Supplemental Material

Classifying settlement types from multi-scale spatial patterns of building footprints

Warren C. Jochem1*
Douglas R. Leasure1
Oliver Pannell1
Heather R. Chamberlain1
Patricia Jones1
Andrew J. Tatem1

* Corresponding Author

Affiliation for all authors

1. WorldPop, School of Geography and Environmental Science, University of Southampton, UK

Email addresses
w.c.jochem@soton.ac.uk

Acknowledgements

This work was conducted as part of the Geo-Referenced Infrastructure and Demographic Data for Development project (GRID3; OPP1182408) funded by the Bill & Melinda Gates Foundation and the United Kingdom’s Department for International Development. The authors acknowledge the use of the IRIDIS Higher Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Contents

1. Spatial Metrics Calculations
2. Principal Components Selection
3. Examples of Spatial Metrics
4. Mixture Model Selection Process
5. Predicted Settlement Types
6. References
1. Spatial Metrics Calculations

The local patterns of the building footprints are quantified within circular areas in a moving window operation. The calculation window is moved to the centre coordinates for each cell of a regular grid with a 3 arc-second (nominally 100 m) spatial resolution. This resolution was chosen to match other spatial demographic and environmental datasets (Lloyd et al., 2019) as our intention in future work is to integrate the settlement classifications with other demographic datasets. Each grid cell thus becomes a processing location to calculate the seven different spatial metrics. The size of the circular window is also set to 10 different radii (from 50 m up to 500 m in 50 m increments), which we refer to as the “spatial scale” of the analysis. The calculations are repeated at all locations for each radius and the results are stored in the cell processing location of the 100 m grid. At larger spatial scales, the circular processing windows do overlap. The results of the calculations are stacks of rasters (spatially-referenced gridded datasets) with 70 values (7 metrics x 10 spatial scales) for each grid cell location.

To improve computational efficiency, the calculations are performed in parallel by tiling the 100 m grid into blocks. The blocks are first buffered to create overlaps along adjacent edges, and these extended areas are used to extract geographic subsets of the building features that are processed in parallel by multicore processors. In the final step, the tiles of results are reassembled into a 100 m grids spanning the study regions. The overlaps (equal in size to the maximum processing radius) around each tile are necessary to prevent edge effects between blocks from the moving window analyses for the spatial metrics. No adjustments, however, are made at the boundaries of the study regions. The tiling and moving window calculations are summarised graphically below.

|   |   | A set of large “tiles” are defined to cover the full extent of the study region. In this example, the grey circle is the study region for which there are building footprints. These regular tiles can be several kilometres on a side and are used to subset the building shapes and divide the workload across multiple computer processors to perform the calculations in parallel. The tiles are reassembled to create the final, complete grids after the calculations. |
|---|---|---|
| 1 |   | The extent of one of the processing tiles is expanded by the maximum radius to create overlapping blocks. The overlap prevents edge effects during the calculations by extracting building features beyond the tile’s extent. |
Each tile (one example tile is shown at left) is defined on a 100 m spatial resolution grid (in this schematic 7 x 7 cells), the cells of which define the set of processing locations for the moving window calculations. The solid square highlights four such cells which will be shown in steps 4 and 5.

Each of the locations in the 100 m grid is the centre of a circular processing window for the calculations (two are shown). The processing steps are repeated as the radius of the window is varied from 50 m up to 500 m in radius (50 m increments). This example shows only 3 scales, but the larger radii will overlap multiple cell locations.

The fragmentation metrics are calculated on building footprints which have been rasterised to a 1 m spatial resolution (shown in this example as black polygons within one circular window). The results of the calculations for each radius are stored for the grid cell location in the 100 m grid. Each metric and each radius are stored in their own grid, creating a multilayer stack of 100 m resolution grids.
2. Principal Components Selection

The results of the spatial metric calculations are spatial gridded data layers (also known as rasters). There are 70 layers (7 metrics x 10 spatial scales). We highlighted the patterns of variation and reduced the number of data layers by using principal components analysis (PCA) prior to the Gaussian mixture modelling. The data were first mean centred and scaled by their standard deviation within each metric. Next, the minimum number of components to explain 90% of the total variation in each fragmentation metric were retained. The processing was completed separately for each study area and the number of retained components are listed in Table S1.

Table S1: Number of components selected per fragmentation metric in each study area.

| Fragmentation Metric       | Number of Principal Components Selected |
|----------------------------|-----------------------------------------|
|                            | Kaduna | Kinshasa | Maputo |
| Mean Patch Area            | 3      | 4        | 5      |
| Patch Density              | 2      | 2        | 2      |
| Patch Size CV              | 2      | 2        | 1      |
| Mean Fractal Dimension Index | 4      | 4        | 4      |
| Landscape Shape Index      | 2      | 1        | 1      |
| Mean Shape Index           | 3      | 3        | 3      |
| Patch Cohesion Index       | 2      | 2        | 3      |
| **Total**                  | **18** | **18**   | **19** |
The principal component analysis transforms the raw feature data using a set of loadings, the values of which are shown in Figure S1, plotted across the scale of the processing windows. These plots show the patterns for the retained principal components for the three study areas. As noted above, the number of components was selected specific for each study area to explain 90% of the variation in each metric. The three areas show remarkable similarities across many metrics, with the exception of the average fractal index and the average patch area.

