Degree of Morphological Informality

1.0 Overview

The degree of morphological informality sub-domain aims to capture the degree to which a settlement is characterised by a highly formal structure (e.g., spaced buildings and gridded road layouts), in contrast with organic, informal structures.

1.1 How is this sub-domain different to the previous version?

In ‘Phase 1’ of the project (June 2023 - February 2024), we produced three versions of deprivation sub-domain models (overall deprivation, small dense settlements and irregular layout). We will combine the best practices of the three individual models and generate a new model in ‘Phase 2’ (Phase 2.1: March - June 2024; Phase 2.2: July- November 2024), which we will refer to as morphological informality.

The crude deprivation model generated in Phase 1 (overall deprivation sub-domain model), used a deep learning model to analyse Sentinel imagery and Google buildings datasets to provide estimations on whether an area (100m cell) was deprived or non-deprived. The binary output was useful in the first phase to provide an initial understanding of areas within our pilot cities that were more deprived than others. However, the outputs did not take into consideration areas that were mixed-use (a mixture of deprived and non-deprived), such as the peripheries or fringes of slums and more informal areas. This meant that during our validation feedback sessions, users found it difficult to provide feedback within contentious/fringe areas. In Phase 1, our stakeholders (including community members) had issues interpreting the methods used to model overall deprivation and how this was translated into a binary output. We also experienced issues explaining the overall deprivation outputs to our SH groups as deprivation was interpreted differently by participants.

To overcome these issues, the next version of the model will produce outputs with three classes of urban areas (highly informal, moderate informal and formal),to help improve interpretability of outputs and enable users to interact with/validate outputs more effectively[1]. The datasets and modelling process used to calculate the high, medium and low categories are outlined in sections 2.0 and 3.0.

[1] 1 Li et al. (2023) https://doi.org/10.1016/j.scs.2023.104863

The model of morphological informality will also consider additional variables to more holistically represent our pilot cities and produce more impactful outputs[2][3][4]. The model will include the following variables:

- Road connectivity (Proxy for services)
- Building density
- "Organic"/irregular layout
- Population density (threshold/controlled variable)

1.2 Modelling approach

In the first part of the Phase 2 (April - end of June 2024), the model will follow a 'rule based' (fig 1). This is to help us to identify our three classes of morphological informality (high, medium and low) as outlined above. Once these classes have been identified, it may then be useful to disseminate the outputs and conduct another participatory/validation activity to attain feedback on areas that the model has classified (in)accurately. In the second part of Phase 2 (July - November 2024), the outputs will be validated via multi-sectoral workshops. These workshops will be developed in more detail once the activities in Phase 2.1 are complete.

1.2.1 Why are we not using more sophisticated modeling techniques such as deep learning and machine learning?

There are arguments for using a rule-based or a deep learning model within the first part of Phase 2. For example, the input data used for rule-based approaches is not updated frequently enough, whilst the data used within deep learning models requires high-resolution and sometimes expensive satellite imagery. However, recent discussions with stakeholders have highlighted that outputs generated via a rule based approach (or simple RF model) are more easily communicable with national governments, decision makers and community members[5][6]. Furthermore, deep learning models require a large amount of high quality reference datasets to produce accurate outputs. It is therefore essential to first define high quality reference datasets for our high-med-low classes and have these validated from those with lived experience of our case studies. Following a rule-based/random forest approach with the first part of Phase 2 will enable us to achieve this.

[2] Kohli et al. (2012) https://doi.org/10.1016/j.compenvurbsys.2011.11.001

[3] Kraff et al. (2019) doi.org/10.1109/JURSE.2019.8808978

[4] Taubenbock et al. (2018) doi.org/10.1016/j.apgeog.2018.02.002

[5] Li et al. (2023) https://doi.org/10.1016/j.scs.2023.104863

[6] Wang et al. (2023) https://doi.org/10.1016/j.landurbplan.2023.104691

2.0 Data

Dataset Description

We will utilise a global dataset of “city segment” boundaries being developed by CIESIN. These city segments are derived from OSM roads and rivers, NOT intersected by a major road or water body, and each contains a minimum of 500 people per GHS-POP R2023A. Road (OSM) and building (Google Footprints) metrics are then calculated within each city segment, and used to label city segments as highly informal, moderately informal, and formal. City segments and whole cities with limited OSM data will be marked as NA.

- Unit of Analysis: 100m grid defined by FUA see section 3.1.
  - Document and use the CASA Methodology for Uncrossable Features + City Segments
- GHS-UCDB
  - Identifying attributes, eg City Name, Country Name, Region Name
  - Exclude cities with insufficient OSM data coverage. Herfort et al 2023 classify completeness of OSM building data by GHS-UCDB.
- Population metrics
  - Segment total population based on GHS-POP 2023
  - Segment population density (metre/hectare) based on GHS-POP 2023
  - City total population
  - City total population classified as follows based on other definitions [7] : <250K, 250-500K, 500K-1M, 1M-10M, 10M+
  - City 2020-2025 annual population growth rate
- Road metrics: OSM
  - Length of internal streets within blocks connected to external street network (metres) [Million Neighborhoods]
  - Length of internal streets within blocks NOT connected to external street network (metres) [Million Neighborhoods]
  - Distance to roads (metres)
  - Intersection of roads
  - Topography?

[7] “Small and medium sized” cities have been defined in wildly different ways in Africa vs Asia by different authors.

- < 5 million in Asia (Birkmann et al 2016, World Bank)
- < 1 million in Asia (UNESCAP, Dahiya 2014)
- < 500,000 in Asia (Hugo 2019) and Global North (OECD)
- < 300,000 in Africa (UNECA)
- < 250,000 in Africa (Africapolis, Satterthwaite 2017). Satterthwaite 2017 describes small Africa cities as 20,000+ pop. Could supplement GHS-UCDB with Africapolis to include Africa cities 20-50K

  - Depthmapx at UCL has some good road metrics that we could follow to produce a map of spatial elements and connect them via relationship (for example, intervisibility, intersection or adjacency)
- Building metrics (more info on individual metrics in appendix 1): Google Buildings
  - Area
  - Perimeter
  - Compactness
  - Corners
  - Squareness
  - Equivalent rectangular index
  - Elongation
  - Centroid corners
  - Orientation
  - Alignment
  - Cell alignment
  - Neighbour distance
  - Mean interbuilding distance
  - Building adjacency
  - Longest axis length
  - Circular compactness
  - Area ratio
  - Neighbours
  - Covered area
  - Block count
- Environmental metrics TBC WP2/3 mtg 21/5/24
  - Greenspace/NDVI

2.1 Reference datasets

Option A

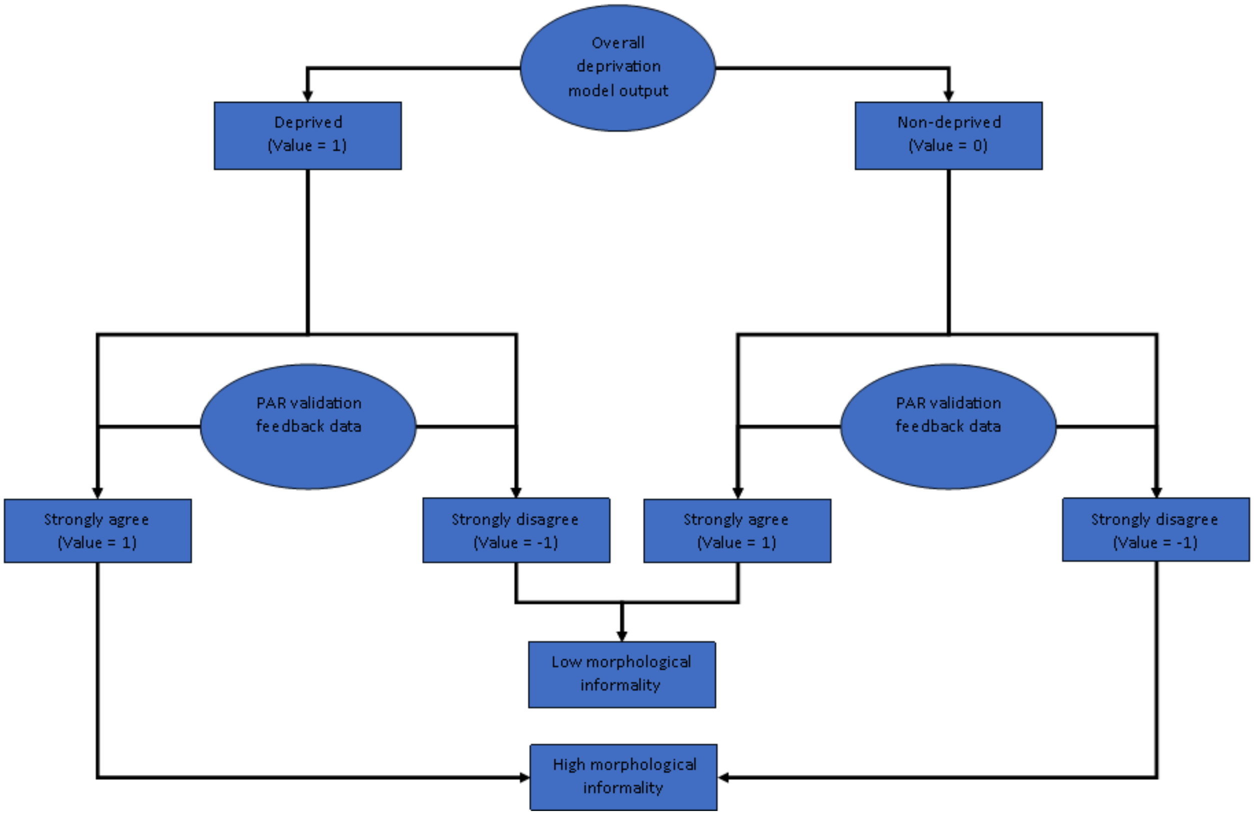

Overall deprivation model output
Deprived (Value = 1)
Non-deprived (Value = 0)
PAR validation feedback data
PAR validation feedback data
Strongly agree (Value = 1)
Strongly disagree (Value = -1)
Strongly agree (Value = 1)
Strongly disagree (Value = -1)
Low morphological informality
High morphological informality

Fig x. Combining model output of overall deprivation with PAR validation feedback data to produce reference data for high and low morphological informality.

