COLUMBIA CLIMATE SCHOOL
CENTER FOR INTERNATIONAL EARTH SCIENCE
INFORMATION NETWORK

Data Release Notes

Name of the dataset	GRID3 COD - Settlement Extents v3.0
Name of the file	GRID3_COD_settlement_extents_v3_0.gpkg GRID3_COD_settlement_grid_v3_0.gpkg
Date of data release	June 28, 2024
File formats	OGC geopackage
Dataset version	v3.0
Abstract	The *GRID3 COD - Settlement Extents v3.0* consists of a geographic representation of settlements in the Democratic Republic of the Congo, in two forms: 1) settlement polygons, and 2) spatial points depicting the centroids of settled grid cells at 3-arc seconds (or ~100 meters) contained within settlement polygons. Both layers include attributes as described in the codebooks. Data inputs and methodology to derive the 2 layers is described in this document.
Dataset citation	Center for International Earth Science Information Network (CIESIN), Columbia University. 2024. *GRID3 COD - Settlement Extents v3.0. New York: GRID3. https://doi.org/10.7916/4wqv-5t02. Accessed [DAY MONTH YEAR].*
Terms of use	Users are free to download, store, access, use, copy, adapt, transform, alter, arrange, build upon, distribute and transmit this work and any derivative works. Attribution of the source must be provided, and further distribution of this work or derived work must maintain the same terms of data use and license as set forth in this Terms of Use. Copyright 2024. The Trustees of Columbia University in the City of New York.
Data license	The data and accompanying document are licensed under a Creative Commons Attribution-ShareAlike 4.0 International, CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0) and specified in legal code (http://creativecommons.org/licenses/by-sa/4.0/legalcode)
Contacts and data queries	The authors of this dataset appreciate feedback regarding the data, including suggestions, discovery of errors, difficulties in using the data, and format preferences. For dataset-related questions, please send an email to: info@ciesin.columbia.edu

I. Methodology

A. Data inputs

We consolidated 11 data sources that provided a spatial extent baseline (see Table 1). We generated metrics of interest such as data stack count, building count, building area, and other covariates (see Table 2), and aggregated the results using 3-arc second (~100-meter) grid cells aligned with WorldPop's global population grid[1]. The alignment with WorldPop's grid will allow further aggregations of population at the settlement level and without spatial inconsistencies. Each grid cell was assigned a unique ID (variable = grid3_id) to keep track of each data input's lineage.

