GRID3

GEO-REFERENCED INFRASTRUCTURE AND DEMOGRAPHIC DATA FOR DEVELOPMENT

Mapping and Classifying Settlement Locations

June 2021

Table of Contents

Glossary

Area-level classification
An approach to classifying settlements that uses high-resolution building footprint data sets to identify the patterns that can be observed in buildings.

Building footprints
Digitised outlines of buildings derived from imagery.

Building-type settlement classification
The most basic level of classifying settlements, it classifies each building in a settlement as either residential or non-residential.

Comprehensive settlement layer
Combines settlement extents with the attributes (e.g. place names or administrative units) from the point layer. The comprehensive settlement layer can be in the form of either a point layer or polygon layer.

Database schema
The table structure of the rows and columns used within a data file.

Ensemble
A combination of predictions across models that is used to improve overall prediction and to avoid “overfitting” from any individual model in the ensemble.

Extent
The geographic areal coverage of a group of settlements.

Intra-settlement categorisation
The application of methods to identify and classify local areas of settlement into different types based on the size, shape, and arrangement of structures, as well as their relationship to infrastructure and other features of the landscape.

Machine learning
A subset of artificial intelligence in which a computer uses one or an ensemble of statistical mathematical models in order to analyse the hidden patterns and inferences within a data set.

Machine learning
A subset of artificial intelligence in which a computer uses one or an ensemble of statistical mathematical models in order to analyse the hidden patterns and inferences within a data set. The model enables the analysis to be applied (i.e. predict) efficiently beyond the training data set.

Points of interest (POIs)
Point features, or geo-referenced points, that depict infrastructure, buildings, and landmarks. These points locate structures and services that are critical to the health and well-being of society, such as health facilities, schools, marketplaces, banks, warehouses, and wells.

Polygon
A digital representation of a settlement’s boundaries. Polygons also describe the extent of a settlement and can include other attribute information, such as the total number of buildings within the given area.

Probability model
A mathematical representation of a random phenomenon. [1]

Raster
In its simplest form, a matrix of cells (or pixels) organised into rows and columns (or a grid) where each cell contains a value representing information, such as temperature or number of buildings. Rasters include digital elevation models, digital aerial photographs, imagery from satellites, digital pictures, or even scanned maps. [2]

Remote sensing
The process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance from the targeted area. Special sensors on satellites or aircraft collect remotely-sensed images of the Earth, which help researchers "sense" information about the Earth. [3]

Settlement
A settled area of permanently inhabited structures and compounds that can range from small rural villages to large urban zones.

Settlement data
Information on the location and characteristics of potentially inhabited structures that is used as input data to produce high-resolution population estimates.

Settlement layer
A dataset that provides settlement points or polygons and their names to spatially locate, identify, and visualise settlement features.

Settlement mapping
The collection, creation, and harmonisation of multiple data sources representing places that are inhabited by people.

Supervised models
One subset of machine learning. Supervised models are trained with small amounts of pre-classified data, as opposed to unsupervised classification methods that do not use a training data set. A model helps to predict the behaviour of something in the real world (termed a "system"). The model explains how the system operates and demonstrates the effect of different components of the system.

Settlement point
A representation of a location of a settlement and its basic attribute data.

Zero-Inflated Poisson (ZIP) regression
Used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the zip model has two parts, a Poisson count model and the logit model for predicting excess zeros. [4]

1. National Institute of Standards and Technology. "Probability model". https://csrc.nist.gov/glossary/term/Probability-model.
2. Definition based on Environmental Systems Research Institute. "What is raster data?". https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm.
3. Definition based on United States Geological Service. "What is remote sensing and what is it used for?". https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used?qt-news_science_products=7#qt-news_science_products
4. Definition based on UCLA Institute for Digital Research & Education Statistical Consulting. "Zero-Inflated Poisson Regression | R Data Analysis Examples". https://stats.idre.ucla.edu/r/dae/zip/.

Executive Summary

In order to identify a country's development priorities, it is crucial to understand as much as we can about the places where people work and live. The data we collect about such settlements can be used to inform a wide range of activities, whether that is planning a government's budget, distributing bed nets in a community, or carrying out a rapid disaster response. The more accurate and detailed settlement data are, the more effective such interventions can be.

Creating the settlement layer

Traditionally, census cartography and operational maps are the main source of settlement data. In some cases, these data are paper maps and non-geo-referenced lists that are incomplete or obsolete. By contrast, the work of GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) centres on creating an improved, comprehensive settlement layer that can enable a real-world picture of communities, which in turn facilitates more accurate, more effective analyses.

The first step in creating a comprehensive settlement layer is to obtain point and polygon data from government authorities, as well as permission to publish these data. Once this is done, GRID3 performs an initial, exploratory analysis to become familiar with the data sets' content. Next, the data are evaluated to determine if they cover the entire area of interest. While government data are often the most authoritative type of information, the data are frequently incomplete. Thus, the next step is to fill in what data GRID3 has by using data from a variety of supplemental sources, including NGOs, private companies, grassroots-led efforts, UN entities, and independent contractors. Once point data have been gathered and evaluated, they are ready to be cleaned. This involves combing through numerous files, many of them in different formats and with duplicates both within and among the files. Having understood what is contained in each file, GRID3 can begin integrating the existing settlement point data. GRID3 examines how data within the files are represented and then standardises which words are used to represent different types of data. Among other things, this allows for the identification and removal of duplicates. Lastly, the table structure of the rows and columns used within a data file (or "database schema") is standardised so that the files containing the different data are placed in a uniform format.

Past settlement extents were provided by Oak Ridge National Laboratory, while currently Ecopia Landbase Africa powered by Maxar's building footprints are used for the production of a boundary that helps define the extents. The next step is to validate the data. First, GRID3 compares the polygons against the points and identifies what is missing or incorrect in the points, checking for misalignments, potential errors, and data gaps. Next, the comparison is reversed—points are compared against the polygons to see what is missing from the polygons. Finally, GRID3 creates a comprehensive settlement layer. This layer combines the settlement extents with the attributes from the point layer. The comprehensive settlement layer can be in the form of either a point layer or polygon layer. This layer is sent out to a series of reviewers; after making whatever adjustments are needed, GRID3 finalises the layer.

Classification approaches

While such data sets can provide details on the presence and location of settlements, they often lack information that can differentiate a place in an urban core from a growing fringe of a city, or a growing fringe of a city from a remote village. However, the size and shape of structures, as well as the arrangement and position of structures (relative to other infrastructure or features), can convey information about different land uses or economic activities. GRID3 is exploring new methods for identifying those patterns and extracting additional information that can improve our understanding of settlements at the scales of individual structures and of neighbourhoods. Specifically, GRID3 is developing two approaches that use building footprints, geospatial data layers, and machine learning algorithms to classify structures and local areas.

The most basic level of classifying settlements is by building type—residential versus non-residential. The characteristics of each structure can give clues about a building's use. To accurately predict the building type, GRID3 uses multiple machine learning methods. Each method provides a prediction and these predictions are combined across models into an ensemble that helps to improve the overall prediction. The statistical methods used are supervised models; i.e. the models are trained with small amounts of pre-classified data. GRID3 uses data from OpenStreetMap and other labelled building datasets.

The binary distinction of residential versus non-residential is important for supporting more accurate population models and for guiding services that need to find residential areas. However, these two classes say little about the type of neighbourhood a building is in, and by extension how that neighbourhood fits into the settlement patterns of the wider geographic area. For that information, GRID3 develops a second, area-level classification model. This approach once again starts with the high-resolution building footprint data sets; but in this case, GRID3 focuses on the patterns that can be observed in the buildings (rather than on the individual structures). When viewed together, the size, density, shapes, and orientations of the structures create a visual "texture" that varies across the landscape, indicating differences in land use or type of neighbourhood. To create these neighbourhood classifications, the pattern and texture is quantified at multiple spatial scales across an entire region of interest by calculating the variation in sizes and shapes, as well as many other metrics on a regular grid of locations. Gaussian mixture models and other machine learning methods are used to find clusters of similar patterns and to classify types across a 100 x 100 metre-resolution gridded surface.