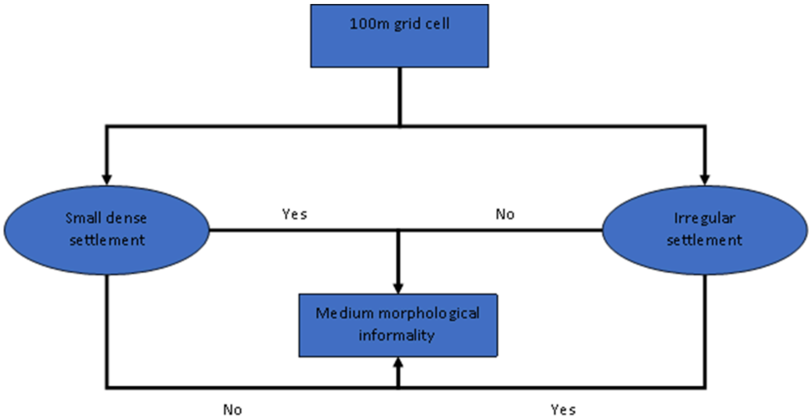


Fig x. Combining irregular settlement layout and small dense structure model outputs to produce reference data for medium morphological informality.

Option B

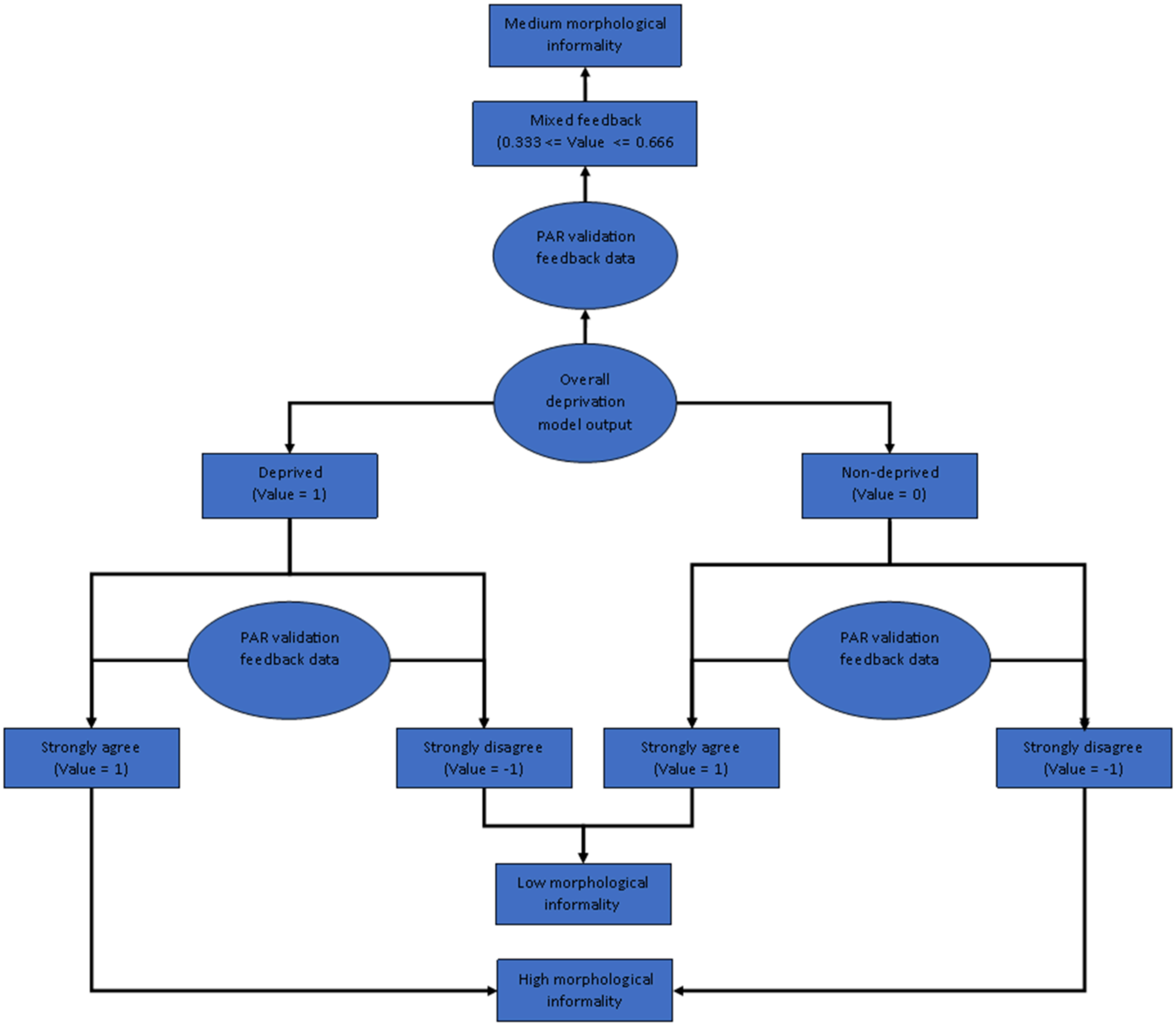

Medium morphological informality
Mixed feedback (0.333 <= Value <= 0.666
PAR validation feedback data
Overall deprivation model output
Deprived (Value = 1)
Non-deprived (Value = 0)
PAR validation feedback data
PAR validation feedback data
Strongly agree (Value = 1)
Strongly disagree (Value = -1)
Strongly agree (Value = 1)
Strongly disagree (Value = -1)
Low morphological informality
High morphological informality

Fig x. Combining overall deprivation model output with PAR validation feedback data to generate three classes of high (value = 1), medium (-0.666 <= value <= 0.666 0.333) and low (value = -1). CHANGE COLOUR TO WHITE.

*****Option C: Recommended option*****

The model outputs for small dense settlements and irregular settlement layout will be combined with PAR validation data to produce reference datasets for high, medium and low morphological informality (similar to the methods described within the figure shown in option B above). This method is recommended due to its explainability/interpretability by end users (using morphometrics), scalability/reproducibility for future modelling practices (using model outputs and validation data) and quantity of available validation data for unique cells (1792 total PAR feedback for SDS and ISL outputs: 1015 for Kano, 238 for Lagos and 645 for Nairobi). It is likely that Kano will be used to generate and train the first model due to its higher number of unique cells/validated outputs. This will be discussed within the backlog review meeting.

Table x. The small dense structures (SDS) and Irregular settlement layout (ISL) outputs with the PAR feedback for each model output (SDS PAR and ISL PAR). Note that for the model outputs 0 represents no and 1 represents yes e.g. for SDS model 1 would reflect a cell that has small dense structures and 0 would reflect a cell that does not have small dense structures. The values within the PAR cells are 1 agree and 0 is disagree.

SDS model	SDS PAR	ISL model	ISL PAR	Aggregated number	Final categories	Nairobi example
1	1	1	1	2	High	309
0	0	0	0	0	High	
1	1	0	1	1	Med	50
0	1	1	1	1	Med	
0	1	0	0	1	Med	
0	0	0	1	1	Med	
1	0	1	0	0	Low	286
0	1	0	1	0	Low	

**Grant to look at 10-15 pixels of each and see if they are accurate.

Accuracy: high and low is good but medium is not good

Feature importance

Partial dependence plot

SHAP

Example: If a cell has small dense structures and the validator agrees AND has irregular settlement layout that the validator agrees, then this cell is categorised as having High morphological informality.

3.0 Analysis Plan (option A)

3.1 Functional Urban Area

Administrative Areas: Geographical regions defined by government boundaries, such as cities or districts, established for legal or administrative purposes.

Functional Urban Areas (FUA): Areas determined by actual urban sprawl and human activities, encompassing the core city and economically or socially integrated surrounding regions.

Using Administrative delimitations for spatial analysis in urban planning can be limiting because the boundaries are usually determined by historical, political, or administrative decisions and may not necessarily reflect the actual patterns of human settlement or economic activity. Table x demonstrates that often less than half of these areas are actually urbanised.

