Using machine learning to classify urban building footprints into residential / non-residential categories, in low income settings

Lloyd CT[*1], Sturrock HJW[†2], Leasure D[1], Jochem WC[1], Lázár AN[1], and Tatem AJ[‡1]

[1] WorldPop Programme, School of Geography and Environmental Science, University of Southampton, UK
[2] Global Health Group, University of California, San Francisco, CA, USA

June 30, 2020

Summary
A recently developed stacked generalisation, ensemble, machine learning model is freshly applied to a combination of new building footprint and label data representing urban areas in two low income countries in Africa. The footprint and label data are of greater completeness and attribute consistency than previously available. We discuss novel GIS workflow, model, and output. Results show that the model correctly classifies between 85% and 93% of structures as residential and non-residential across both countries. Classified outputs are likely to be valuable across a range of applications, including urban planning, resource allocation, service delivery, and modelling population distributions.

KEYWORDS: machine learning, settlement classification, superlearner, residential, building footprint

1. Extended Abstract

Analyses of modelled human population distributions aid the monitoring of progress towards the achievement of UN Sustainable Development Goals (SDGs) and related agendas such as urban planning, resource allocation, and service delivery. Population models are better informed when correctly labelled high resolution residential building footprint data are available and used as an input covariate. However, buildings particularly in low income countries are often inadequately mapped in terms of coverage and delineation accuracy. Identifying residential buildings via machine learning provides a much needed alternative to time consuming and labour intensive manual (i.e. human) delineation of such objects.

Geospatial characteristics of residential buildings (e.g. pattern, size, proximity to roads, proximity to similar adjacent buildings and land uses) are good signals to inform a machine learning model; and thus to potentially identify residential buildings with better model performance than so far achieved.

Using new building footprint and label data as input, via a newly developed GIS workflow, we present fresh application of an existing algorithm to differentiate residential from non-residential buildings in urban areas in two low income countries (The Democratic Republic of the Congo, COD; and Nigeria, NGA). We then communicate future work pertaining to development of the model.

[*] C.T.Lloyd@soton.ac.uk
[†] hugh.sturrock@ucsf.edu
[‡] A.J.Tatem@soton.ac.uk

We utilise the object based, binary, stacked generalisation, ensemble classification algorithm of Sturrock et al. (2018), and apply it separately to urban areas in COD and NGA. The model predicts on Maxar and OpenStreetMap (OSM) building footprint data, and is trained and tested in country using simplified OSM and UCLA/ORNL/World Bank building 'type' labels in order to produce a binary residential/non-residential building classification. The combined building footprint and label datasets have significantly greater completeness and attribute consistency than has previously been available for these countries. OSM highway data, and US NASA Global Man-made Impervious Surface data (GMIS from Landsat) are also used to develop further attributes (i.e. area, sides, sub-polygons, dist. to road, dist. to neighbouring structure, urbanicity, etc.) from the building footprint data in order to inform the model.

The tweaked classification model workflow runs in the 'Superlearner' package (Polley et al, 2019) within the R environment (R Core Team, 2017) either locally (where computationally viable) on a Windows machine or at Linux command line using the Iridis 5 High Performance Computer located at the University of Southampton. Newly developed GIS workflow prepares the data for input to the model via semi-automated batch GDAL scripts running at Windows or Linux command line. Python, Grass GIS, Saga GIS, and Spatialite scripts form part of the GIS workflow. The workflow will be of most use to those who apply the model to further countries and who use input data from diverse sources, allowing potential expansion of use of the model to low income countries across the world as footprint data become more widely available.

In order to improve the accuracy of building identification in low income countries, forthcoming development of the model will involve the modification and expansion of the classification algorithm to include a greater range of appropriate (data inputs and so) generated attributes in order to improve predictive power. As is frequently the case with such algorithms, the existing model performs very well from a statistical point of view when trained, tested, and predicting within a given country, but shows signs of requiring some (real world) improvement when the output is visually assessed by a human operator. Statistical results show that the model correctly classifies between 85% and 93% of structures as residential and non-residential across both countries. However, visualisation and subsequent analyses of output (via comparison to satellite imagery) highlight that whilst the model is generally very effective in classifying at neighbourhood scale, at street scale some suburban areas can suffer apparent misclassification.

In the future, the model is to be adapted to classify a wider variety of building use (such as informal settlement, mixed use, as well as formal residential/ non-residential) in order to better inform population models.

2. Acknowledgements

The authors acknowledge the use of building footprint and highway data provided by OpenStreetMap (© 2020 OpenStreetMap contributors; geofabrik.de); building footprint data provided by Maxar Technologies (DigitizeAfrica data © 2020 Maxar Technologies, Ecopia.AI); building label data for Kinshasa and North Ubangi, COD, provided by the World Bank (World Bank Group, 2018); and impervious surface data provided by US NASA (SEDAC) (Brown de Colstoun et al, 2017).

The University of California, Los Angeles (UCLA)-Democratic Republic of the Congo (DRC) Health Research and Training Program, the Kinshasa School of Public Health (KSPH), and the Bureau Central du Recensement (BCR) coordinated and conducted the two microcensus rounds in the provinces of Kinshasa, Kongo Central, Kwango, Kwilu, and Mai-Ndombe, COD, during 2018. The Oak Ridge National Laboratory (ORNL) contributed to the first round of microcensus. ORNL designed, and eHealth Africa implemented, the collection of microcensus data in the states of Abia, Adamawa, Akwa Ibom, Bauchi, Ebonyi, Edo, Gombe, Kaduna, Kebbi, Ogun, Oyo, Sokoto, Yobe, and Zamfara, NGA, during 2016-17.

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated

support services at the University of Southampton, in the completion of this work. We also acknowledge the help of the WorldPop Modelling Team for their critique of the GIS workflow discussed in this paper, as well as Heather R. Chamberlain at WorldPop.

This work is part of the GRID3 project (Geo-Referenced Infrastructure and Demographic Data for Development), funded by the Bill and Melinda Gates Foundation and the United Kingdom Department of International Development (DFID) (#OPP1182408). Project partners include the United Nations Population Fund (UNFPA), Center for International Earth Science Information Network (CIESIN) in the Earth Institute at Columbia University, and the Flowminder Foundation.

3. Biography

Christopher T Lloyd is an early career researcher, with diverse research interests that include the application of GIS in support of analyses of human population distribution and disease migration, the use of machine learning to classify settlement, and glacial geomorphology and geology.

References

Brown de Colstoun, E. C., C. Huang, P. Wang, J. C. Tilton, B. Tan, J. Phillips, S. Niemczura, P.-Y. Ling, and R. E. Wolfe (2017). Global Man-made Impervious Surface (GMIS) Dataset From Landsat. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4P55KKF

Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M (2019). Package 'SuperLearner' Documentation. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf

R Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, Available at: https://www.-R-project.org

Sturrock H, Woolheater K, Bennett A, Andrade-Pacheco R, Midekisa A (2018). Predicting residential structures from open source remotely enumerated data using machine learning. PLoS ONE, 13(9), e0204399. https://doi.org/10.1371/journal.pone.0204399

World Bank Group (2018). The World Bank Data Catalog, DRC - Building points for Kinshasa and North Ubangi https://datacatalog.worldbank.org/dataset/building-points-kinshasa-and-north-ubangi