a specific geography: mangroves in West Africa. Most population models take a country-wide or large region (continent, global) approach and are heavily influenced by large urban areas and more dispersed rural or ex-urban population distribution patterns. Using similar approaches to other models, this project revealed relationships between population and spatial covariates that are different than broader-scale patterns seen in models for larger regions.

Has the project contributed to the use and/or production of gender statistics? If yes, please describe.

No