Table x: Summary of administrative extensions per study city, and of unbuilt-up areas within the administrative boundaries.

City	**ADM extension**	**Non-built-up**	**% Non-built-up**
Nairobi	695 km^2	229 km^2	32%
Lagos	3.815 km^2	2.057 km^2	53%
Kano	20.069 km^2	18.358 km^2	91%

In contrast, using FUA reflects the actual urban centres and patterns, based on objective characteristics (i.e., travel time to the Urban Centres, area of the Urban Centres, population, and country GDP per capita). As can be seen in Figure x, the FUA does not follow the extension set by the administrative areas.

In many cities, there are outlying urbanised areas not directly linked to the city centre. A common method is to apply a uniform 1-kilometre buffer. This distance includes nearby non-urban regions and potential future urban expansion, ensuring peripheral urban zones are considered in planning (REF). The building, road and population metrics outlined in the ensuing sections (3.2, 3.3 and 3.4) are therefore summarised within the FUA 100m grid for each pilot city (Lagos, Kano and Nairobi).

ADD EXTRA TEXT ON GENERATING FISHNET 100M GRID OF FUA AND REMOVING ISLANDS.

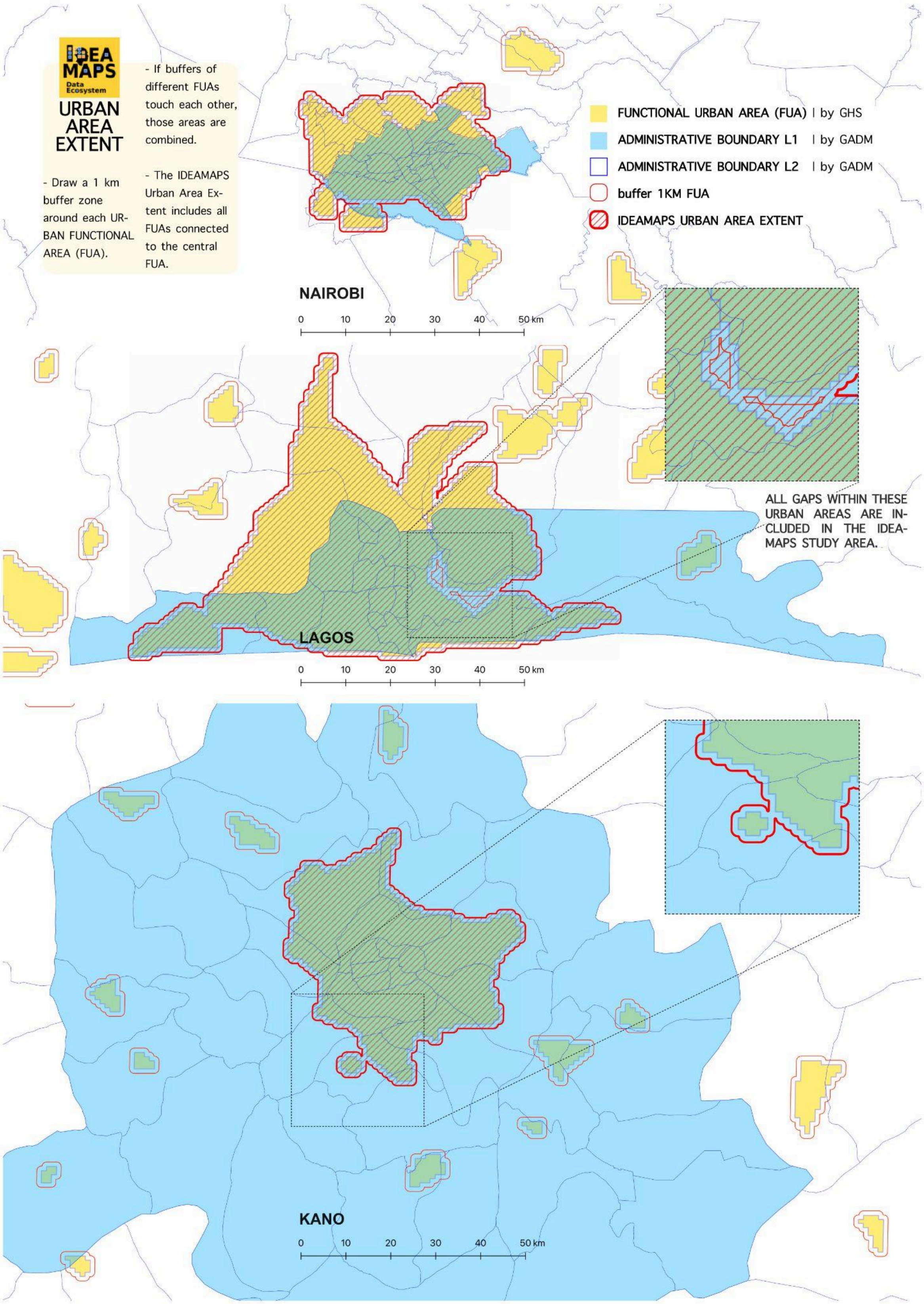

IDEA MAPS Data Ecosystem
URBAN AREA EXTENT
- Draw a 1 km buffer zone around each URBAN FUNCTIONAL AREA (FUA).
- If buffers of different FUAs touch each other, those areas are combined.
- The IDEAMAPS Urban Area Extent includes all FUAs connected to the central FUA.
FUNCTIONAL URBAN AREA (FUA) | by GHS
ADMINISTRATIVE BOUNDARY L1 | by GADM
ADMINISTRATIVE BOUNDARY L2 | by GADM
buffer 1KM FUA
IDEAMAPS URBAN AREA EXTENT
NAIROBI
0 10 20 30 40 50 km
ALL GAPS WITHIN THESE URBAN AREAS ARE INCLUDED IN THE IDEAMAPS STUDY AREA.
LAGOS
0 10 20 30 40 50 km
KANO
0 10 20 30 40 50 km

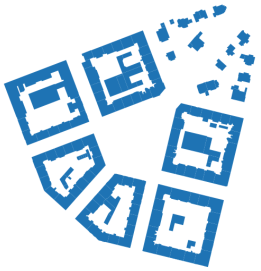
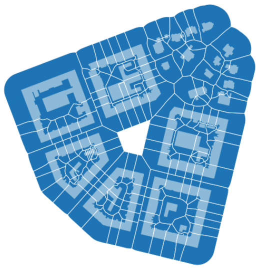
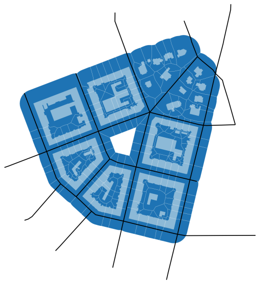
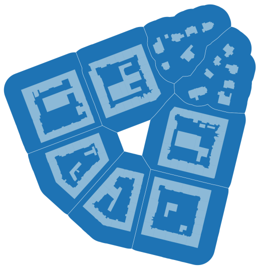
Figure x: Nairobi, Lagos and Kano administrative boundary with the FUA.

3.2 Building morphometrics

Building features have been shown to exhibit the strongest correlation with slum occurrence (Li et al. etc etc) and will therefore be used as the primary explanatory variable for identifying high, medium and low clusters of morphological informality.

1. We will use the open building dataset provided by Google (Sirko et al., 2021) and extract building morphological variables (app 1) within our pilot city functional urban areas (~~INSERT FIGURE BELOW OF FUA~~). Google open buildings v3 was used as it has greater coverage and completeness within urban areas when compared to other openly accessible datasets such as Ecopia, OSM and Microsoft (Chamberlain et al., 2024). The google buildings dataset also had a greater median count of building footprints per grid cell for Nigeria and Kenya, when compared to the other aforementioned datasets (fig x below). The Google buildings dataset also focuses predominantly on classifying buildings within the continent of Africa (and the Global South at large) and is therefore suitable for current and future pilot cities (Sirko et al., 2021).
2. The building and road dataset (OSM highways: outlined in section 2.0) will be used to generate building features, tessellations and blocks and extract 24 different morphology metrics (app 1). The Momepy library in python was used for calculating tessellations and blocks using building and road features (fig.x below). This library has been used to calculate building morphometrics in slums and informal settlements (Wang et al. xxx). The library is part of PySAL (Python Spatial Analysis Library) and is built on top of GeoPandas, other PySAL modules and networkX.

Table x. An example of outputs generated using the Momepy library and calculating building tessellations and tessellation-based blocks.