Figure S1: Principal component loadings for each fragmentation metric across spatial scales of the moving window.
3. Examples of Spatial Metrics

In this section we provide examples of areas with high (and low) values of the seven spatial metrics. The aim of these examples is to provide greater clarity on the spatial metrics, connecting the calculated values with patterns observed in the study regions. The example areas were selected manually from within the upper (and lower) quartiles of the 250 m scale layer in the Kinshasa study area. In the images below the Ecopia building footprints (© 2020 Maxar Technologies, Ecopia.AI) are overlaid on ESRI basemap imagery.

| Metric          | High                  | Low                   |
|-----------------|-----------------------|-----------------------|
| Mean Patch Area | ![High Mean Patch Area Image](image1) | ![Low Mean Patch Area Image](image2) |
| Patch Density   | ![High Patch Density Image](image3) | ![Low Patch Density Image](image4) |
The examples of the spatial metrics shown above highlight the potential overlap or correlation of some information from the different metrics; however, the classification step identifies the important differences found in the combination of metrics. For clarity we provide a simple, two-metric comparison. The two areas shown in the images below both have low patch densities based on the number of patches per area (in the lowest quartile), but while the scene on the left also has a low average patch area and clearly comes from a rural area, the scene on the right has a high average patch area and is a set of warehouses and hangars near an airport. The combination of different information helps to differentiate and classify areas.

| Low patch density – Low patch area | Low patch density – High patch area |
|------------------------------------|------------------------------------|
| ![Image of low patch density – low patch area](image1.png) | ![Image of low patch density – high patch area](image2.png) |

Source: Esri, DigitalGlobe, GeEye, Earthstar Geographics, CNES/Astrium DS, iView, ULAIR, AEX, Geomapping, Arengi, I2I, IGF, swisstopo, and the GIS
4. Mixture Model Selection Process

The classification procedure used a Gaussian finite mixture model implemented in the R package mclust (Fraley et al., 2016; Scrucca et al., 2016). Gaussian mixture models (GMM) are probabilistic models for describing data which arise from a mixture of unimodal Gaussian distributions. The number and size of the Gaussian distributions are treated as unknowns, and the results are an unsupervised clustering. The Gaussian distributions are assumed to be multivariate, and therefore, the clusters can become ellipsoidal in shape with a covariance matrix (instead of a variance for a univariate distribution) describing a volume, shape, and orientation for each cluster. This design gives GMMs flexibility to find irregularly shaped or overlapping clusters. GMMs can be similar to the k-means clustering algorithm; however, k-means does not use covariance information and thus represents only circular shaped clusters (Hastie et al., 2009).

The GMMs in mclust are fit using the expectation-maximisation (EM) algorithm to estimate the likelihood of observing the data given a set of parameters (Fraley et al., 2016). Because GMMs are a model-based solution, they also provide a probability of cluster membership and can be considered a “soft” clustering technique (Hastie et al., 2009).

In order to select the best performing clustering model, Gaussian mixture models with between 2 and 18 components were fit using 14 different covariance structures in MCLUST version 5 software (Fraley et al., 2016). A Bayesian information criterion (BIC) score is used to compare the models, and the top 6 BIC values are taken as potential candidate models. Because GMMs can be computational demanding, we implemented a data sampling and model comparison approach described by Wehrens et al. (2004). The steps are outlined below:

1. Sample 2000 locations in the 100 m grid as a training set using a balanced spatial design which ensures that sample locations are spatially distributed and representative of the range of feature metrics (Grafström and Tillé, 2013)
2. Initialise the model using hierarchical clustering to partition the data
3. Run up to 100 EM steps using 14 different covariance structures and between 2 and 18 mixture components
4. Model selection to identify the top 6 models (number of components and covariance structure) by BIC
5. Repeat up to 100 EM steps on the top 6 models using the full dataset for the study region
6. Select the single best performing model form that maximises the log-likelihood

Figure S2 shows the performance results graphically for each study area. The sharp declines in BIC values visible in the plots can occur when small sample sizes within components cause unstable model results. The top candidate models, maximising the BIC, are summarised in Table S2. Across all three study areas the fully varying (“VVV”) model was preferred with between 6 and 13 components. The final model was selected by maximising the log-likelihood when applying the top candidate models to the full datasets for each study area. The final model used for predicting the settlement types is indicated with an asterisk (“*”) in the leftmost column of Table S2.
Table S2: Results of the mixture model selection steps. Six candidate models with the best Bayesian Information Criterion (BIC) were selected by fitting models with between 2 and 18 components and 14 covariance structures on a sample of observations from each study area. The best performing model (indicated by ‘*’) was selected by maximising the log-likelihood when fitting the candidate models to the full datasets.