GRID3 continues to develop this line of research—the programme's efforts will eventually include expanding the range of attributes generated by (and so input data used by) the classification model, which will improve predictive power. While research on classification is still in a relatively early stage, it is already providing a dramatically different—and improved—perspective on building footprints, and is helping to situate those footprints within wider settlements patterns.

Applications

GRID3's settlements products have a wide range of uses. They can generate open-source data, provide support for data applications to ensure effective impact, or enable training to strengthen national geospatial foundations for future evidence-based development and humanitarian decision-making. These products also fit into the programme's other work on census enumeration, population models, boundaries' delineation, and the identification of infrastructure locations. GRID3's settlements data have already begun to make impacts in several different areas of sustainable development; examples of these impacts are described below.

- While most of the settlements in the eastern Democratic Republic of the Congo (DRC) are visible on satellite imagery, many do not have names attached to them. GRID3 has worked with various partners to change this, and has so far successfully identified the names of all settlements in a selected set of priority health zones in Haut-Lomami and Tanganyika provinces.

- GRID3 has also improved settlement mapping in eastern DRC by organising participatory cartography sessions. By creating a more complete settlement map in the two provinces, GRID3 enabled health workers to improve the completeness of the settlements included in their microplans, allowing for additional target populations to be included in future vaccination campaigns.

- In Nigeria, GRID3 supported vaccination teams to monitor settlement coverage, reduce the number of missed settlements, and improve the teams' outcomes. This work enabled GIS microplans maps to be generated for vaccination teams and for settlements to be visualised.

- GRID3 has worked closely with the Zambian government to create a comprehensive settlement layer that will provide the most comprehensive repository to date for the country's settlement names and locations. A wide range of benefits will result from this still-developing set of settlement data, including: assistance with planning for a nationwide census, more effective immunisation campaigns, and increased capacity for disaster response.

Conclusion

GRID3's activities are adding value to efforts to understand settlements, both by improving data collection/analysis techniques and by improving the data itself. Many challenges still need to be addressed in the campaign to collect comprehensive, actionable data on settlements. These include: capturing temporary settlements, temporary population shifts, long-term population shifts, nomadic settlements, diurnal settlements, and seasonal settlements; distinguishing residential from non-residential populations; reconciling differences between official and local conceptions of places, place names, and settlement boundaries; and negotiating political sensitivities and security issues both within and between countries around places, place names, and settlement boundaries.

In the coming years, GRID3 will continue to work with partners to make its current data more user-friendly and to seek out new ways of collecting and validating information on settlements.

Introduction

In order to identify a country's development priorities, it is crucial to understand as much as we can about the places where people work and live. The data we collect about such settlements can be used to inform a wide range of activities, whether that is planning a government's budget, distributing bed nets in a community, or carrying out a rapid disaster response. The more accurate and detailed settlement data are, the more effective such interventions can be.

"Settlements are dynamic places that often exist across multiple contexts: settlements can be defined by administrative/political boundaries, functional spaces, their relations to other places, or how people use and experience a space."

And yet, ensuring that settlement data are accurate and detailed is no simple matter. Settlements are dynamic places that often exist across multiple contexts: settlements can be defined by administrative/political boundaries, functional spaces, their relations to other places, or how people use and experience a space. For these and many other reasons, the methods used to measure settlements need to be as dynamic as the settlements themselves.

GRID3 settlements work

GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) is a programme funded by a grant from the Bill & Melinda Gates Foundation and the United Kingdom's Department for International Development. It is implemented by Columbia University's Center for International Earth Science Information Network, the United Nations Population Fund, WorldPop at the University of Southampton, and the Flowminder Foundation. GRID3's primary mission is to build spatial data solutions that make development goals achievable. To accomplish this, GRID3 draws on the expertise of its partners in government, the United Nations, academia, and the private sector to design adaptable and relevant geospatial solutions that are based on developing countries' capacity and development needs.

GRID3 defines a settlement as a settled area of permanently inhabited structures and compounds that can range from small rural villages to large urban zones. Settlement data, in turn, provide information on the location and characteristics of potentially inhabited structures and are used as input data to produce high-resolution population estimates.

Data can be derived from high-resolution aerial/satellite imagery or other building maps, and can be represented in several ways: settlement points/polygons, probabilistic modelling, or intra-settlement neighbourhood categorisation:

- **Points/polygons:** Points indicate the location of a settlement and basic attribute data (e.g. the settlement's name). Polygons represent a settlement's boundaries; they also describe the extent (i.e. the geographic areal coverage) of a settlement and can include other attribute information, such as the total number of buildings within the given area.
- **Probability model:** A mathematical representation of a random phenomenon. To predict occurrence of the settlements, a zero-inflated Poisson model (ZIP) is used. [5]
- **Intra-settlement categorisation:** While the previous two settlement data representations focus on locating settlements in space and identifying key attributes, additional information can be attached to these locations. Intra-settlement categorisation is the application of methods to identify and classify local areas of settlement into different types based on the size, shape, and arrangement of structures, as well as their relationship to infrastructure and other features of the landscape.

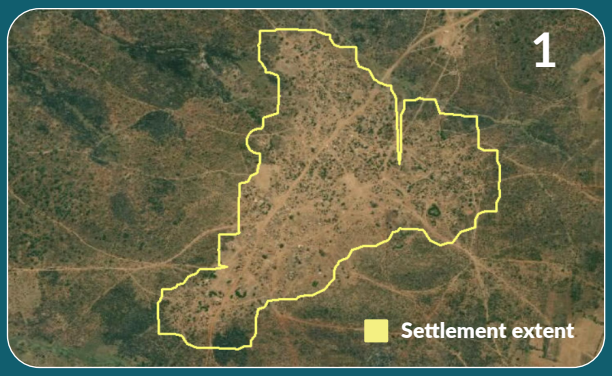
Settlement Data Representations

1. Points/polygons

2. Probability model

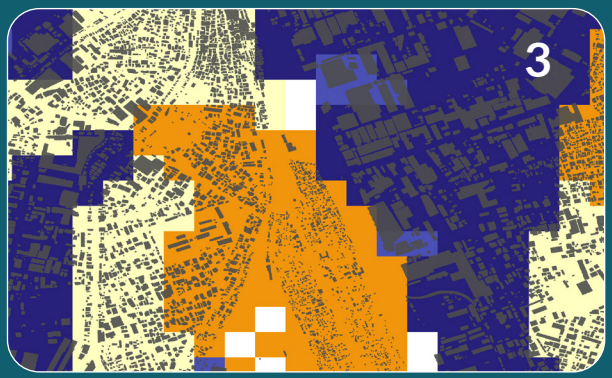
3. Intra-settlement neighbourhood categorisation

Area types shown in different colours and overlaid with building footprints. Using only information extracted from building locations and their patterns, this can help distinguish different areas of settlement within a city.



Settlement extent: CIESIN, Columbia University under the GRID3 programme using Ecopia Vector Maps Powered by © 2020 Maxar; base image: Esri

Settlement extent: Oak Ridge National Laboratory; Probability model: Markus Walsh; base image: Esri



Building footprints © 2020 Maxar Technologies, Ecopia.AI

5. Used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the zip model has two parts, a Poisson count model and the logit model for predicting excess zeros.

Traditionally, census cartography and operational maps are the main source of settlement data. In some cases, these data are paper maps and non-geo-referenced lists that are incomplete or obsolete. Recently, as part of census modernisation, digital cartography is encouraged; but this approach is not yet used widely. By contrast, GRID3's work centres on creating an improved, comprehensive settlement layer [6] that can enable a real-world picture of communities, which in turn facilitates more accurate, more effective analyses. Rather than generate new settlement data via expensive and time-intensive large-scale fieldwork, GRID3 harmonises data that have already been collected from governmental and non-governmental organisations and integrates them with newly collected, targeted field data.