Building features	Building tessellation	Tessellation and road features	Blocks
			

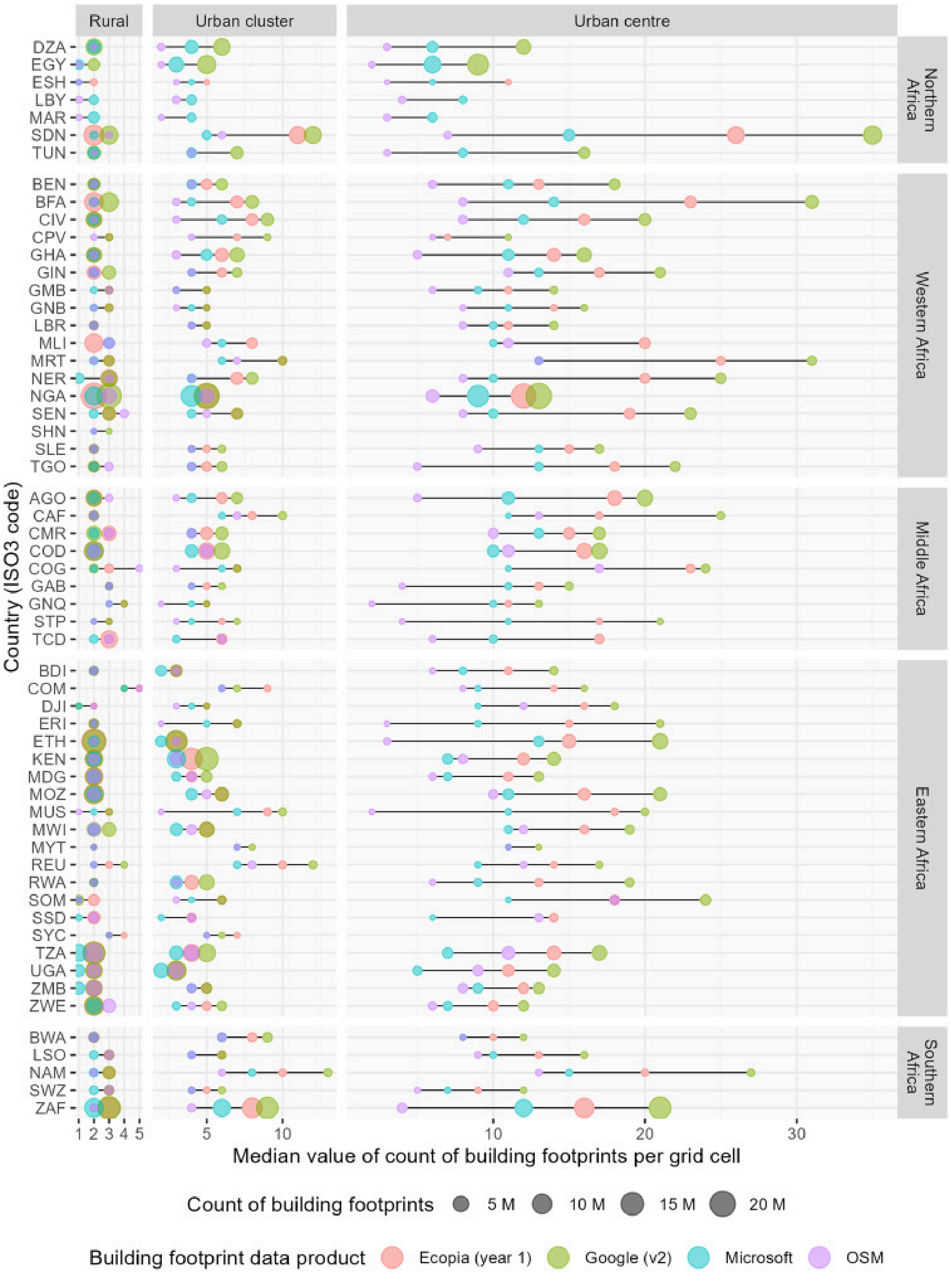


Figure x. Median building count features for OSM, Microsoft, Ecopia and Google buildings datasets summarised for African countries. Taken from Chamberlain et al. (2024).

~~MORE CLEAR ON CALCULATING BLOCK, TESS, BUILDING.~~ ~~MORE CLEAR ON THE AGGREGATION FOR EACH METRIC TOO.~~ ~~CITE JOAOS PAPER ON THE ROAD COMPLETENESS.~~ ~~GOOGLE BUILDINGS WHY WE CHOOSE THEM, HOW GOOD THEY ARE ETC.~~ ~~POPULATION DATA JUSTIFICATION. WHY WE USE IT AND HOW THE METRICS ARE CALCULATED.~~ ~~DOES THE OSM DATA HAVE WIDTH. GIVE ROAD WIDTH BASED ON TYPE IF NOT.~~

3. Using the reference data outlined in section 2.0, an importance analysis of these metrics will be calculated for slum areas using a variety of methods including autocorrelation, correlation analysis, logistic robust regression analysis, multicollinearity, random forest, gradient boosting and decision tree. The results will be used to determine which method is most optimal for reducing variables of building morphometrics within our pilot cities.
    a. MORE INFO/CLARITY ON THE METHOD WE USED AND HOW/WHY.
    b. DO A VIF UNDER REGRESSION FRAMEWORK
    c. DO FEATURE IMPORTANCE UNDER MACHINE LEARNING FRAMEWORK
4. 15 metrics will be used to define two key clusters of building morphologies: ‘small dense structures’ and ‘irregular settlement layout’.

Table x. The metrics used to calculate small dense settlements and irregular settlement layout. Note: b, t and bl refer to building, tessellation and block features respectively.

Irregular Settlement	**Small Dense Structures**
Centroid corners	Area b
Orientation	Elongation b
Alignment between b	Distance to neighbour
Alignment between b and t	Weighted inter building distance
Number of neighbours t	Adjacency b
Orientation	Area t
Number of bl neighbours	Equivalent rectangular index t
	Area ratio t
	Number of neighbours t
	Weighted area of t neighbours
	Number of bl neighbours

5. 15 Metric statistics from SDS and ISL are aggregated to the 100m *100m FUA grid.

Table x. The statistics used to aggregate metrics within each cell of the 100m FUA grid for both small dense settlements and irregular settlement layout based upon Wang et al. (2023).

Statistical summary type	**Irregular Settlement**	**Statistical summary type**	**Small Dense Structures**
sd	Orientation t	sum & med	Area b
	Orientation b	med	Elongation b
med	Centroid corners		Distance to neighbour
	Alignment between b		Weighted inter building distance
	Alignment between b and t		Adjacency b
	Number of neighbours t		Area t
	Number of bl neighbours		Equivalent rectangular index t
			Area ratio t
			Number of neighbours t
			Weighted area of t neighbours
			Number of bl neighbours

~~Explain that we chose 11 metrics SDS and 7 metrics for ISL.~~

6. 15 metrics unsupervised clustering (k-means) will be used to create clusters of small dense structures and irregular settlement layout across the three pilot cities. USE JOHN WANG PAPER TO EXPLAIN WHY HE USED 8,10,12 CLUSTERS. Why we did it etc. collected one class each from the set of clusters.

7. Focal mean analysis will be used to reduce salt-and-pepper noise and ensure high medium and low areas are grouped more consistently (***this is an update from Phase 1, we didn’t use focal mean and we had fragmented areas of dep vs non dep, focal means should make things a bit more uniform INSERT REF. Maybe this doc should have a section at the beginning that outlines the technical developments between P1 and P2 model but in more detail i.e. highlighting the specific methods used and a comparison betw the two?)

This process is outlined in fig x below.