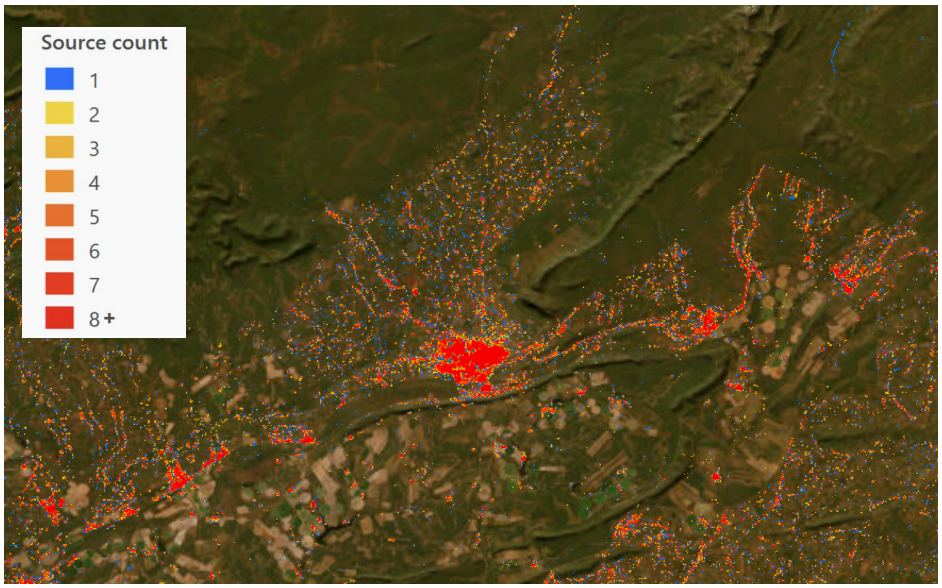
Table 1. Data inputs

Name	Data type/ format	Release year	Input data year	Resolution (in meters)
1- Data collected during anti-malaria bednet campaigns in the DRC (PLNP)	Household points, vector	2021 - 2023	2018-2023	n/a
2- Google Building Footprints v3	Building footprints, vector	2023	2023 and earlier	n/a
3- Microsoft Building Footprints	Building footprints, vector	2022	2014-2023	n/a
4- OSM Building Footprints	Building footprints, vector	2023	varies	n/a
5- ESRI Land Cover, urban class	Raster	2023	2023	10 m
6- Global Human Settlements Layer (GHSL)	Raster	2023	2018	10 m
7- Google's Dynamic World, urban class (DW)	Raster	2023	2023	10 m
8- World Settlement Footprint (WSF)	Raster	2019	2016	10 m
9- High Resolution Settlement Layer (HRSL)	Raster	2020	2011-2019	~30 m (1 arc-second)
10- Digitized building points using Maxar's imagery around lake areas in Haut-Lomami and Tanganyika provinces, DRC (CIESIN)	Building points, vector	unpublished	2022- 2024	n/a
11 - GRID3 DRC Health facilities, Schools, and other points of interest (POI) datasets collected in the field (GRID3)	Point data, vector	2021-2022	2019-2022	n/a

[1] WorldPop, University of Southampton https://www.worldpop.org/

The consolidation of all data sources resulted in an output with many millions of grid cells which showed no ambiguity to detect urban areas given that most data sources agreed on their location and spatial extent. However, a high variability across sources was found in rural areas, especially those representing hamlets (see figure 1). From past experiences, we have learned about the importance of mapping small settlements or hamlets with relatively high confidence in order to plan for in-country interventions. Therefore, our next step was to filter out grid cells with low confidence of representing true settled areas.

Figure 1 - Examples of the data stack by data source count.



B. Classification modeling to predict settlement probability

We generated a classification model (XG Boost) to obtain a probability value for any given grid cell to be settled. Training data was developed by randomly sampling ~9,000 grid cells (weighted by stack count to avoid bias due to an unbalanced distribution[2]) that were visually inspected 3 times by independent mappers against Google imagery. When discrepancies between settled/ non-settled grid cells were found (e.g. 2 mappers agreeing and 1 disagreeing), CIESIN made a final determination.

Table 2 - Covariates used to predict the presence of false positives among grid cells assumed as part of a settlement.

Covariates	**Definition**	**Data inputs (**see notes for key).									
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Stack_count	Count of data inputs classifying the grid cell as being settled.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
[datainput] _bld_count	Count of buildings' centroids contained within a grid cell.	No	Yes	Yes	Yes	No	No	No	No	No	No
[datainput] _bld_area	Sum of built-up area from buildings' centroids contained within a grid cell.	No	Yes	Yes	Yes	No	No	No	No	No	No
google_max _confidence	Maximum confidence from google v3 BF.	No	Yes	No	No	No	No	No	No	No	No
[datainput] _grid_count	Count of grid cells from raster data inputs, based on the specific resolutions.	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No
Grid_count _[radius]	Count of contiguous grid cells, within 100, 300, and 500 meters radius, where at least one data source classifies the grid cell as settled.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
source_count _[radius]	Within the same contiguity as above (100, 300, and 500 meters), count of data inputs where at least one data source classifies the grid cell as settled.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dist_to _[variable]	Distance from the grid cell centroid to the nearest: road, coast line, inland water body (i.e. river, lake), in meters.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Road_lenght _m	Total length of roads by grid cell, in meters	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
dist_roads_ intersections	Distance from the grid cell centroid to the nearest road intersection, in meters	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

[2] Not all the data sources have a consistent national coverage. For instance, there are specific areas where HRSL or Microsoft have data gaps. In addition, the PLNP or GRID3 datasets are not comprehensive for the entire country but only a few provinces.