Once a cleaned and standardised settlement dataset is in hand, GRID3 conducts the above described harmonisation and integration via settlement mapping. This is the collection, creation, and harmonisation of multiple data sources representing places. GRID3 works collaboratively with national stakeholders to integrate new data into settlement maps, a process that often involves the use of remote sensing [7] and GPS field data to generate accurate, up-to-date, and complete settlement data. Creating, linking, and/or harmonising settlement maps can improve development initiatives in myriad ways. It aids census enumeration by helping census managers learn where people live. It is instrumental to delineating boundaries and identifying infrastructure locations (e.g. helping to identify all points that are known to fall within a given administrative or health catchment area). And, more generally, mapping settlements is essential to sustainable planning, as accurately mapped and accurately named settlements enable monitoring the progress of development activities, interventions, and programmes in critical sectors such as agriculture, health, and education.

Structure of this paper

This paper describes GRID3's approach to mapping and gathering data on settlements. Section one describes the process of creating the comprehensive settlement layer. Section two describes the motivations behind and different approaches to classifying settlements. Section three discusses the various real-world applications GRID3's settlements mapping have for development initiatives. Section four concludes the paper with a survey of other potential benefits of settlements work, the current shortcomings in GRID3's methods, and plans for the programme's future settlements work.

6. A settlement layer is a dataset that provides settlement points or polygons and their names to spatially locate, identify, and visualise settlement features. A comprehensive settlement layer combines the settlement extents with the attributes (e.g. place names or administrative units) from the point layer. The comprehensive settlement layer can be in the form of either a point layer or polygon layer.
7. "Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance from the targeted area. Special cameras on satellites collect remotely sensed images of the Earth, which help researchers 'sense' information about the Earth." See United States Geological Service reference above.

Creating the Settlement Layer

Many developing countries rely on outdated, incomplete, or inaccurate maps to make their policy-making decisions, which results in inefficient, inequitable planning, as well as unequal resource distribution. However, new methods are being developed to create new, improved settlement maps that can help ground population counts and improve development. GRID3 is helping to lead the effort to collect, organise, and use data that can make settlement maps more effective tools for development. A key part of GRID3's work in this area is creating a comprehensive settlement layer consisting of polygon extents and point attribute data.

Assessing the existing data

The first step in creating a comprehensive settlement layer is to obtain point and polygon data from government authorities (e.g. administrative maps or operational maps) as well as permission to publish these data. Once this is done, GRID3 performs an initial, exploratory analysis to become familiar with the data sets' content; establishing, for example, what format the data are in. Next, the data are evaluated to determine if they cover the entire area of interest. GRID3 looks to see if an area has few data points because it is not settled, or because GRID3 is missing data. GRID3 also looks to see if data are more complete in some areas than others.

While government data are often the most authoritative type of information, the data are frequently incomplete—for example, the data will only cover half the country, or it will cover the entire country but with uneven levels of granularity (e.g. one state will have 100 settlement points, while another has 100,000). Thus, the next step is to fill in what data GRID3 has by using data from a variety of supplemental sources, including NGOs, private companies, grassroots-led efforts, UN entities, and independent contractors. [8] GRID3 iterates this process as many times as necessary, filling gaps in the data as much as possible. During this stage, GRID3 might return to the government to obtain more data, continue searching for other data from non-governmental sources, conduct targeted fieldwork to obtain data, or partner with an organisation or company that is already in-country and obtain data from them.

Creating the settlement locations database (points)

Once point data have been gathered and evaluated, they are ready to be cleaned. This involves combing through numerous files, many of them in different formats and with duplicates both within and among the files. Having understood what is contained in each file, GRID3 can begin integrating the existing settlement point data. GRID3 examines how data within the files are represented—for example, the word "village" might be spelled differently in different files—and then standardises which words are used to represent different types of data. Among other things, this allows for the identification and removal of duplicates. Lastly, the table structure of the rows and columns used within a data file (or "database schema") is standardised so that the files containing the different data are placed in a uniform format. This involves merging hundreds of files and standardising on column names and data types so that the tables in each file contain data that can be consolidated in a uniform way, each containing a unique ID, alternate names, the geographic location of the data point, and a unique settlement ID.

8. If there are several providers of raw settlement data (points and polygons) and conflicting information is provided, field work is conducted to resolve the issues, or the data are reviewed by a local stakeholder. If this is not possible, data provided by the country are defined as the authoritative source and will take precedence; but the schema provides space for alternate names.

Creating the settlement areas database (polygons)

Past settlement extents were provided by Oak Ridge National Laboratory [9], while currently Ecopia Landbase Africa powered by Maxar's [10] building footprints (i.e. digitised outlines of buildings) enable the production of a boundary that helps to define the extents. [11] This is done either by aggregating features that fall within 25m to 100m of each other, or by utilising building density to draw contours around groups of building footprints. Settlement extents are separated into four categories: 1. built-up areas, 2. small settlement areas, 3. hamlets, and 4. hamlet areas.

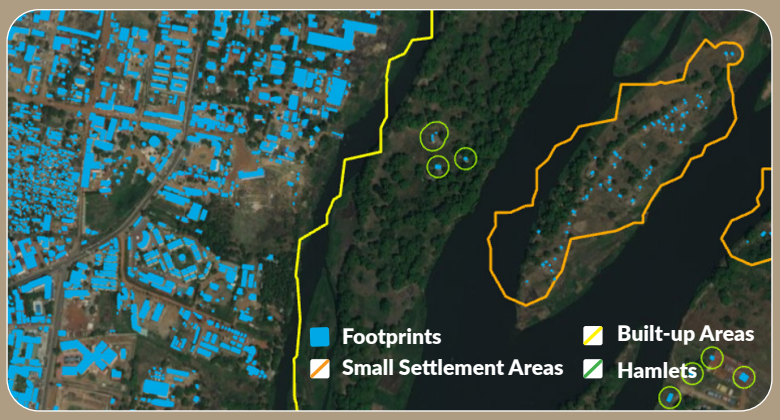
Four Categories of Settlement Extents

1. Built-up Areas (BUAs) polygon feature
An area of urbanisation with moderately-to-densely-spaced buildings and a visible grid of streets and blocks.

2. Small Settlements (SSAs) polygon feature + point feature
A settled area of permanently inhabited structures and compounds of roughly a few hundred to a few thousand inhabitants. The housing pattern in SSAs is an assemblage of family compounds adjoining other similar habitations.

3. Hamlets polygon feature + point feature
A single-family compound or several compounds or sleeping houses in isolation from small settlements or urban areas.

4. Hamlet Areas (HAs) polygon feature
A group of two or more hamlets, with each hamlet being within 200 metres of another hamlet in the HA.



Settlement Extent: CIESIN, Columbia University under the GRID3 programme; Building Footprint: Ecopia Landbase Africa powered by Maxar; Image: Esri

*Definitions based on Barau, I., M. Zubairu, M.N. Mwanza, and V.Y. Seaman. 2014. Improving polio vaccination coverage in Nigeria through the use of geographic information system technology. *The Journal of Infectious Diseases* (November): S102–S110.

Validation and fieldwork

By this stage, GRID3 has collected a set of data points that have been cleaned, normalised, and placed in a standard database schema, as well as a set of polygons that have been created in collaboration with one of GRID3's partners. No data set is perfect, though, and the next step is to validate the data.

First, GRID3 compares the polygons against the points and identifies what is missing or incorrect in the points, checking for misalignments, potential errors, and data gaps. GRID3 takes this information to the partner government to see if it can fill in the gaps, as well as checks non-governmental sources like OpenStreetMap. [12] If none of these sources has the needed information, GRID3 can work with a partner that is already in the field to collect the missing data, or do fieldwork itself.

9. See Oak Ridge National Laboratory. https://www.ornl.gov/.
10. See Maxar Technologies, Inc. and Ecopia Tech Corporation. 2020. Ecopia Landbase Africa powered by Maxar. DigitizeAfrica.ai.
11. Some areas may have a large population but small extent, which would indicate that many people are clustered in a small area—the population density is therefore significant. Some areas may have the same population but larger extent, meaning that the population is spread across a greater area.
12. See OpenStreetMap. https://www.openstreetmap.org/about.