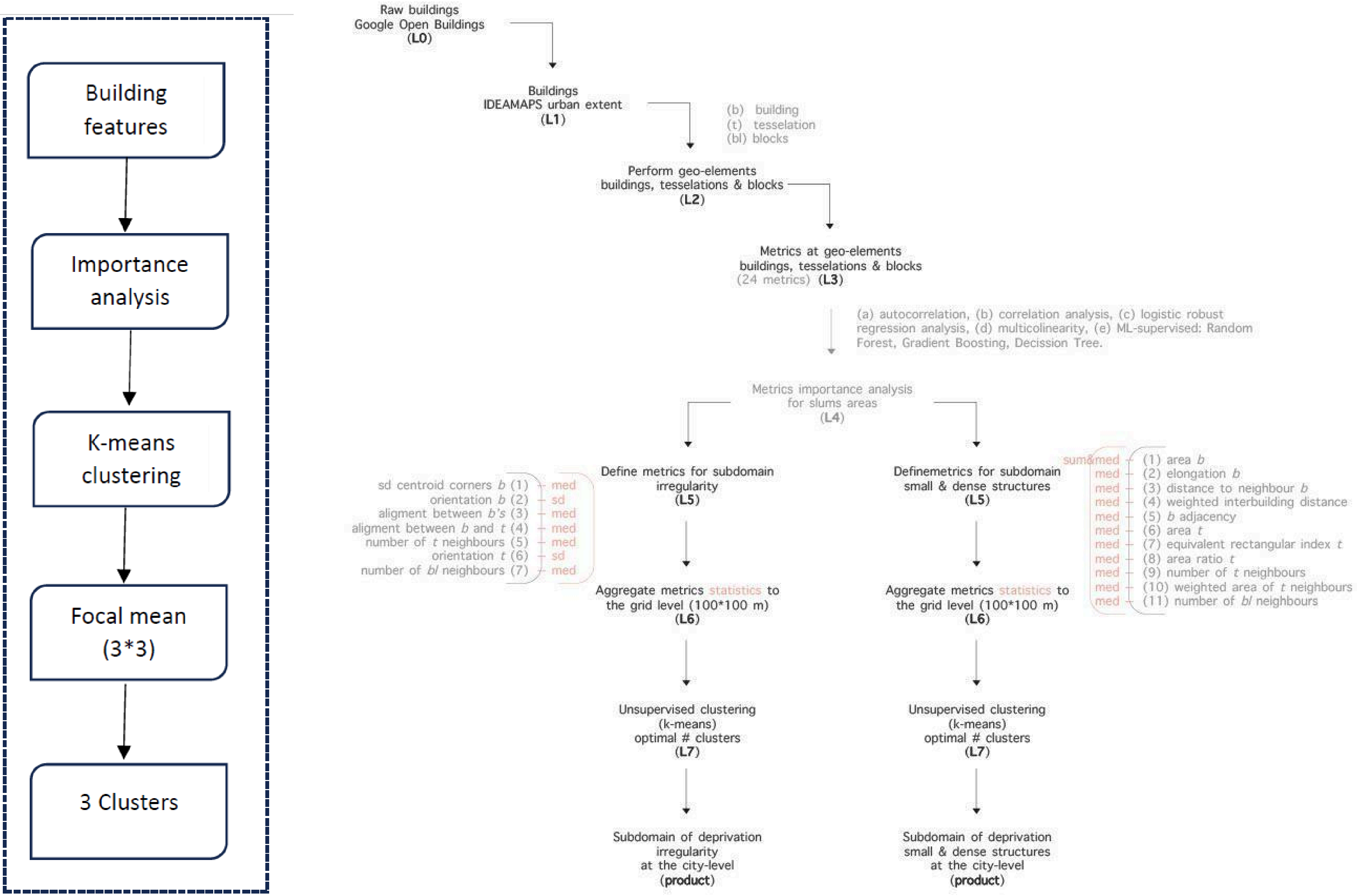


Fig x. The different steps required to generate building metrics and develop products of small dense structures and irregular settlement layout using Google Open Building data.

3.3 Road metrics

Road and accessibility metrics can influence and reflect the level of deprivation experienced in a community (REFs). They have also been used in conjunction with building metrics to prevent built-up formal areas, such as central business districts or downtown areas, from being misclassified as slum areas during modelling processes (REFs). Road and accessibility metrics have therefore been used to support the classification of high, medium and low morphological informality as described below.

The second part was focused only on OpenStreetMap (OSM) since it showed the best results in the previous stage (Jovanovic et al., 2019)

1. Road network data was sourced from OpenStreetMap (OSM) using the following query parameters:

   Key: Highway
   Value: City Name
   Selection Criteria: “highway” IN "highway" IN ('motorway' , 'motorway_link' , 'primary' , 'primary_link' , 'raceway' , 'road' , 'secondary' , 'secondary_link' , 'tertiary' , 'tertiary_link' , 'trunk' , 'trunk_link')

   For Lagos, it was observed that several roads appeared to function as tertiary roads but were categorised as residential. This was likely due to these roads being unpaved and were therefore not included within the final road dataset.

   OSM data was used in comparison to other datasets due to its global availability and high performance in urban areas (Minaei, 2020). Furthermore, OSM has been shown to have a higher content density of road data in large sums when compared to government sourced datasets (Mahabir et al., 2017). However, OSM road data is not suitable for modelling rural areas due to its limited coverage and completeness in these areas.

2. The road dataset was used to calculate road and accessibility metrics including road density, road length, road intersections and distance to roads. MORE CLARITY GRANT TO CALCULATE ROAD WIDTH BASED ON HIGHWAY TYPE I.E. RESIDENTIAL, SECONDARY, TERTIARY ROAD AND USE THIS TO RECALC DENSITY.
3. Metric statistics are aggregated to the 100m *100m grid.

4. Existing reference data sets are then used to define thresholds of road metrics within high, medium and low MI areas i.e. slum sites are used to define road thresholds for high MI areas in each pilot city.
5. Thresholds are then used to generate outputs of high, med and low MI at the city level.

   ~~NB: if thresholds overlap for a particular cell, the cell will use the definitions of the lower MI class.~~

Fig x. TRACEABILITY chain for calculating road/accessibility metrics.

3.4 Population metrics

High population density areas can experience more intense deprivation due to increased pressure on resources and services like housing, healthcare, and education (REFs). Population metrics have been used in conjunction with building metrics and road metrics to help differentiate between formal built up areas and non-formal built up areas (REFs) Within this model, population metrics have therefore been used to support the classification of high, medium and low morphological informality as described below.

1. Population data was sourced from GHS-POP 2023. The dataset contains estimates of the number of people living in 100m Mollweide grid cells. This dataset was used as it has global coverage at 100m resolution and therefore allows for inter-city and inter-country comparison. The GHS population grid is also compatible with the FUA grid (section 3.1) which was used to calculate the urban extents to which summary statistics are aggregated
   Note: The spatial raster dataset depicts the distribution of residential population, expressed as the number of people per cell. Residential population estimates between 1975 and 2020 in 5-year intervals and projections to 2025 and 2030 derived from CIESIN GPWv4.11 were disaggregated from census or administrative units to grid cells, informed by the distribution, volume, and classification of built-up as mapped in the Global Human Settlement Layer (GHSL) global layer per corresponding epoch.
   Note: This dataset is an update of the product released in 2022. Major improvements are the following: use of built-up volume maps (GHS-BUILT-V R2022A); use of more recent and detailed population estimates derived from GPWv4.11 integrating both UN World Population Prospects 2022 country population data and World Urbanisation Prospects 2018 data on Cities; revision of GPWv4.11 population growth rates by convergence to upper administrative level growth rates; systematic improvement of census coastlines; systematic revision of census units declared as unpopulated;

integration of non-residential built-up volume information (GHS-BUILT-V_NRES R2023A); spatial resolution of 100m Mollweide (and 3 arcseconds in WGS84); projections to 2030.

2. Population density is calculated by dividing the number of people by the area of the spatial unit, expressed as people per square metre (people/m²) within the provided grid cells.
3. Due to the overlap of grid cells with the FUA 100m grid cells used in the MI model, average density values are aggregated based upon the proportional coverage of cells as described below:

$$PopDens(Ai) = \sum_{k=1}^{n} (Overlap(A_i, B_k) \times PopDens(Bk))$$

Where

- A_i represent cell i in the pilot city FUA fishnet.
- B_k represent cell k in the GHS-POP 2023 dataset.
- k ranges over all cells in map B that overlap with cell A_i
- $Overlap(A_i, B_k)$ is the fraction of the area of cell A_i that overlaps with cell B_k
- $PopDens(Bk)$ is the population density of cell B_k

4. Existing reference data sets are then used to define thresholds of population density within high, medium and low MI areas i.e. slum sites are used to generate population density thresholds for high MI areas in each pilot city.
5. Thresholds are then used to generate outputs of high, med and low MI at the city level.

Fig x. TRACEABILITY chain for calculating population metrics.

3.5 Combining multiple covariates of morphological informality

In this section, the methods developed to combine the three variables and multiple covariates (outlined in the previous sections) will be described. A rule-based machine learning algorithm will be adopted to develop three categories of morphological informality (high-medium-low).

3.5.1 Python libraries/packages for data processing

The source code will be developed in Python and Jupyter notebook will be used to develop and record the framework. Numpy and pandas (geopandas) will be used to process the data, whilst scikit-learn will be used to deploy the random forest machine learning algorithm. Within this phase (2.1) we have introduced a number of new variables to the modelling process (population, road metrics etc). It is therefore important to understand whether these covariates have a significant contribution to the prediction (morphological informality) within our case study areas. This will be repeated for each pilot city. Scikit-learn offers a package for calculating shap values and will therefore be used in this study (outlined in more detail below).

3.5.2 Reference, training and test data.

Reference datasets (described within section 2.0) will be used to identify high, medium and low reference areas of morphological informality. These areas will be used to develop a set of training and test data for the model. Covariate data within these areas will be randomly sampled at 70% and 30% for training and testing of the model. At present, the reference data developed by combining PAR validation data with model outputs has a different number of samples due to the feedback collected from PAR sessions. I.e. Kano received the highest number of feedback and validation of previous model outputs and will therefore have a larger number of samples used within the training and test datasets when compared to Lagos which received the lowest number of validated cells.

Table x. The number of validated samples for Irregular settlement layout and small dense structures. Note that some validated cells will overlap due to participating providing feedback on the same cell. The table therefore also provides the number of unique samples to accommodate for this.

City extent	Number of validated samples	Number of unique samples	SDS	ISL	Both SDS and ISL
Kano	1015	829	301	714	
Lagos	238	202			
Nairobi	539	355			

If we only have one person providing feedback then we will follow this feedback i.e. agree or disagree.

3.5.3 Hyperparameter tuning

The morphological informality model developed within phase 2.1 has introduced new covariates when compared to the variables modelled within phase 1. It is therefore important to adopt hyperparameter tuning to improve model performance and avoid overfitting of outputs. This model will use RandomisedSearchCV algorithm to determine which combination of hyperparamter values (number of trees in random forest and maximum number of levels in a tree) perform best and will be used in the model going forward.

3.5.4 Training of model

The parameters defined within section 3.4.3 will be used to train the Random Forest model. The covariates (section 3.2, 3.3 and 3.4) and prediction target data will be used to train and analyse the performance of the model in the initial phase. R-squared values will be calculated to determine the fit between the observations and the predicted values. We will also use the remaining test data (3.4.2) to determine model generalisation and overfitting.