Covariates	Definition	Data inputs (see notes for key).									
		A	B	C	D	E	F	G	H	I	J
dist_to_[land use/ land cover type]	Distance from the grid cell centroid to the nearest: forest area, bare ground, cropland, wetland, and rangeland.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: A- PLNP, B- Google, C- Microsoft, D- OSM, E- ESRI Land Cover, F- GHSL, G- Google's DW, H- WSF, I- HRSL, J- GRID3. CIESIN's building points around the Haut-Lomami/ Tanganyika lake area were not considered as part of the data inputs for the probability model. The road's data source was a combination of OSM and Facebook. The inland water bodies data source used was ESRI's land cover water class.

With respect to building count and building area metrics, we took the centroid of each building footprint as the way to determine where the building belonged to. In that sense, the building count is really a count of building centroids, dismissing any portions of any given building crossing over more than one grid cell. Likewise, the building area metric considered the entirety of the building's area, irrespective of whether portions crossed over more than one grid cell. The latter resulted in some grid cells with built-up areas higher than the maximum possible area value for a grid cell (100 x 100 m = 10,000 m2). At present, we opted to leave these values as they are data values considered for settlement classification.

Figure 1 - XG Boost model result

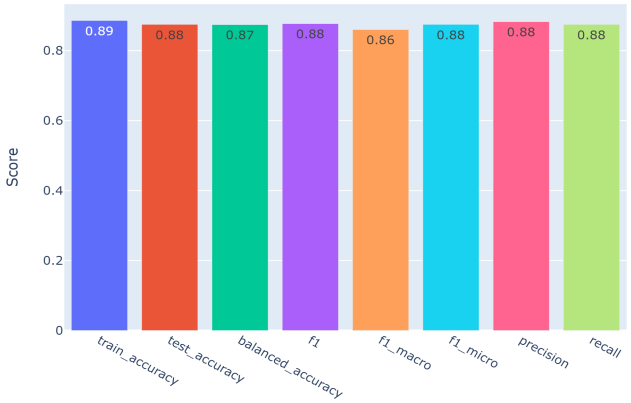
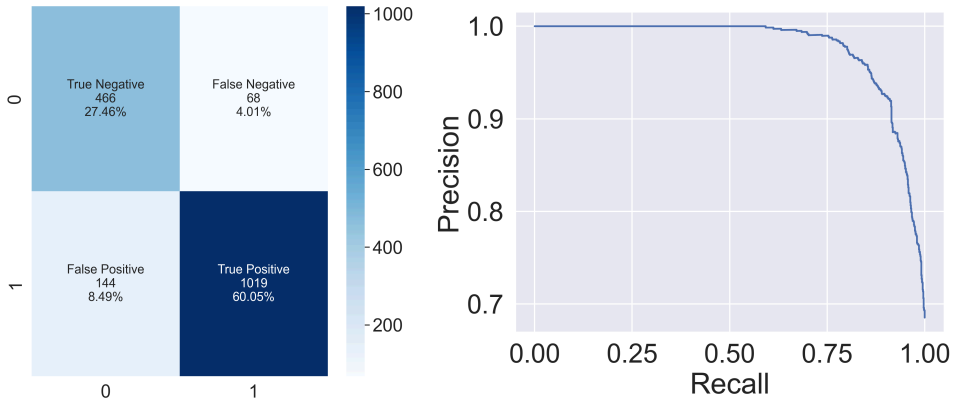


Figure 2 - Confusion matrix, and precision-recall curve from XG Boost model



C. Treatment of grid cells with low probability values

We assumed that grid cells with a probability value of less than 0.5 were less likely to represent true settlements. Therefore we removed grid cells with probability values of less than 0.5. Figure 3 shows representations of the distribution by stack count and by data input between grid cells under this assumption. Figure 4 shows the distribution of grid cells by stack count and data input.

Figure 3 - Distribution of settled/ not settled grid cells by stack count (left) and data input (right). Green = true positive: grid cells with a prediction value of 0.5 or above, assumed to be part of a settlement; red = false positive: grid cells with a prediction value of less than 0.5, assumed to not be part of a settlement.

