Applying GIS and Machine Learning to Classify Residential Status of Buildings in Low and Middle Income Countries

Lloyd CT[*1], Sturrock HJW[†2] and Tatem AJ[‡1]

[1] WorldPop Programme, Department of Geography and Environmental Science, University of Southampton, SO17 1BJ, UK
[2] Locational, Lytchett House, 13 Freeland Park, Wareham Road, Poole BH16 6FA, UK
* Correspondence: C.T.Lloyd@soton.ac.uk

November 19th, 2021

Abstract

A previously developed, machine learning, building classification model is applied for the first time to each of several lower and middle income countries in Africa; for which new building footprint/ label, and impervious surface data have recently become available. Buildings are classified according to residential or non-residential use. The footprint and label data are of greater completeness and attribute consistency than previously available, and impervious surface data have greater accuracy. The existing modelling has been extended to include semi-urban and rural buildings as well as those in urban settings (i.e. the modelling is now countrywide). We discuss model workflow, statistical results/ classified outputs, and ongoing development of the model to enhance real world predictive performance. Classified outputs are likely to be valuable across a range of applications, including urban planning, resource allocation, service delivery, and modelling population distributions.

KEYWORDS: Geographical Information Systems (GIS), Harmonized data sets, Population geography, Spatial analysis, Data and Methods

1. Extended Abstract

Where there is an absence of recent population and housing census data, human population models can provide high resolution and reliable estimates of population distribution (Wardrop et al. 2018). Such estimates aid the monitoring of progress towards the achievement of UN Sustainable Development Goals (SDGs) and related agendas associated with human health, livelihoods, changes in family patterns, and the local environment (WHO and UN, 2010; UN, 2016).

When population models are better informed by correctly labelled high resolution residential building footprint data, it is possible to equip national, regional, and local levels of government (or non-governmental organisations or private subcontractors) with more accurate datasets of population distribution. These are essential to ensure successful management of urban areas, resource allocation, and service delivery (Lloyd et al. 2020). However, buildings particularly in lower and middle 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 financially costly, time consuming, and labour intensive location surveys.

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.

* C.T.Lloyd@soton.ac.uk
† hugh@locational.io
‡ A.J.Tatem@soton.ac.uk

Using new building footprint/label, and impervious surface data as input, via refinement of a previously developed workflow, we discuss fresh applications of the building classification algorithm to differentiate residential from non-residential buildings in lower income countries in Africa (such as Ghana, GHA, and Guinea, GIN). The existing modelling has been extended to include semi-urban and rural buildings as well as those in urban settings (i.e. output building classifications are now countrywide). Previous modelling (Lloyd et al. 2020), undertaken for the Democratic Republic of the Congo (COD), is repeated and extended using new data in order to improve the quality of outputs.

We utilise the object based, binary, stacked generalisation, ensemble classification algorithm of Sturrock et al. (2018), and apply it separately to urban, semi urban, and rural areas in each of COD, GHA, and GIN, then combining output classified buildings to form contiguous labelled building datasets per each country. The model predicts on Maxar and OpenStreetMap (OSM) building footprint data and is trained and tested in country using simplified OSM building labels. In COD, UCLA/KSPH, World Bank, and new GRID3 Mapping for Health (M4H) building survey labels are utilised in training and testing (in addition to OSM building 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. OSM highway data, and World Settlement Footprint (WSF) Impervious Surface data (DLR, 2021 – pending public release)/ GRID3 Settlement Extents (CIESIN, Columbia University and Novel-T, 2021) are used to inform the model.

The 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. An adaption of existing GIS workflow (Lloyd et al. 2020) prepares data for input to the model via semi-automated batch GDAL scripts running at Windows or Linux command line. Python, Grass GIS, and Spatialite scripts form part of the GIS workflow. The GIS workflow is of 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 and middle income countries across the world as footprint data become more widely available.

We discuss ongoing development of the model to enhance accurate building identification in low and middle income countries (i.e. 'real world' predictive performance), including the modification and expansion of the classification algorithm to comprise a greater range of data inputs and corresponding 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 have previously shown that the model correctly classifies between 85% and 93% of structures as residential and non-residential in different countries (Sturrock et al. 2018; Lloyd et al. 2020). However, visualisation and analysis of output highlights that whilst the model is observed to be generally very effective at neighbourhood scale, that at street scale some suburban areas can suffer apparent misclassification.

Further, we discuss possibilities in terms of adapting the model 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 (© 2021 OpenStreetMap contributors; geofabrik.de); building footprint data provided by Maxar Technologies (DigitizeAfrica data © 2021 Maxar Technologies, Ecopia.AI); building label data for Kinshasa and North Ubangi, COD, provided by the World Bank (World Bank Group, 2018); impervious surface data provided by the German Aerospace Center (DLR). 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 household survey rounds in COD during 2017–2018. The Oak Ridge National Laboratory (ORNL) contributed to the first round of household survey. GRID3 Mapping for Health (M4H) and partners coordinated and conducted the household survey rounds undertaken in COD during 2021.

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 Population Modelling Team for critique of GIS workflow used in this study, as well as project partners and Heather R. Chamberlain at WorldPop for liaison with Maxar and the Bill and Melinda Gates Foundation regarding acquirement of data.

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 Foreign, Commonwealth and Development Office (#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.

References

Center for International Earth Science Information Network (CIESIN), Columbia University and Novel-T (2021). GRID3 Settlement Extents, Version 1.0. Palisades, NY: Geo-Referenced Infrastructure and Demographic Data for Development (GRID3). https://academiccommons.columbia.edu/search?f%5Bseries_ssim%5D%5B%5D=GRID3

Lloyd C.T, Sturrock H.J.W, Leasure D.R, Jochem W.C, Lázár A.N, Tatem A.J (2020). Using GIS and Machine Learning to Classify Residential Status of Urban Buildings in Low and Middle Income Settings. Remote Sens. 12, 3847. https://doi.org/10.3390/rs12233847

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

UN Habitat (2016). World Cities Report: Urbanization and Development - Emerging Futures; United Nations Human Settlements Programme (UN-Habitat): Nairobi, Kenya, p. 262.

University of California - Los Angeles (UCLA) and Kinshasa School of Public Health (KSPH) (2018). Kinshasa, Kongo Central and Former Bandundu Household Surveys in 2017 and 2018; University of California: Los Angeles, CA, USA.

Wardrop N.A, Jochem W.C, Bird T.J, Chamberlain H.R, Clarke D, Kerr D, Bengtsson L, Juran S, Seaman V, Tatem A.J (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proc. Natl. Acad. Sci. USA, 115, 3529–3537.

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

World Health Organization & United Nations (2010). Human Settlements Programme. Hidden Cities: Unmask and Overcoming Health Inequities in Urban Settings; World Health Organization: Geneva, Switzerland, p. 126.