3.5.5 SHAP values and covariate reduction

Due to the introduction of new variables within the MI model in Phase 2.1, it is important to determine which variables contribute significantly to the prediction target/outputs. A specific version of SHAP (TreeSHAP) has been developed to be compatible with the Random Forest decision tree model used in this phase of work.

The SHAP values generated for each city domain/model the variables will be used to determine the contribution of each covariate to prediction accuracy. It is therefore an opportunity to reduce the number of variables (if necessary) used in the modelling process and create a more efficient mode of modelling and generating outputs at a large city scale (particularly Lagos where the datasets are large).

Different plots/visualisation techniques (beeswarm plots/partial dependence plots) will be generated to visually inspect outputs and determine whether the relationships between independent and dependent variables being described are plausible and relevant to our modelling process.

3.5.6 Run model using reduced variables

Using the methods outlined in section 3.4.1 to 3.4.5 the morphological informality model will be re-run using a reduced number of covariates/variables and the predictions will be plotted spatially to generate outputs/maps of morphological informality at high, medium and low MI. SHAP values will also be recalculated to compare the impact of reduced variables upon prediction accuracy when compared to the model using the full set of covariates.

Fig below for assigning values to high med low and using weighted sum to aggregated values (50% for building metrics, 25% population and 25% road metrics).

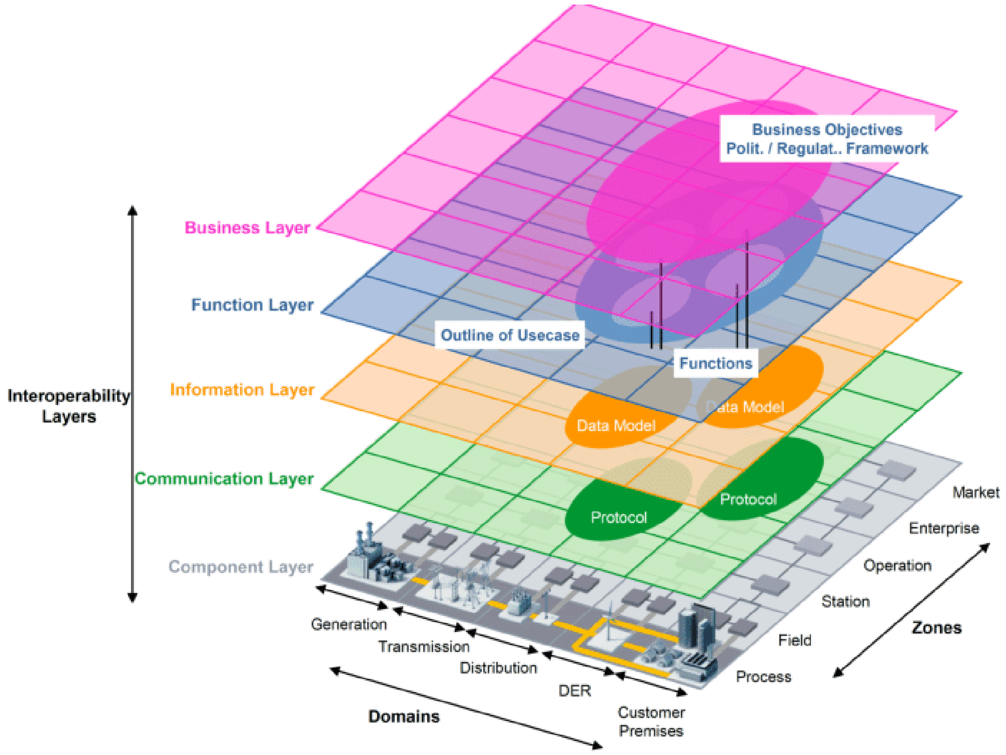
Need to create a link between tech spec and the laypersons doc.

Google buildings data	Building metrics	Summarised at grid level	Clustering	SDS	(binary)	Rule based ML model (training validation process)	Final high medium low output of MI
				ISL			
		OSM road data	Road metrics	Summarised at grid level	100m grid (continuous)		
			GHS population data	Summarised at grid level	100m grid (continuous)		

Bringing this back to the PAR sessions

How do we weight the variables?

Which ones out of these 27 do you think is the most/least important


Business Objectives
Polit. / Regulat.. Framework
Business Layer
Function Layer
Outline of Usecase
Functions
Interoperability Layers
Information Layer
Data Model
Data Model
Communication Layer
Protocol
Protocol
Component Layer
Market
Enterprise
Operation
Station
Field
Process
Zones
Generation
Transmission
Distribution
DER
Customer Premises
Domains

3.0 Analysis Plan (option B)

The analysis will be performed in 5 or 6 steps.

First, we will generate a global dataset of sub-city segments within GHS-UCDB (2019) urban centres, and calculate a set of road, building, and population metrics. In the subset of cities where we have complete citywide “slum” maps, segments will be classified as follows:

- Highly informal = Segment area is 90%+ covered by a slum
- Moderately informal = Segment area is 5-90% covered by slum (discuss if this makes sense during WP2/3 mtg on 21/5/24…), and
- Formal = segment is <5% area covered by a slum

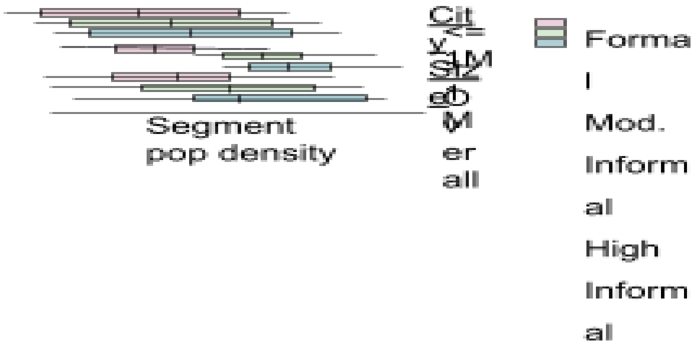
Second, we will perform bivariable analysis in ALL SEGMENTS to understand correlations among road, building, population metrics (multicollinearity). This will help identify any metrics that contribute “duplicate” information to the analysis and might be dropped to simplify the analysis.

- Sensitivity analysis: Check bivariate correlations by Region, city size, and city growth rate. If differences are observed, why?

OPTIONAL. **Third**, we will perform PCA analysis on the remaining road, building, and population metrics in ALL SEGMENTS and interpret 1+ principal components. Two potential insights might be gained from this exercise. First, discussing interpretations of PCs might give insight about expressions of urban form at approximately the neighbourhood scale. Second, factor weightings indicate variable importance/contribution to each PC.

- OPTIONAL. Sensitivity analysis: Perform PCA analysis by region, city size, and city growth rate. Discuss differences in urban forms (PCs) by city type.

Fourth, in cities where complete citywide “slum” maps are available, we will plot retained metrics by slum category and city characteristics and calculate bivariable test statistics (eg t-test / X2). For example:



Fifth, in cities where complete citywide “slum” maps are available, we will perform multivariable regression analysis to understand the strength and direction of associations between road/building/pop metrics and classes of “morphological informality”. Note, other modelling approaches (eg random forest) might be considered for this step.

- Sensitivity analysis: Run models by Region, City size, and City growth rate. Does direction or strength of association between metrics & “slumness” differ city type?

Sixth, Decide on 2-5 sets of potentially viable segment metrics + thresholds (by city type as appropriate). Use these parameter sets to classify city segments by degree of slumness and (a) evaluate which most consistent with KYC “slum” settlements, and (b) grid the output and test with IDEAMAPS colleagues in select cities

Note: We need to think carefully about if/how “slumness” presents itself at the 100m resolution (approximately uniform areas, different total pop) versus city segments (approximately uniform total pop, different area). City segments are likely closer to the conceptualisations of “slums” on the ground, and are likely a more sensible resolution and unit of analysis than grid cells. Relationships that exist between segment metrics and “slum” classes at the city segment scale might not hold (or be calculable) at a 100m grid scale to be discussed at WP2/3 mtg 21/5/24. Alternative method (Appendix 3) will be discussed during WP2/3 mtg 21/5/24.

4.0 Bibliography

Birkmann, J., Welle, T., Solecki, W., Lwasa, S., Garschagen, M., 2016. Boost resilience of small and mid-sized cities. Nature. https://doi.org/10.1038/537605a

Chamberlain, H.R., Darin, E., Adewole, W.A., Jochem, W.C., Lazar, A.N., Tatem, A.J., 2024. Building footprint data for countries in Africa: To what extent are existing data products comparable? Computers, Environment and Urban Systems. https://doi.org/10.1016/j.compenvurbsys.2024.102104

Dahiya, Bharat. (2014). Southeast Asia and sustainable urbanization (Strategic Review: The Indonesian Journal of Leadership, Policy and World Affairs, ISSN: 2477-1813, Jakarta). 4. 125-134.

Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., Zipf, A., 2023. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nat Commun. https://doi.org/10.1038/s41467-023-39698-6

Jovanovic, S., Jovanovic, D., Bratic, G., Brovelli, M.A., 2019. Analysis of free road data in Tanzania, Uganda and Kenya using free and open source software. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. https://doi.org/10.5194/isprs-archives-xlii-2-w13-1567-2019

Kraff, N.J., Taubenbock, H., Wurm, M., 2019. How dynamic are slums? EO-based assessment of Kibera's morphologic transformation. 2019 Joint Urban Remote Sensing Event (JURSE). https://doi.org/10.1109/jurse.2019.8808978

Kohli, D., Sliuzas, R., Kerle, N., Stein, A., 2012. An ontology of slums for image-based classification. Computers, Environment and Urban Systems. https://doi.org/10.1016/j.compenvurbsys.2011.11.001

Li, C., Yu, L., Oloo, F., Chimimba, E.G., Kambombe, O., Asamoah, M., Opoku, P.D., Ogweno, V.W., Fawcett, D., Hong, J., Deng, X., Gong, P., Wright, J., 2023. Slum and urban deprivation in compacted and peri-urban neighborhoods in sub-Saharan Africa. Sustainable Cities and Society. https://doi.org/10.1016/j.scs.2023.104863

Mahabir, R., Stefanidis, A., Croitoru, A., Crooks, A., Agouris, P., 2017. Authoritative and Volunteered Geographical Information in a Developing Country: A Comparative Case Study of Road Datasets in Nairobi, Kenya. IJGI. https://doi.org/10.3390/ijgi6010024

Minaei, M., 2020. Evolution, density and completeness of OpenStreetMap road networks in developing countries: The case of Iran. Applied Geography. https://doi.org/10.1016/j.apgeog.2020.102246

Satterthwaite, D., 2017. The impact of urban development on risk in sub-Saharan Africa's cities with a focus on small and intermediate urban centres. International Journal of Disaster Risk Reduction. https://doi.org/10.1016/j.ijdrr.2017.09.025

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y.S.E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., Quinn, J., 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery. https://doi.org/10.48550/ARXIV.2107.12283

Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. Applied Geography. https://doi.org/10.1016/j.apgeog.2018.02.002

Wang, J., Fleischmann, M., Venerandi, A., Romice, O., Kuffer, M., Porta, S., 2023. EO + Morphometrics: Understanding cities through urban morphology at large scale. Landscape and Urban Planning. https://doi.org/10.1016/j.landurbplan.2023.104691

5.0 Appendices

Appendix 1

The proposed building metrics used to analyse Google Buildings data. These have been used to calculate metrics for SDS and ISL in the previous phase of IDEAMAPS. Metrics were calculated based upon research by Wang et al. (2023).

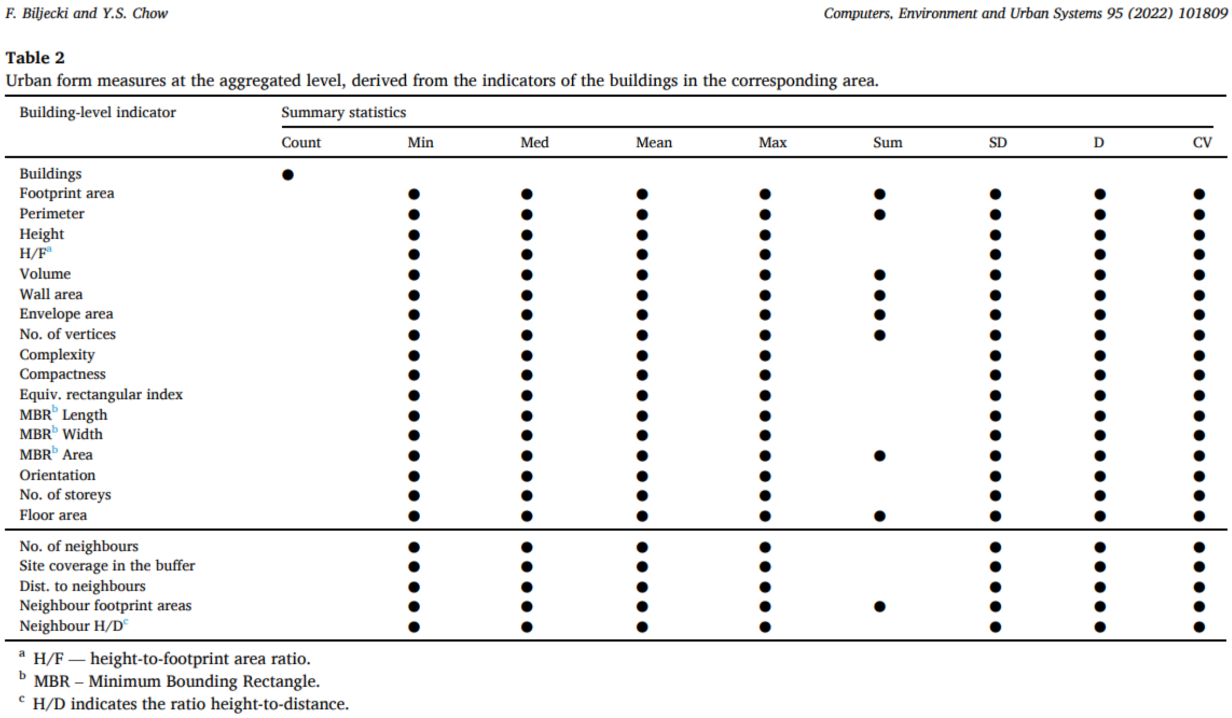
Parameter	**Description**
Area	Calculates the area of each object in a given GeoDataFrame. It can be used for any suitable element (building footprint, plot, tessellation, block). It is a simple wrapper for GeoPandas (.area) for the consistency of momepy.
Perimeter	Calculates perimeter of each object in a given GeoDataFrame. It can be used for any suitable element (building footprint, plot, tessellation, block).

Compactness	Calculates the compactness index of each object in a given GeoDataFrame.
Corners	Calculates the number of corners of each object in a given GeoDataFrame. Uses only external shape (shapely.geometry.exterior), courtyards are not included.
Squareness	Calculates the squareness of each object in a given GeoDataFrame. Uses only external shape (shapely.geometry.exterior), courtyards are not included.
EquivalentRectangularIndex	This index aims to measure deviation of a polygon from an equivalent rectangle,improving a drawback of rectangularity. ssbERI largely overcomes this problemby scaling the MABR until its area equals to the polygon’s area.
Elongation	Calculates the elongation of each object seen as elongation of its minimum bounding rectangle. * (fom Basaraner 2017) Rectangularity is too sensitive to the protrusions along the boundary of a BF polygon. This can cause a significant increase in the size of the polygon’s minimum area bounding rectangle, and thus can produce a misleading value for the shape.
CentroidCorners_mean	Calculates the mean distance centroid corners (ssbCCM)
CentroidCorners	Calculates the standard deviation distance centroid corners (ssbCCD).
Orientation	Calculate the orientation of object. The deviation of orientation from cardinal directions are captured. Here ‘orientation’ is defined as an orientation of the longest axis of bounding rectangle in range 0 - 45. The orientation of LineStrings is

	represented by the orientation of the line connecting the first and the last point of the segment.
Aligment	Calculate the mean deviation of solar orientation of objects on adjacent cells from an object.
Cell Aligment	Calculate the difference between cell orientation and the orientation of object.
NeighborDistance	Calculate the mean distance to adjacent buildings (based on spatial_weights).
MeanInterbuildingDistance	Calculate the mean interbuilding distance. Interbuilding distances are calculated between buildings on adjacent cells based on spatial_weights, while the extent is defined as order of contiguity.
BuildingAdjacency	Calculate the level of building adjacency. Building adjacency reflects how much buildings tend to join together into larger structures. It is calculated as a ratio of joined built-up structures and buildings within the extent defined in spatial weights higher.
LongestAxisLength	Calculates the length of the longest axis of object. Axis is defined as a diameter of minimal circumscribed circle around the convex hull. It does not have to be fully inside an object.
Area	Calculates the area of each object in a given GeoDataFrame. It can be used for any suitable element (building footprint, plot, tessellation, block). It is a simple wrapper for GeoPandas (.area) for the consistency of momepy.
CircularCompactness	Calculates the compactness index of each object in a given GeoDataFrame.

EquivalentRectangularIndex	Calculates the equivalent rectangular index of each object in a given GeoDataFrame.
AreaRatio	Calculate covered area ratio or floor area ratio of objects.
Neighbours	Calculate the number of neighbours captured by spatial_weights.
CoveredArea	Calculates the area covered by neighbours, which is total area covered by neighbours defined in spatial_weights and the element itself.
Orientation	Calculate the orientation of object. The deviation of orientation from cardinal directions are captured. Here ‘orientation’ is defined as an orientation of the longest axis of bounding rectangle in range 0 - 45. The orientation of LineStrings is represented by the orientation of the line connecting the first and the last point of the segment.
BlocksCount	Calculates the weighted number of blocks. The number of blocks within neighbours defined in spatial_weights divided by the area covered by the neighbours.

There are additional metrics here in this ref (Biljecki and Chow, 2022). Should we merge some of these indicators and conduct another importance analysis? To be discussed during WP2/3 mtg 21/5/24.