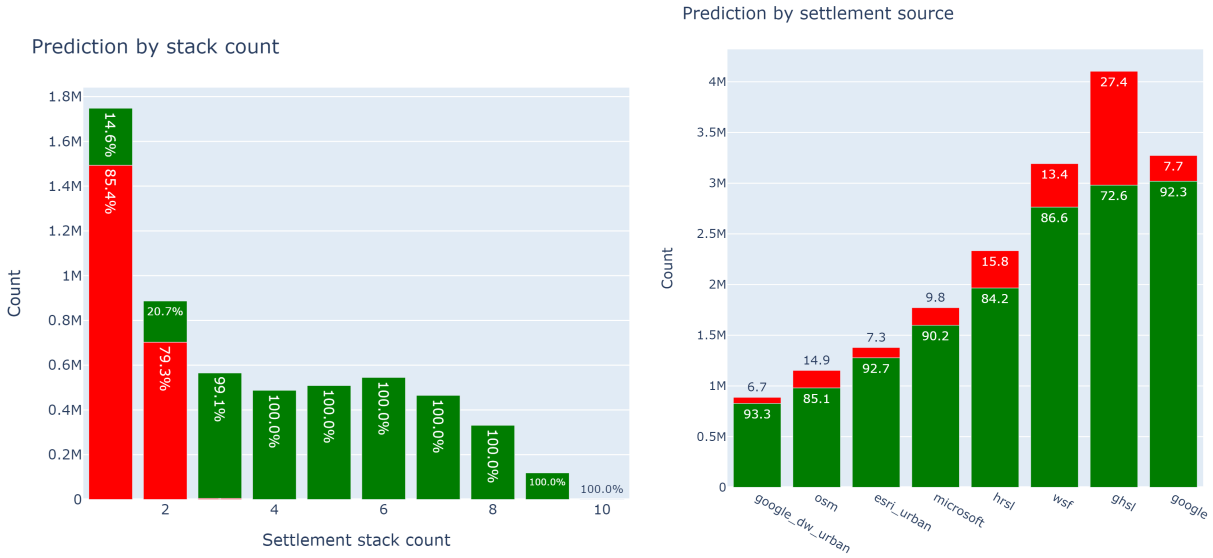
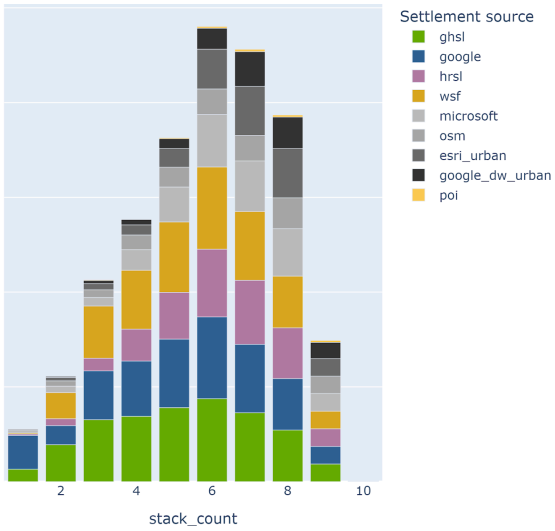


Figure 4 - Distribution of grid cells predicted as *settled* by stack count and data input.



D. Treatment of grid cells with raster-based provenance.

After conducting random checks of the preliminary output we noticed that grid cells conformed exclusively by raster data inputs (i.e. ESRI urban class, GHSL, DW, or WSF) were not rendering satisfactory results, even though the predicted values equaled 0.5 or above. Some of these grid cells were located along the outskirts of settlements; others were pointing out to completely new (and erroneous) locations. Therefore, we decided to conduct a second filtering and removed grid cells that belonged to this category, irrespective of their predicted probability value.

E. Treatment of grid cells with field data provenance.

In that same token, after inspection of results from the following data sources: PLNP and GRID3, we observed that the model assigned low probability scores to grid cells when these sources were the only ones present. Because the major predictor is the variable "stack_count" these grid cells were penalized without considering how these data were generated (i.e. via fieldwork). Therefore, we decided to add all grid cells containing data points from PLNP, CIESIN or GRID3, irrespective of the predicted probability value[3]. Our assumption is that data generated from fieldwork has an intrinsically higher value and therefore a higher probability of representing a true settlement than data generated from satellite imagery or machine learning processes.

It is worth noting that for grid cells with data sources including PLNP, GRID3, or CIESIN, the probability value was manually set to null. Even though, empirically, the probability value of this class equals 1, at this time we opted to preserve the model-based probability only.