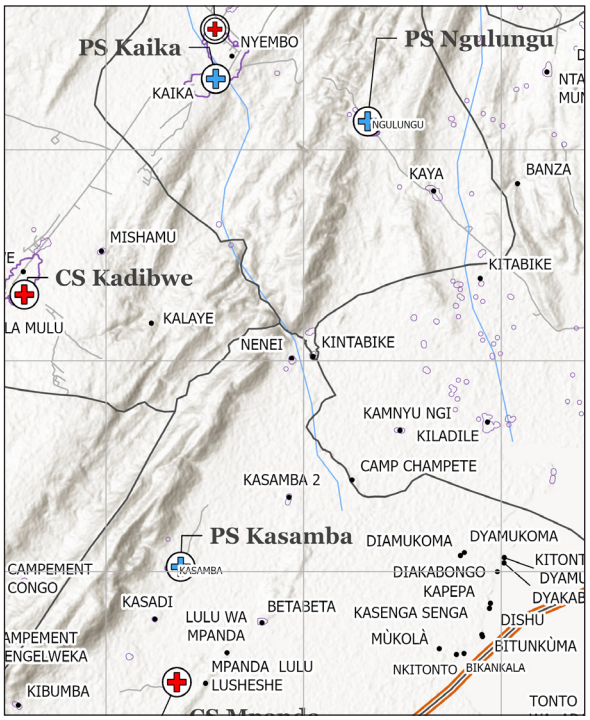
Next, the comparison is reversed—points are compared against the polygons to see what is missing from the polygons. GRID3 visually inspects those points against satellite imagery, which allows for 1) the identification of settlements that were missed by the feature extraction and then 2) the creation of polygons for them.

GRID3 repeats the exercise of linking place names with settlements' extents to get a best-possible settlement layer that incorporates all inputs. Visual and logic checks are automated in order to flag inconsistencies between the two layers (e.g. places that are shown in the raster grid [13] but are not part of the baseline settlement, and vice versa). Conflicts are resolved in coordination with local stakeholders. When applicable, general consultation with local stakeholders is done to validate results.

Comprehensive settlement layer

Finally, GRID3 creates a comprehensive settlement layer. This layer combines the settlement extents with the attributes (e.g. place names or administrative units) from the point layer. The comprehensive settlement layer can be in the form of either a point layer or polygon layer. This layer is sent out to a series of reviewers; after making whatever adjustments are needed, GRID3 finalises the layer.

ZS Bukuma, Haut-Lomami Province, DRC



ZS Mbulula, Tanganyika Province, DRC

Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, GEoland, FEMA, Intermap and the GIS user community

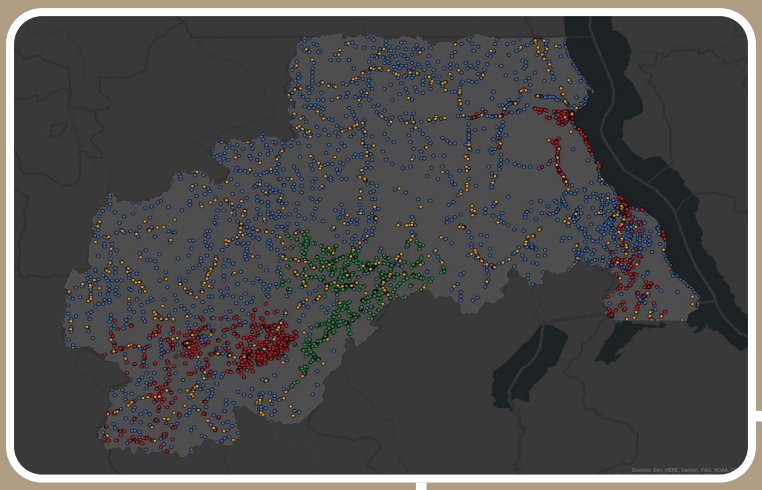
13. "In its simplest form, a raster consists of a matrix of cells (or pixels) organised into rows and columns (or a grid) where each cell contains a value representing information, such as temperature. Rasters are digital aerial photographs, imagery from satellites, digital pictures, or even scanned maps." See Environmental Systems Research Institute reference above.

GRID3 Settlement Mapping Process

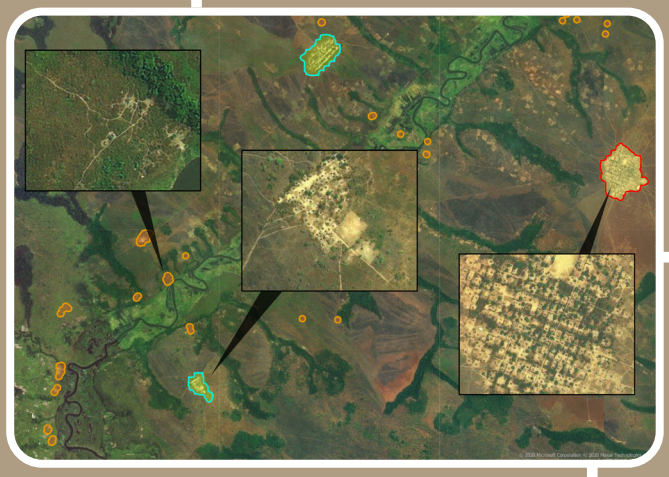
1 Assess the exising data

Obtain point and polygon data from government authorities

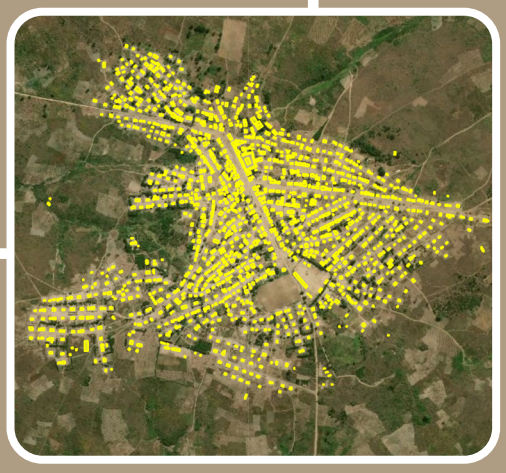
If data are incomplete, fill in data gaps from a variety of supplemental sources, including government authorities, NGOs, private companies, grassroots-led efforts, UN entities, and independent contractors



Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, OpenStreetMap contributors, and the GIS User Community



Sources: © 2020 Microsoft Corporation © 2020 Maxar © CNES (2020) Distribution Airbus DS



Sources: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

2 Create databases (locations and areas)

Standardise which words are used to represent different types of data

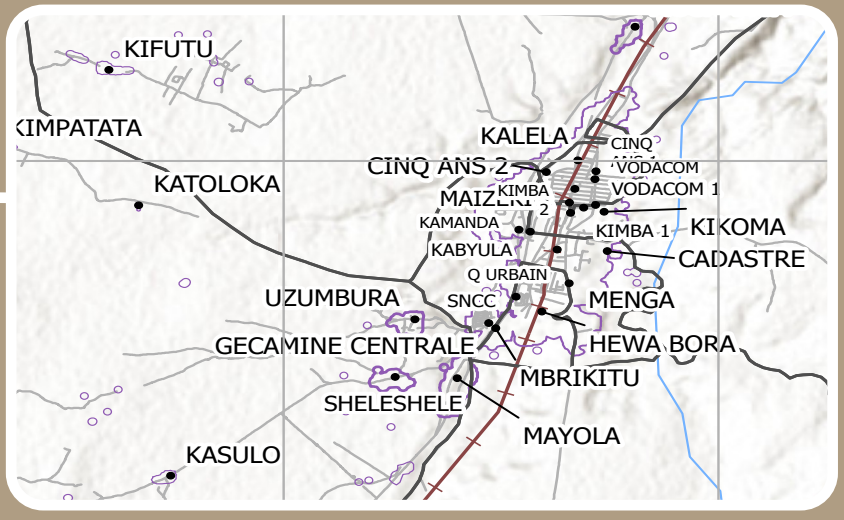
Aggregate features that fall within 25m to 100m of each other, or by utilising building density to draw contours around groups of building footprints

3 Validation and fieldwork

Compare polygons against the points and identify what is missing or incorrect in the points, checking for misalignments, potential errors, and data gaps; and vice versa

4 Settlement Layer

This layer combines settlement extents with attributes from the point layer. The comprehensive settlement layer can be in the form of either a point layer or polygon layer.



Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, GEoland, FEMA, Intermap and the GIS user community

Classification Approaches

The world continues to become more urbanised—55 percent of all people live in urban areas today, a figure that is projected to increase to 68 percent by 2050. [14] But urban areas are not homogeneous—neither between settlements nor within built-up areas—and so their development needs and challenges are not the same. Computational advances are allowing us to gain a more nuanced understanding of urban areas. As mentioned in the previous section, machine learning algorithms are able to automatically detect and map all building features from high-resolution satellite and aerial imagery, and these footprints in turn provide a high-resolution map of potentially settled areas (identifiable by the presence of detectable structures).

"GRID3 is exploring new methods for identifying those patterns and extracting additional information that can improve our understanding of settlements at the scales of individual structures and of neighbourhoods."

While such data sets can provide details on the presence and location of settlements, they often lack information that can differentiate a place in an urban core from a growing fringe of a city, or a growing fringe of a city from a remote village. However, the size and shape of structures, as well as the arrangement and position of structures (relative to other infrastructure or features), can convey information about different land uses or economic activities. GRID3 is exploring new methods for identifying those patterns and extracting additional information that can improve our understanding of settlements at the scales of individual structures and of neighbourhoods. Specifically, GRID3 is developing two approaches (described below) that use building footprints, geospatial data layers, and machine learning algorithms to classify structures and local areas. Such improved pictures of settlements can in turn improve our study of urbanisation patterns, enhance our understanding of the similarities and differences found in local contexts, help distinguish between urban and rural areas, and make population modelling and field surveys more robust (settlement data is a key input to GRID3's bottom-up population models).

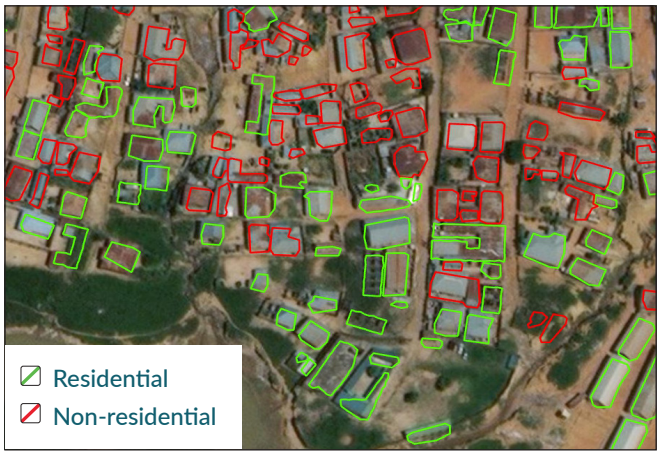
Building-level classification (residential vs. non-residential)

The most basic level of classifying settlements is by building type—residential versus non-residential. The characteristics of each structure (such as their size and shape as well as the presence of infrastructure and other features of interest) can give clues about a building's use.

To accurately predict the building type, GRID3 uses multiple machine learning methods. Each method provides a prediction and these predictions are combined across models into an ensemble [15] that helps to improve the overall prediction. The statistical methods used are supervised models; i.e. the models are trained with small amounts of pre-classified data. GRID3 uses data from OpenStreetMap and other labelled building datasets.

14. See United Nations Department of Economic and Social Affairs. 2018. 2018 Revision of World Urbanizaton Prospects.
15. A combination of predictions across models that is used to improve overall prediction and to avoid "overfitting" any individual model in the ensemble.

The binary distinction of residential versus non-residential is important for supporting more accurate population models and for guiding services that need to find residential areas. However, these two classes say little about the type of neighbourhood a building is in, and by extension how that neighbourhood fits into the settlement patterns of the wider geographic area. For that information, GRID3 develops a second, area-level classification model.



An object-based, binary classification, machine learning approach is utilised to train a model in order to predict building structures as residential or non-residential types.

Finding similar areas from building patterns (neighbourhood classification)

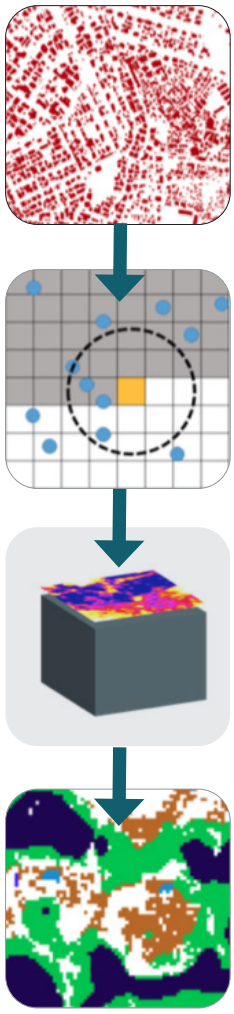
GRID3 is developing methods that can go beyond the residential/non-residential classification. Once again, GRID3 starts with the high-resolution building footprint data sets; but in this case, GRID3 focuses on the patterns that can be observed in the buildings (rather than on the individual structures). When viewed together, the size, density, shapes, and orientations of the structures create a visual "texture" that varies across the landscape, indicating differences in land use or type of neighbourhood.

To create these neighbourhood classifications, the pattern and texture is quantified at multiple spatial scales across an entire region of interest by calculating the variation in sizes and shapes, as well as many other metrics on a regular grid of locations. Gaussian mixture models and other machine learning methods are used to find clusters of similar patterns and to classify types across a 100 x 100 metre-resolution gridded surface. [16]

GRID3 continues to develop this line of research—the programme's efforts will eventually include expanding the range of attributes generated by (and so input data used by) the classification model, which will improve predictive power. While research on classification is still in a relatively early stage, it is already providing a dramatically different—and improved—perspective on building footprints, and is helping to situate those footprints within wider settlements patterns.

16. GRID3 works on a 3 arc-second grid, which in actuality translates closer to 90m x 90m. GRID3 nominally rounds the figures up to 100m x 100m.

Quantifying and Classifying Patterns



Stage 1: Settlement Layer
The settlement layer is comprised of building footprints; the goal is to classify areas of similar patterns.

Stage 2: Moving Window Calculations
At every 100m x 100m location across the settlement map, a series of calculations quantify the size, shape, clustering of buildings, and other measures of the patterns. The calculations are performed within differently sized "windows" around the locations.

Stage 3: Fragmentation Statistics
Collectively, the calculations are referred to as a "stack of fragmentation statistics." At each location on the map, there is a range of values describing the local settlement patterns.

Stage 4: Unsupervised classification
Unsupervised classification methods are used to explore the fragmentation statistics and to find groupings of similar data. These groupings are mapped back to the real-world locations to produce a settlement classification.

Images created by Warren C. Jochem, using building data from Ecopia Landbase Africa powered by Maxar.

Applications

GRID3's settlements products have a wide range of uses. They can generate open-source data, provide support for data applications to ensure effective impact, or enable training to strengthen national geospatial foundations for future evidence-based development and humanitarian decision-making. These products also fit into the programme's other work on census enumeration, population models, boundaries' delineation, and the identification of infrastructure locations.

GRID3's settlements data have already begun to make impacts in several different areas of sustainable development. Some of these impacts are described below.

Service provision and operations

Vaccine campaigns, supply chain routing, and disaster response/risk reduction can all be strengthened using GRID3's settlements data.