F. Biljecki and Y.S. Chow Computers, Environment and Urban Systems 95 (2022) 101809

Table 2
Urban form measures at the aggregated level, derived from the indicators of the buildings in the corresponding area.

Building-level indicator	Summary statistics								
	Count	Min	Med	Mean	Max	Sum	SD	D	CV
Buildings	●								
Footprint area		●	●	●	●	●	●	●	●
Perimeter		●	●	●	●	●	●	●	●
Height		●	●	●	●		●	●	●
H/F[a]		●	●	●	●		●	●	●
Volume		●	●	●	●	●	●	●	●
Wall area		●	●	●	●	●	●	●	●
Envelope area		●	●	●	●	●	●	●	●
No. of vertices		●	●	●	●	●	●	●	●
Complexity		●	●	●	●		●	●	●
Compactness		●	●	●	●		●	●	●
Equiv. rectangular index		●	●	●	●		●	●	●
MBR[b] Length		●	●	●	●		●	●	●
MBR[b] Width		●	●	●	●		●	●	●
MBR[b] Area		●	●	●	●	●	●	●	●
Orientation		●	●	●	●		●	●	●
No. of storeys		●	●	●	●		●	●	●
Floor area		●	●	●	●	●	●	●	●
No. of neighbours		●	●	●	●		●	●	●
Site coverage in the buffer		●	●	●	●		●	●	●
Dist. to neighbours		●	●	●	●		●	●	●
Neighbour footprint areas		●	●	●	●	●	●	●	●
Neighbour H/D[c]		●	●	●	●		●	●	●

[a] H/F — height-to-footprint area ratio.
[b] MBR – Minimum Bounding Rectangle.
[c] H/D indicates the ratio height-to-distance.

Appendix 2.

Building derived metrics that are easily accessible in raster format from WorldPop, Meta and Million Neighbourhoods

These may be good to validate/cross check our outputs that we calculate via GB data and compare? To be discussed during WP2/3 mtg 21/5/24.

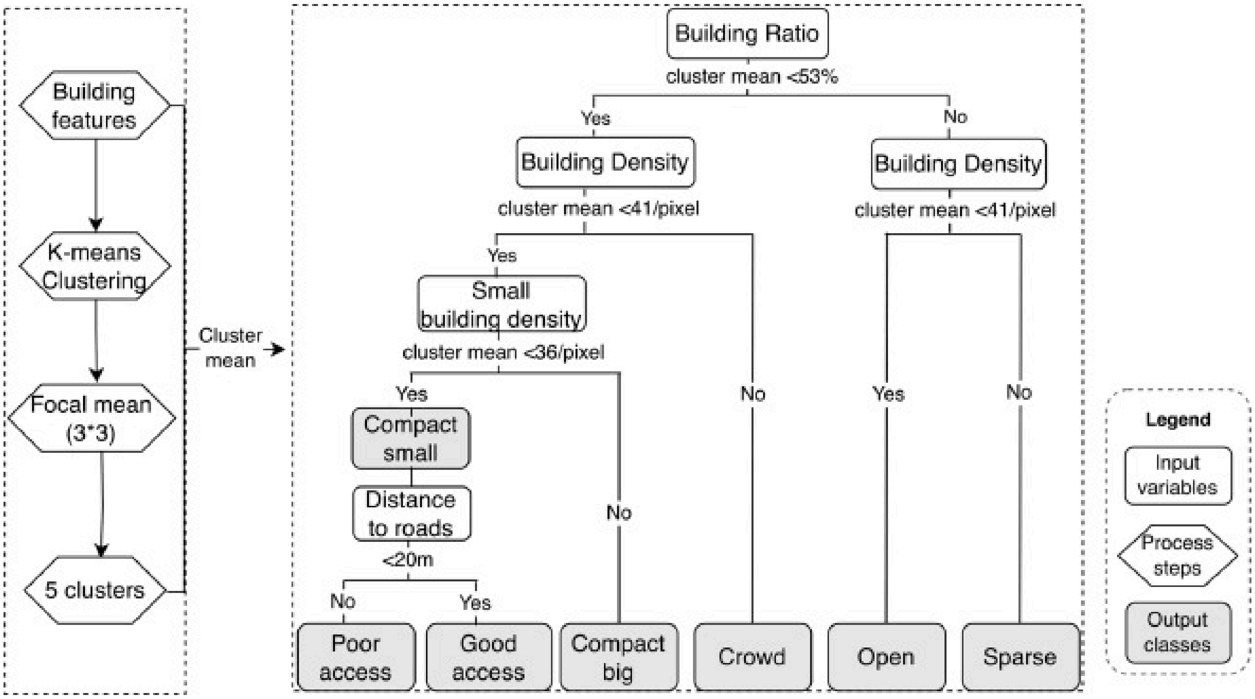
variable	description	source
any_blg	Any building is present	Facebook / Meta
XXX_buildings_v2_0_count.tif	This raster contains counts of buildings that fall within a grid cell. Each buildings was counted in the grid cell that contained the centroid of its building footprint.	WorldPop WOPR
XXX_buildings_v2_0_density.tif	This raster contains a measure of the number of buildings per grid cell area in square kilometres (km2), i.e. building count divided by the total number of square kilometres in the grid cell. If needed, the grid cell area can be retrieved by dividing the count raster by the density raster.	WorldPop WOPR
XXX_buildings_v2	This raster contains an urban / rural classification for each grid	WorldPop

_0_urban.tif	cell based on building patterns in their surroundings. A value of 1 indicates an urban grid cell and a value of zero indicates a rural grid cell. NAs indicate grid cells without any building-footprint centroid. Note: Details on the settlement classification are provided in the methods section.	WOPR
XXX_buildings_v2_0_total_area.tif	This raster contains a grid cell level sum of the building areas for all buildings whose centroid falls inside a grid cell. Total area is given in m2. Note: Total building area could exceed the area of a grid cell if the centroid of a large building falls within the grid cell.	WorldPop WOPR
XXX_buildings_v2_0_mean_area.tif	This raster contains a grid cell level mean of the building areas for all buildings whose centroid falls inside a grid cell. Mean area is given in m2.	WorldPop WOPR
XXX_buildings_v2_0_cv_area.tif	This raster contains a grid cell level coefficient of variation of building areas for all buildings whose centroid falls inside a grid cell. Coefficient of variation is the standard deviation divided by the mean. Note: This measure informs about the heterogeneity in building areas within a grid cell.	WorldPop WOPR
XXX_buildings_v2_0_total_length.tif	This raster contains a grid cell level sum of the building lengths for all buildings whose centroid falls inside a grid cell. Here, length refers to a building's perimeter. Total length per grid cell is therefore a sum of the perimeters of all the buildings whose centroid falls inside a grid cell. Total length is given in metres.	WorldPop WOPR
XXX_buildings_v2_0_mean_length.tif	This raster contains a grid cell level mean of the building lengths for all buildings whose centroid falls inside a grid cell. Here, length refers to a building's perimeter. Mean length per grid cell is therefore the mean perimeter length of all the buildings whose centroid falls inside a grid cell. Mean length is given in metres.	WorldPop WOPR
XXX_buildings_v2_0_cv_length.tif	This raster contains a grid cell level coefficient of variation of building lengths for all buildings whose centroid falls inside a grid cell. Coefficient of variation is the standard deviation divided by the mean. Here, length refers to a building's perimeter. Note: This measure informs about the heterogeneity in building lengths within a grid cell.	WorldPop WOPR
bldg_area_count	Count of buildings with X-X area (categorized)	Million Neighborhoods
bldg_area_m2	Cumulative area of buildings X-X area (categorized)	Million Neighborhoods
worldpop_population_un_per_building_area_m2	Population per building area in meters square (WorldPop 2020, UN adjusted). Based on dividing: worldpop_population_un / building_area_m2	Million Neighborhoods

worldpop_population_un_per_building	Population per building (WorldPop 2020, UN adjusted). Based on dividing: worldpop_population_un / building_count	Million Neighborhoods
building_to_block_area_ratio	Building area to block area ratio. Based on dividing: building_area_m2 / block_area_m2	Million Neighborhoods
parcel_count	Parcel count	Million Neighborhoods
k_complexity	Block complexity measures the number of building parcels between the least accessible structure and the street network	Million Neighborhoods

Appendix 3

Recent study by Li et al. (2023) that combines multiple input feature classes (buildings, roads, access to services etc) and datasets to create 5 different clusters of urban morphologies. The study combines k-means clustering, focal mean statistics and statistical regression methods within a decision tree process to generate distinct classes of morphology. This method is an additional option to the analysis plan in section 3.0 and will be discussed during WP2/3 mtg 21/5/24.


Building features
K-means Clustering
Focal mean (3*3)
5 clusters
Cluster mean
Building Ratio
cluster mean <53%
Yes
No
Building Density
cluster mean <41/pixel
Building Density
cluster mean <41/pixel
Yes
Small building density
cluster mean <36/pixel
Yes
Compact small
Distance to roads
<20m
No
Yes
No
No
Yes
No
No
Poor access
Good access
Compact big
Crowd
Open
Sparse
Legend
Input variables
Process steps
Output classes