F. Building count and building area value assignment for the settlement grid.

The final step includes the assignment of building count and building area values for each grid cell, as we have a mixed bag of cases: some grid cells contain more than one source with building footprints; some grid cells contain only building counts but not building areas; other grid cells do not have either value.

We decided to assign building count and building area values based on data provenance in ranked order; at present, we did not produce composite values for these metrics. In that sense, when a grid cell contained Google building footprints, we assigned values from Google. If no Google data was present, we used Microsoft. If no Microsoft data was present, we used OSM. When no OSM values were present, we used CIESIN (only for lakeshore grid cells in Haut Lomami and Tanganyika). When no building source was present, we imputed the value using zonal statistics taking into account the values of adjacent grid cells within 200 meters. If no neighboring value was found at 200 meters, we imputed the global median value for building count (for DRC, the median global building count value is 3), and building area (for DRC, the median building area is 90 m2).

Final counts for imputed values are as follows:

[3] CIESIN's building points around the Haut-Lomami/ Tanganyika lake area were not considered as part of the data inputs for the probability model.

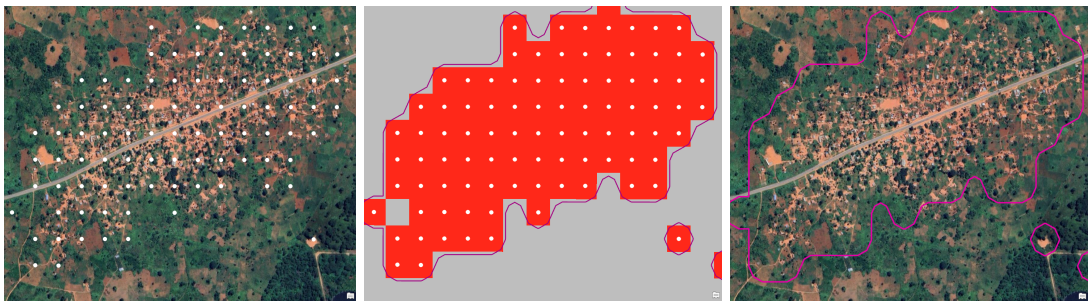
Total count of settled grid cells	3,595,709 (100%)
Count of settled grid cells with imputed building count or building area values	351,958 (9%)
Count of settled grid cells with imputed building count or building area values using zonal statistics at 100 meters	203,890 (6%)
Count of settled grid cells with imputed building count or building area values using zonal statistics at 200 meters	48,068 (1%)
Count of settled grid cells with imputed building count or building area values using global median values	100,000 (3%)

G. Geometric aggregation of settled grid cells

Our next step is to generate settlement polygons (or extents) out of the settled grid cells. We used 2 inputs to delineate the border and further classify each polygon: a raster of settled grid cells and a raster of building counts. We utilized the contour shell-up function in ArcGIS Pro[4] to generate the desired outputs.

For the delineation of settlement polygons, we used the raster of settled grid cells. We figured that assigning a fixed cell value all across as opposed to taking any of the metrics produced above produces the best results and allows for a continuous integration of diagonal corners and smooth rendering of edges. After some trial-and-error, the grid cell value = 2 produced the best output. For the estimation of built-up density, we used the raster of building counts. We utilized the same contour function in ArcGIS Pro, and we took the value calculated for building density. Figure 5 shows an illustration of the process.

Figure 5. Visual representations of: 1) settled grid cell centroids; 2) contour delineation with underlying centroids/ raster; 3) settlement polygon or extent, in a given settlement



Settled grid cell centroids Contour delineation Settlement extent

H. Estimating a probability value for settlement extents

The settlement extent probability is the output of the combined probability from model-based probabilities at the grid cell level. If we consider P(S) to be the probability of settlement, and $P(N_i)$ to be

[4] https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/contour.htm

the probability that a given grid cell has no buildings, then the probability that all cells have no buildings is the product of all $P(N_i)$. Therefore, the probability of settlement is one minus the product of $P(N_i)$.