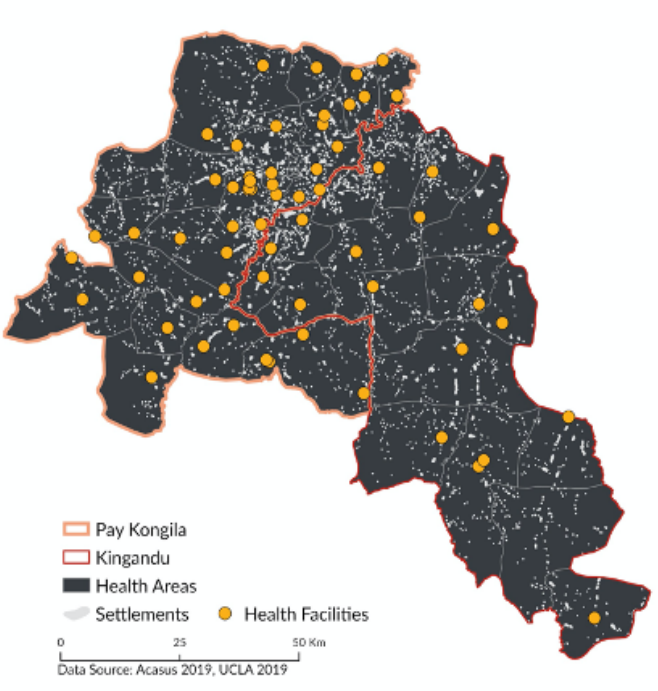
For example, in eastern Democratic Republic of the Congo (DRC), political instability and associated infrastructure shortfalls have made it difficult to create accurate settlement maps for the region. While most of eastern DRC's settlements are visible on satellite imagery, many do not have names attached to them. To help change this, the Bill & Melinda Gates Foundation, GRID3, PATH, [17] and Novel-T [18] partnered with the DRC Ministry of Health's polio vaccination campaign to identify the names of all settlements in a selected set of priority health zones in Haut-Lomami and Tanganyika provinces. Smartphones were provided to vaccinators; while recording the needed data on child vaccinations, the vaccinators also used their phones to capture the names of the settlements they travelled to, as well as the health area and administrative unit that a given settlement fell within. With about 300 vaccinators travelling for 3-4 days each in each of the selected health zones, GRID3 built on the considerable human resources deployed in the field to successfully and substantially augment the number of collected settlement names, as well as improve knowledge about settlement locations. The data derived from this project will be used by not only the Ministry of Health for future polio vaccination campaigns, but are also likely to improve routine immunisation, mass measles and anti-malaria campaigns, and health service delivery in general. The data will serve as a basemap for any NGO or humanitarian organisation working in these health zones, and also for the National Institute of Statistics (INS, Institut National de la Statistique) as it prepares to carry out DRC's first nationwide census since 1984.

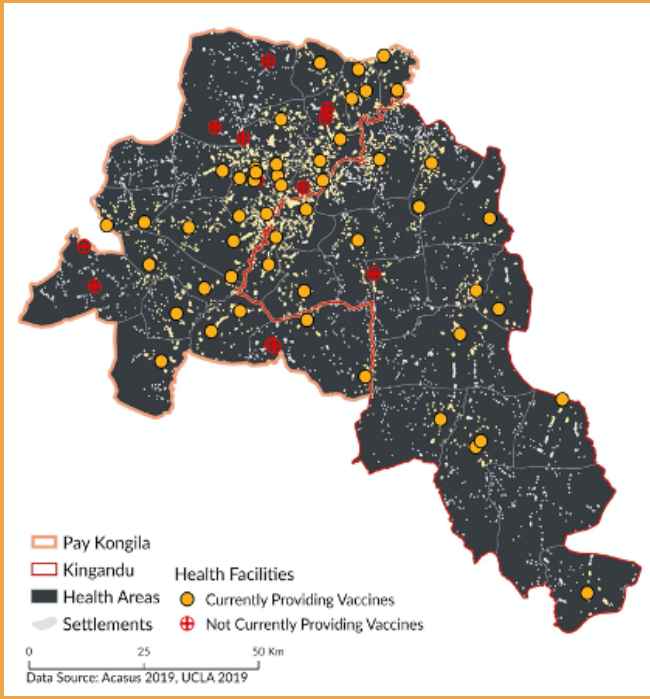
"GRID3 enabled health workers to improve the completeness of the settlements included in their microplans, allowing for additional target populations to be included in future vaccination campaigns."

GRID3 has also improved settlement mapping in eastern DRC by organising participatory cartography sessions. Health workers from a specific health zone—ideally during their monthly meeting or routine activities—sat in on the initial two-day session to identify and map unnamed villages. Two GRID3 mappers with GIS training were provided with all data that had been collected on the region so far (via satellite imagery uploaded onto the trainees' laptops, as well as the feature extractions of settlements whose quality had been improved by US-based GRID3 researchers); they used this data to guide the process. The GRID3 mappers used the local knowledge of the health workers along with their own technical skills to combine these layers and locate settlements. To identify unknown villages or villages whose location and names remained uncertain at the end of the participatory mapping session, GRID3 equipped health workers with smartphones and trained them on an easy-to-use data collection application provided by Novel-T. The health workers returned to their assigned health areas and collected information on whatever settlements remained unnamed. The two GRID3 mappers would then collect these additional data and build the comprehensive settlement layer. Once this work was completed, a validation meeting with the head of the health zone (Médecin Chef de Zone) was held in order to obtain sign-off on new settlement names. By creating a more complete settlement map in the two provinces, GRID3 enabled health workers to improve the completeness of the settlements included in their microplans, allowing for additional target populations to be included in future vaccination campaigns.

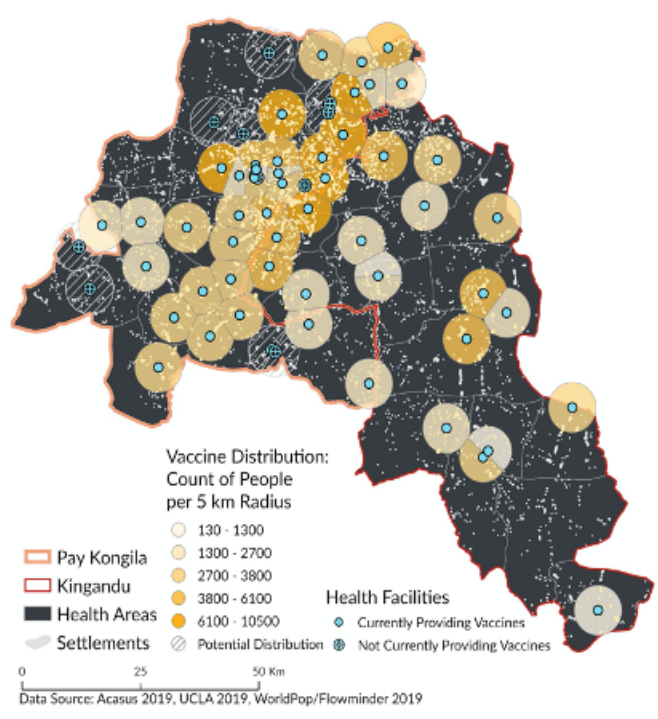
17. PATH: The Bill & Melinda Gates Foundation's logistical/coordinating partner in Haut-Lomami and Tanganyika provinces. See https://www.path.org/where-we-work/democratic-rep-of-the-congo/.
18. Novel-T: technical partner configuring the Android phones and providing the data collection application. See http://www.novel-t.ch/en#Home.