**A. Kaduna**

| Model | Structure (Components) | BIC     | Log-likelihood |
|-------|------------------------|---------|----------------|
| 1     | ‘VVV’ (9)              | -33813.3| -1597916       |
| 2     | ‘VVV’ (7)              | -33862.1| -1788340       |
| 3     | ‘VVV’ (8)              | -34123.4| -1696295       |
| 4     | ‘VVV’ (10)             | -34224.4| -1485265       |
| 5 *   | ‘VVV’ (11)             | -34488.1| -1421494       |
| 6     | ‘VVV’ (6)              | -35121.0| -1944037       |

**B. Kinshasa**

| Model | Structure (Components) | BIC     | Log-likelihood |
|-------|------------------------|---------|----------------|
| 1     | ‘VVV’ (9)              | -6970.5 | 52827.0        |
| 2     | ‘VVV’ (10)             | -7390.2 | 70642.8        |
| 3     | ‘VVV’ (8)              | -7582.7 | 29992.1        |
| 4     | ‘VVV’ (11)             | -8201.1 | 87168.7        |
| 5     | ‘VVV’ (7)              | -8350.5 | -3364.7        |
| 6 *   | ‘VVV’ (12)             | -8934.1 | 107677.2       |

**C. Maputo**

| Model | Structure (Components) | BIC     | Log-likelihood |
|-------|------------------------|---------|----------------|
| 1     | ‘VVV’ (8)              | 291.6   | 570175.1       |
| 2     | ‘VVV’ (11)             | -903.3  | 733970.7       |
| 3 *   | ‘VVV’ (13)             | -953.9  | 838953.4       |
| 4     | ‘VVV’ (9)              | -972.6  | 598128.4       |
| 5     | ‘VVV’ (12)             | -1170.6 | 782569.2       |
| 6     | ‘VVV’ (7)              | -1245.7 | 476834.3       |
Figure S2: Results of Gaussian mixture model selection steps using Bayesian Information Criterion (BIC).
5. Predicted Settlement Types

The main manuscript presents results from Kinshasa, Democratic Republic of the Congo. Below we present additional results from Kaduna, Nigeria and Maputo, Mozambique. As discussed in the methods section, these study regions were analyses in the same manner, but separately. The cluster numbers are not necessarily equal, but we have relabelled all study areas in order of descending patch density at 100 m scale in order to facilitate some relative comparisons. Therefore, predicted settlement type 1 mapped across all three areas refers to likely similar conditions.

The clustering results for the entire state of Kaduna are shown in Figure S3 for an 11 cluster solution and in Figure S4 for a 4 cluster solution which merged the larger number of clusters in order to minimise entropy. In both maps, the predicted settlement types clearly differentiate between densely settled urban areas and sparse, rural areas. However, after merging to the 4 cluster solution, much of the variation within the urban areas is no longer visible. The same urban areas (see inset maps) are now differentiated into just two classes. A comparison with imagery showed that these classes largely conform to large versus small structures.

The clustering results for the area around Maputo, Mozambique are shown in Figures S5 and S6 for the full set of cluster types and the merged set, respectively. In Maputo, the urban core area is clearly distinguished from rural areas (Figure S5). Within this area there are several different predicted settlement types. After merging the clusters to the reduced solution (Figure S6), much of the city is predicted into settlement type 2, though types 3 and 4 remain differentiated within the core area.
Figure S3: Predicted settlement types for Kaduna, Nigeria using an 11 cluster solution.
Figure S4: Predicted settlement types for Kaduna, Nigeria using a 4 cluster solution after merging based on entropy.
Figure S5: Predicted settlement types for Maputo, Mozambique using a 13 cluster solution.
Figure S6: Predicted settlement types for Maputo, Mozambique using a 4 cluster solution.
6. References

Fraley C, Raftery AE and Scrucca L. (2016) mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. R package version 5.2 ed.
Grafström A and Tillé Y. (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 24: 120-131.
Hastie T, Tibshirani R and Friedman J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag.
Lloyd CT, Chamberlain H, Kerr D, et al. (2019) Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data* 3: 108-139.
Scrucca L, Fop M, Murphy TB, et al. (2016) mclust5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* 8: 289-317.
Wehrens R, Buydens LMC, Fraley C, et al. (2004) Model-Based Clustering for Image Segmentation and Large Datasets via Sampling. *Journal of Classification* 21: 231-253.