In the current version, the settlement extents probability ranges between 0.5 and >1. Given that we filtered out settlement grids with probability values of less than 0.5, then the minimum probability value for any given settlement = 0.5. Likewise, given that we took the model-based probability value as is (and did not manipulate it to assign value =1 to field data, regardless of the model output), the maximum settlement probability is close to but less than 1.

I. Categorical classification of settlements based on building density

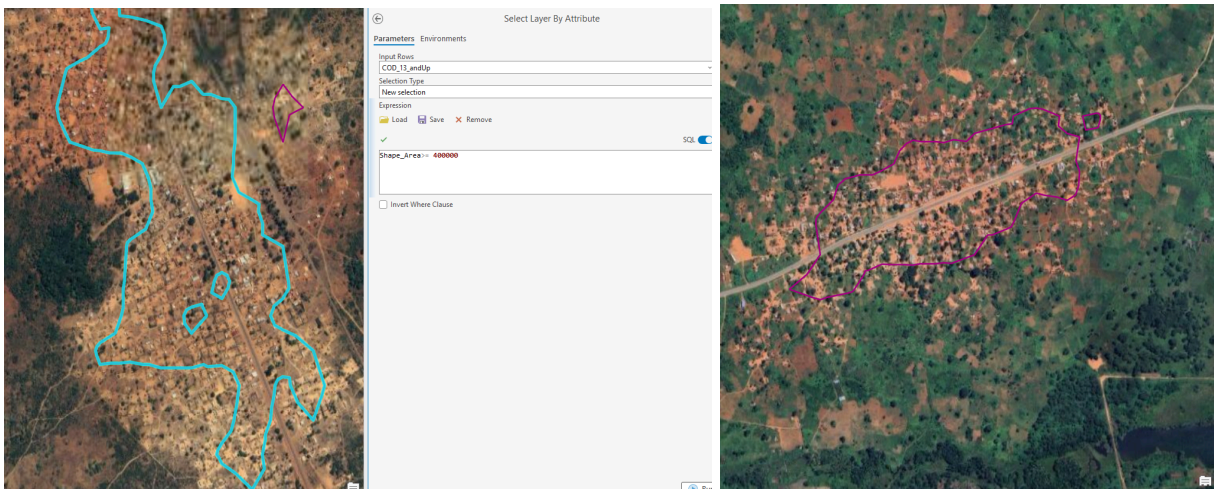
As with previous versions, the GRID3 classifies settlements into 3 categories: built-up areas, small settlement areas, and hamlets. To generate these data fields, we utilized the contour outputs as described in section F.

A built-up area (BUA) is generally an area of urbanization with moderately-to-densely-spaced buildings and a visible grid of streets and blocks. We define BUAs as areas of at least 400,000m2 (or 40 hectares) that have a building density of at least 13 buildings per hectare.

A small settlement area (SSA) is defined as an assemblage of 50 or more buildings that are not classified as BUA. SSAs usually consist of semi-urban areas with moderately urbanization patterns, in some cases resulting from urban sprawling over time. An illustrative example between BUA and SSA is depicted in figure 6, below.

A hamlet is defined as areas with a building count of up to 49. Usually these areas constitute rural, low-density settlements. The isolated, small hamlets are usually classified as hard-to-reach settlements.

Figure 6. Illustration of BUA and SSA. In both instances the density value = 13 buildings/ hectare. However, the image in the left has a spatial extent of more than 40 hectares, whereas the image in the right denotes a settlement of size <40 hectares. The window in the middle shows the selection query.



II. Dataset Description

The *GRID3 COD - Settlement Extents v3.0* consists of 2 datasets:

- GRID3_COD_settlement_extents_v3_0.gpkg: a spatial layer representing settlement polygons.
- GRID3_COD_settlement_grid_v3_0.gpkg: a spatial layer representing the centroids of settled grid cells.