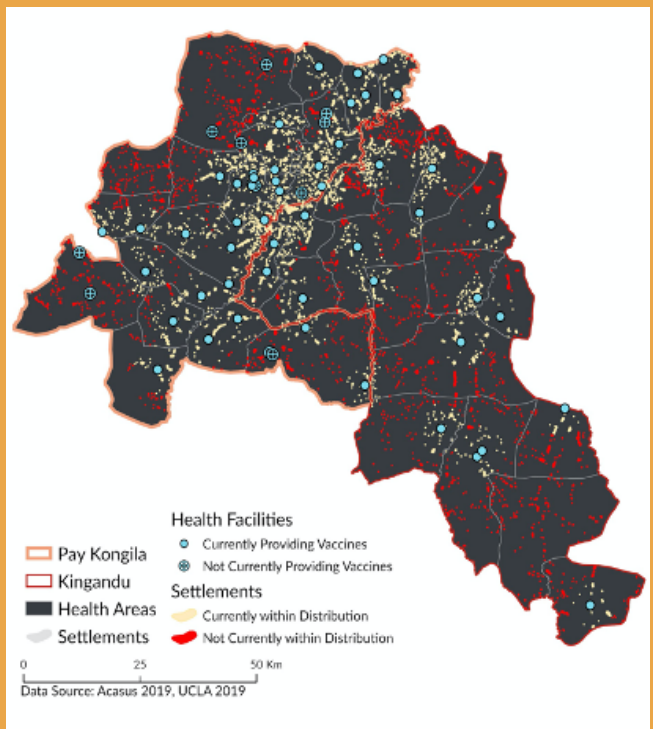
Improving Health Service Delivery in Kwilu, DRC





Efficient planning for vaccine delivery makes use of combined, up-to-date data layers featuring: place names, settlement extents, population estimates, location, and infrastructure available at health centres.





When all information is consolidated, it is possible to observe and quantify the settlements and populations out of range from health centres' catchment areas. With this information, **mobile health posts can be adequately positioned and planned** for effective coverage.

GRID3 has done similar work in Nigeria. [19] Since 2012, the Bill & Melinda Gates Foundation has supported polio eradication efforts via the application of GIS technology—this method has enabled vaccination teams to properly monitor settlement coverage, reduce the number of missed settlements, and improve team performance. Extensive data (including ward boundaries, settlements, and points of interest [20]) were collected in ten northern Nigerian states where there were significant polio outbreaks. This fieldwork—in addition to feature extraction from satellite imagery—enabled GIS microplan maps to be generated for vaccination teams, and allowed settlements to be visualised. The Vaccination Tracking System (VTS) dashboard was developed to follow vaccination teams and produce reports that identified missed settlements, which enabled supervisors to evaluate team performance and identify settlements that needed to be revisited for follow-up vaccinations. A Post Campaign Coverage Survey found that the states where GIS microplans and the VTS were used had better vaccination outcomes and higher population immunity. As a result, data collection efforts were expanded to the remaining states in 2016 and these technologies were also applied in subsequent polio vaccination campaigns. From late 2019 to early 2020, the VTS was enhanced and utilised for additional campaigns, including yellow fever, measles, and other diseases.

Information and research

GRID3's settlements data and associated efforts provide crucial information for development research, including: identifying priority areas for development interventions, conducting market research and locations for potential investments, and generating data about the trajectory and scale of settlement growth.

With GRID3's support, Zambia is making strides towards creating a geospatial layer representing the distribution of its settlements. Improved settlement mapping is a matter of particular importance in the country; Zambia is deeply rural and comprised overwhelmingly of traditionally-owned lands. A small percentage of these lands have been officially mapped. To complicate matters further, it can be difficult to know what to map in the first place (a Zambian village can sometimes be as small as a single household). To address these challenges, GRID3 has worked closely with the Zambian government to create a comprehensive settlement layer that will serve as the most complete repository to date for the country's settlement names and locations.

To create a layer that has government ownership, GRID3 gathered point data from various governmental sources. The majority of settlement names was extracted from a 2010 government mapping exercise, during which POIs nationwide were collected. GRID3 categorised the POIs from the mapping exercise by type in order to extract points that represented villages. GRID3 standardised data from various sources in order to compile it into a single file.

19. For more information on the following, see Touray, K., P. Mkanda, S.G. Tegegn, P. Nsubuga, T.B. Erbeto, R. Banda, A. Etsano, F. Shuaib, and R.G. Vaz. 2016. Tracking Vaccination Teams During Polio Campaigns in Northern Nigeria by Use of Geographic Information System Technology: 2013-2015. Journal of Infectious Diseases (May: S67-72).

20. Points of interest (POI) refer to point features, or geo-referenced points, that depict infrastructure, buildings, and landmarks. These points locate structures and services that are critical to the health and well-being of society, such as health facilities, schools, market places, banks, warehouses, and wells.

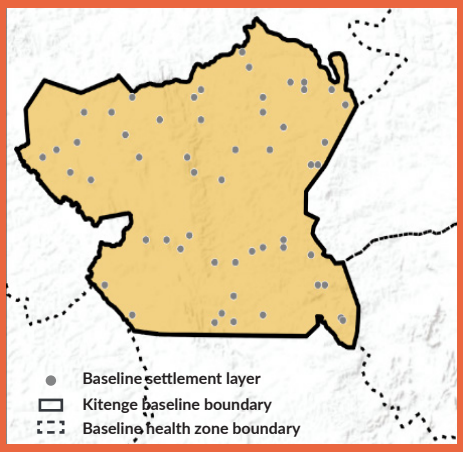
Close collaboration with the Zambian government not only ensures that GRID3 has better access to data and resources, but more importantly fosters Zambian ownership over the project. A wide range of benefits will result from this still-developing set of settlement data, including: assistance with planning for a nationwide census, more effective immunisation campaigns, and increased capacity for disaster response.

Other impacts of GRID3's settlements work:

Programme and intervention design
GRID3's data can help identify optimal locations for new public facilities or public services; establish a foundation for streamlined, more-efficient work plans; and help foster a population-centred design.

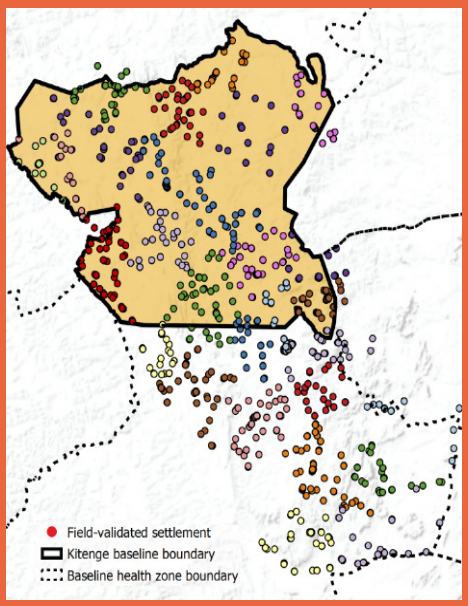
Long-term benefits
Over time, GRID3's data and efforts associated with it can lead to improved long-term development outcomes, including the lowering of various types of mortality and infection rates, increased investment and job growth, increased access to networks and services, and other elements pertaining to the United Nations' Sustainable Development Goals.

Case Study: Strengthening Vaccine Campaigns with Better Settlement Maps in DRC



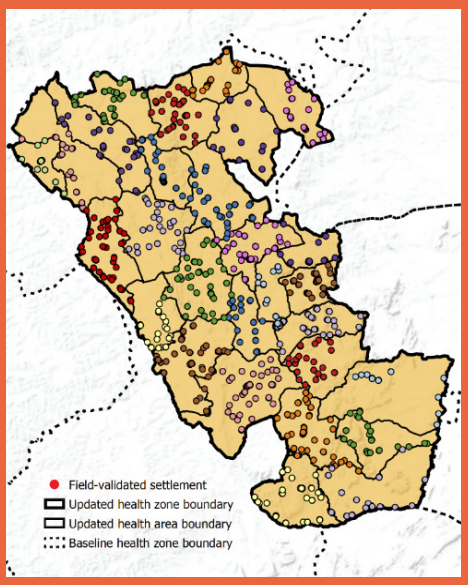
Often, practitioners are forced to make do with settlement maps that are incomplete or inaccurate. This baseline map from eastern DRC was considered inadequate to support a recent polio vaccination campaign.

Sources: Programme Elargi de Vaccination (PEV), Ministry of Public Health, DRC, 2019; Walsh, Chen, Wu, Simbila, et al, 2018; Novel-T (VTS / POI Tracker), 2019; Bureau Central Recensement, 2018; Référentiel Géographique Commun, 2018; Système National d'Information Sanitaire, 2018; UCLA, 2018; UNICEF, 2018; WHO, 2018; Bill & Melinda Gates Foundation, 2019



GRID3 and local partners gathered all sources of GIS settlement data; assessed their quality using satellite imagery and building feature extractions; and consolidated them into a new, harmonised basemap. Then satellite data and machine learning were used to locate previously unmapped settlements. The result was a more reliable settlement map that polio vaccinators could use in their efforts to reach everyone in the region.

Sources: Programme Elargi de Vaccination (PEV), Ministry of Public Health, DRC, 2019; Walsh, Chen, Wu, Simbila, et al, 2018; Novel-T (VTS / POI Tracker), 2019; Bureau Central Recensement, 2018; Référentiel Géographique Commun, 2018; Système National d'Information Sanitaire, 2018; UCLA, 2018; UNICEF, 2018; WHO, 2018; Bill & Melinda Gates Foundation, 2019



Vaccination campaigns are organised according to health boundaries. By identifying which health areas are associated with each settlement, it becomes possible to correct boundary errors and delineate missing boundaries. In this case, a major error in the Kitenge health zone boundary was corrected, and validated health area boundaries were generated. Such clear, validated boundaries enable local managers to plan vaccination campaigns that efficiently reach everybody.

Sources: Programme Elargi de Vaccination (PEV), Ministry of Public Health, DRC, 2019; Walsh, Chen, Wu, Simbila, et al, 2018; Novel-T (VTS / POI Tracker), 2019; Bureau Central Recensement, 2018; Référentiel Géographique Commun, 2018; Système National d'Information Sanitaire, 2018; UCLA, 2018; UNICEF, 2018; WHO, 2018; Bill & Melinda Gates Foundation, 2019

Conclusion

GRID3's activities are adding value to efforts to understand settlements, both by improving data collection/analysis techniques and by improving the data itself. By combining various data, strengthening capacity, and understanding needs, GRID3 is helping to establish new, important data flows for building locations and uses not only for its partner countries, but potentially for other countries at similar stages of development around the world.

Many challenges still need to be addressed in the campaign to collect comprehensive, actionable data on settlements. These include:

- Capturing temporary settlements (such as camps for internally displaced persons, refugees, and migrant workers), temporary population shifts (e.g. post-disaster displacement), long-term population shifts (e.g. urban growth), nomadic settlements, diurnal settlements, and seasonal settlements.

- Distinguishing residential from non-residential populations.

- Reconciling differences between official and local conceptions of places, place names, and settlement boundaries. Currently, GRID3's approach is essentially a spatial conception of settlements and overlooks the on-the-ground experiences of places.

- Negotiating political sensitivities and security issues both within and between countries around places, place names, and settlement boundaries. GRID3 is considering questions such as: What places have been included in past data layers, and should they be included in new ones? What names (and in which language) are places given? How well are indigenous and ethnic minority communities being represented?

In the coming years, GRID3 will be working to address these challenges by continuing to work with partners to make its current data more user-friendly and by seeking new ways of collecting and validating information on settlements.

Acknowledgements and Attribution

Contributing authors

This paper was prepared by Corey Sobel, Jolynn Schmidt, Warren C. Jochem, Kevin Tschirhart, Emilie Schnarr, and Olena Borkovska, with additional inputs provided by Susana Adamo, Sandra Baptista, Sophie Delaporte, Justine Dowden, Paola Kim-Blanco, Matthew Heaton, Attila Lazar, Etienne Leue, Marc Levy, Chris Lloyd, Maxwell Madzikanga, Polly Marshall, Chisimdi Onwuteaka, Silvia Renn, Cathy Riley, Markus Walsh, Greg Yetman, and Apphia Yuma.

Recommended Citation

Center for International Earth Science Information Network (CIESIN), Columbia University; Center for International Earth Science Information Network (CIESIN), Columbia University; Flowminder Foundation; United Nations Population Fund (UNFPA); WorldPop, University of Southampton. 2021. Mapping and Classifying Settlement Locations. Palisades, NY: Georeferenced Infrastructure and Demographic Data for Development (GRID3). https://doi.org/10.7916/d8-gzxf-s834. Accessed DAY MONTH YEAR.

License and Copyright

Works Cited

Barau, I., M. Zubairu, M.N. Mwanza, and V.Y. Seaman. 2014. Improving polio vaccination coverage in Nigeria through the use of geographic information system technology. The Journal of Infectious Diseases (November): S102–S110.

Environmental Systems Research Institute. “What is raster data?”. https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm.

Maxar Technologies, Inc. and Ecopia Tech Corporation. 2020. Ecopia Landbase Africa powered by Maxar. DigitizeAfrica.ai.

National Institute of Standards and Technology. “Probability model”. https://csrc.nist.gov/glossary/term/Probability-model.

Novel-T. http://www.novel-t.ch/en#Home.

Oak Ridge National Laboratory. https://www.ornl.gov/.

OpenStreetMap. https://www.openstreetmap.org/about.

PATH. https://www.path.org/where-we-work/democratic-rep-of-the-congo/.

Programme Elargi de Vaccination (PEV), Ministry of Public Health, DRC, 2019; Walsh, Chen, Wu, Simbila, et al, 2018; Novel-T (VTS / POI Tracker), 2019; Bureau Central Recensement, 2018; Référentiel Geographique Commun, 2018; Système National d'Information Sanitaire, 2018; UCLA, 2018; UNICEF, 2018; WHO, 2018; Bill and Melinda Gates Foundation, 2019.

Touray, K., P. Mkanda, S.G. Tegegn, P. Nsubuga, T.B. Erbeto, R. Banda, A. Etsano, F. Shuaib, and R.G. Vaz. 2016. Tracking Vaccination Teams During Polio Campaigns in Northern Nigeria by Use of Geographic Information System Technology: 2013-2015. Journal of Infectious Diseases (May: S67-72).

UCLA Institute for Digital Research & Education Statistical Consulting. "Zero-Inflated Poisson Regression | R Data Analysis Examples". https://stats.idre.ucla.edu/r/dae/zip/.

United Nations Department of Economic and Social Affairs. 2018. 2018 Revision of World Urbanizaton Prospects.

United States Geological Service. "What is remote sensing and what is it used for?" https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used?qt-news_science_products=7#qt-news_science_products.

Further reading

Dougherty, L., M. Abdulkarim, F. Mikailu, T. Usman, K. Owolabi, K. Gilroy, A. Naiya et al. 2019. From paper maps to digital maps: enhancing routine immunisation microplanning in Northern Nigeria. BMJ Global Health 4.

Jochem, W.C., T.J. Bird, and A.J. Tatem. 2018. Identifying residential neighbourhood types from settlement points in a machine learning approach. Computers, Environment and Urban Systems 69: 104-113.

United Nations Children's Fund. 2018. Guidance on the use of geospatial data and technologies in immunization programs: overview and managerial considerations for in-country strengthening (October).

Weber, E.M., V.Y. Seaman, R.N. Stewart, T.J. Bird, A.J.Tatem, J.J. McKee, B.L. Bhaduri, J.J. Moehl, and A.E. Reith. 2018. Census-independent population mapping in northern Nigeria 204: 786-798.

Paper Version History

- Updated Table of Contents on page 2
- Removed map under ***Polio immunisation microplan for Dundubus Ward in Jigawa State, Nigeria*** on page 19. It is a measles fixed post cluster map created by Novel-T.
- Updated logos and About the Partners section on page 25

GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development) works with countries to generate, validate and use geospatial data on population, settlements, infrastructure, and subnational boundaries. GRID3 combines the expertise of partners in government, United Nations, academia, and the private sector to design adaptable and relevant geospatial solutions based on capacity and development needs of each country.

About the partners: The GRID3 programme is funded by a grant from the Bill & Melinda Gates Foundation and the United Kingdom's Foreign, Commonwealth & Development Office (FCDO). It is implemented by Columbia University's Center for International Earth Science Information Network (CIESIN), the United Nations Population Fund (UNFPA), WorldPop at the University of Southampton and the Flowminder Foundation.

grid3.org info@grid3.org

@GRID3Global