Codebook for the *GRID3_COD_settlement_extents_v3_0* data file

Variable Names	**Definition**	**Type**
OBJECTID	Software- generated unique code	numeric
country	Country name	text
iso3	Three letter ISO code	text
building_count	Total building count within the settlement extent	numeric
building_area	Sum of building areas within the settlement extent	numeric
type	Categorical settlement classification: Built-up area (BUA), Small settlement Area (SSA), or Hamlet	text
probability	Combined probability value of being a settlement for any given settlement polygon.	numeric
date	Date when the data product was derived or edited last	text
source	Name of the data producer	text
mgrs_code	Unique name generated using the Military Grid Reference System	text

Codebook for the *GRID3_COD_settlement_grid_v3_0* data file

Field name	**Definition**	**Type**
OBJECTID	Software- generated unique code	Numeric
Shape	Software- generated geometric type	Text
country	Country name	Text
iso3	ISO3 country code	Text
building_count	Count of building centroids within the grid cell, as per data source ranking	Numeric
building_area	Sum of building areas in m2 within the grid cell, as per data source ranking	Numeric
building_count_source	Data source used to estimate a total building count	Text
building_area_source	Data source used to estimate a total building area	Text
probability	Model-based probability value of grid cell to be considered settled.	Numeric
google	Boolean value when building footprints from Google are available for the grid cell	Boolean

Field name	Definition	Type
google_building_count	Building count using Google building footprints only. Buildings are considered in the count when the centroid is completely within the grid cell. When Google building footprints are not available within the grid cell, the value is set to null.	Numeric
google_building_area	Building area in m2 using Google building footprints only. A building's area is considered in the grid cell aggregation when the building's centroid is completely within the grid cell. When Google building footprints are not available within the grid cell, the value is set to null.	Numeric
microsoft	Boolean value when building footprints from Microsoft are available for the grid cell	Boolean
microsoft_building_count	Building count using Microsoft building footprints only. Buildings are considered in the count when the centroid is completely within the grid cell. When Microsoft building footprints are not available within the grid cell, the value is set to null.	Numeric
microsoft_building_area	Building area in m2 using Microsoft building footprints only. A building's area is considered in the grid cell aggregation when the building's centroid is completely within the grid cell. When Microsoft building footprints are not available within the grid cell, the value is set to null.	Numeric
osm	Boolean value when building footprints from OSM are available for the grid cell.	Boolean
osm_building_count	Building count using OSM building footprints only. Buildings are considered in the count when the centroid is completely within the grid cell. When OSM building footprints are not available within the grid cell, the value is set to null.	Numeric
osm_building_area	Building area in m2 using OSM building footprints only. A building's area is considered in the grid cell aggregation when the building's centroid is completely within the grid cell. When OSM building footprints are not available within the grid cell, the value is set to null.	Numeric
ciesin	Boolean value when building points from CIESIN are available for the grid cell.	Boolean
ciesin_building_count	Building count using CIESIN building points only. Buildings are considered in the count when the point is completely within the grid cell. When CIESIN building points are not available within the grid cell, the value is set to null.	Numeric
grid3	Boolean value when the presence of spatial points represented as POI collected in the field are available for the grid cell.	Boolean
plpn	Boolean value when the presence of spatial points representing households from PLNP are available for the grid cell.	Boolean
longitude	Longitude in decimal degrees	Numeric
latitude	Latitude in decimal degrees	Numeric

III. Known Data Limitations and Disclaimer

CIESIN, Columbia University, and its co-authors follow procedures designed to ensure that data disseminated by the project are of reasonable quality. If, despite these procedures, users encounter apparent errors or misstatements in the data, they should contact CIESIN, info@ciesin.columbia.edu.

CIESIN, Columbia University, its co-authors, and their sponsors do not guarantee the accuracy, reliability, or
completeness of any data provided. We provide these data without warranty of any kind whatsoever, either expressed or implied, and shall not be liable for incidental, consequential, or special damages arising out of the use of any data provided.

IV. Acknowledgments

CIESIN thanks the following institutions that provided input and/or assistance during the development of this data product:

Acasus
Bluesquare
Centers for Disease Control and Prevention, CDC
Division of National Health Information System, DSNIS
Elongated Programme of Immunization, EPI/ PEV
IMA World Health
Kinshasa School of Public Health, ESPK
Ministry of Public Health, Government of the DRC
National Malaria Elimination Programme, PLNP
National Sleeping Sickness Control Programme, PLNTHA
Soins de Santé primaires en milieu Rurale, SANRU
WorldPop Research Group, University of Southampton

Funding for the development and dissemination of this dataset was provided by GRID3 Inc under the Bill & Melinda Gates Foundation's project INV-044979