





# QGIS pypopRF Plugin: A tool for the construction of gridded population distribution maps

Tom McKeen<sup>1</sup>, Rhorom Priyatikanto<sup>1</sup>, Borys Nosatiuk<sup>1</sup>, Wenbin Zhang<sup>1</sup>, Elena Vataga<sup>1</sup>, Natalia Tejedor-Garavito<sup>1</sup>, Andrew J. Tatem<sup>1</sup>, and Maksym Bondarenko<sup>1</sup>

<sup>1</sup> WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK  Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 16 December 2025

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/))

## Summary

Accurate data on the spatial distribution of population is an essential source of information to a wide range of environmental, health and sustainable developmental applications. Consequently, there has been considerable work to develop modelled gridded datasets that accurately capture population distributions at subnational scales. However, generating these datasets generally demands specialised expertise of statistical modelling approaches and computational programming methods. To overcome these challenges, the Spatial Data Infrastructure (SDI) Team at WorldPop have developed the QGIS pypopRF Plugin, a high-resolution population mapping tool that uses machine learning and dasymetric techniques within open-source GIS software ([QGIS.org](https://qgis.org), 2025). The QGIS pypopRF Plugin transforms input data into detailed gridded population distribution maps using a random forest (RF) dasymetric modelling approach ([Stevens et al., 2015](#)) that combines census data, building information and various spatial constraints. The core computational functionality of the plugin is provided by the pypopRF Python package; however, to enable non-technical users without programming expertise, the QGIS pypopRF Plugin offers a suite of tools within a graphical interface to facilitate the implementation of top-down population disaggregation methods ([Wardrop et al., 2018](#)). Users can easily adjust input data and settings within a customisable interface. Population modelling can be computationally intensive, therefore the plugin has been developed to subdivide work among multiple subtasks for parallel processing. This paper aims to describe the core functionality and features of the QGIS pypopRF Plugin, and how the plugin can be leveraged to create high-resolution gridded population data.

## Statement of need

Globally, population dynamics display considerable variation across local-regional scales due to the interface of a series of phenomena. Notable amongst these phenomena are demographic trends ([OECD, 2024](#)), urbanisation ([Sun et al., 2020](#)), and migration patterns ([Qiao et al., 2024](#)) which shape human population distribution and change in distinct ways across countries and regions. In this setting, subnational population datasets are essential to capturing these dynamics and therefore a swathe of applications, including public health strategy, disaster risk management, urban planning and resource allocation ([Maneepong et al., 2025](#); [Wardrop et al., 2018](#)).

However, generating gridded population typically requires expertise in statistical methods and programming skills ([Leyk et al., 2019](#); [Tatem et al., 2007](#)). In-fact tools that precede this plugin, pypopRF ([Priyatikanto et al., 2025](#)) and popRF for R ([Bondarenko et al., 2021](#)), are primarily operable at a command-line or scripting level, thereby limiting the accessibility

of these methods from a wider community. Therefore, open-source and replicable tools for population modelling are crucial to bridging the gap between data curators and data users (Mobasheri et al., 2020).

To address these challenges, the QGIS pypopRF plugin seeks to provide an integrated tool that is accessible to a broad range of users, including GIS practitioners, analysts and educators/students. The Plugin embeds the pypopRF package (Priyatikanto et al., 2025) functionality directly into QGIS, providing a single accessible environment for parameter control, model execution and output visualisation. The plugin facilitates a consistent, reproducible workflow for gridded population data creation with detailed logging options. This enhances the transparency of methods, enabling users to document and share model configurations as well as outputs.

## Overview of the QGIS pypopRF Plugin Functionality

The QGIS pypopRF plugin provides a user-friendly interface to the underlying pypopRF Python library (Figure 1), enabling users to generate high-resolution population distribution maps using machine learning and dasymetric mapping techniques.

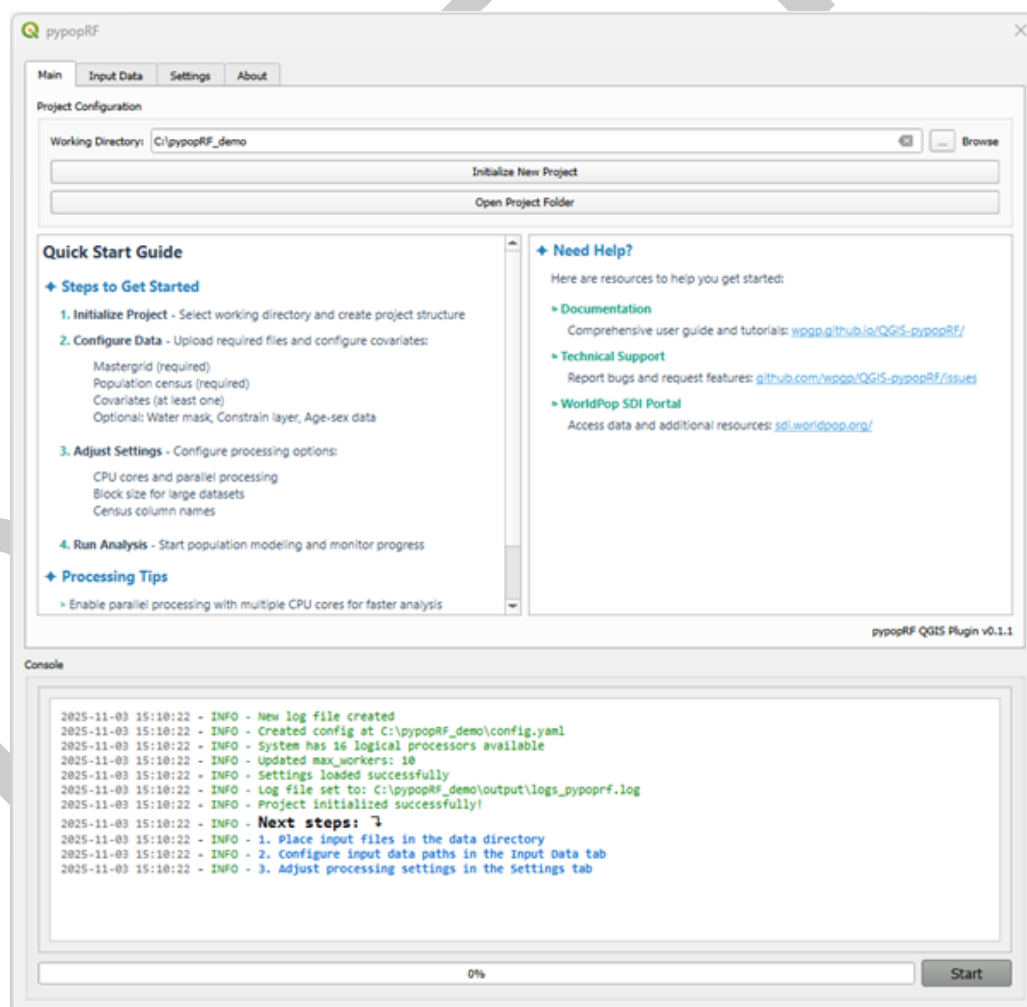


Figure 1: User interface of the pypopRF Plugin for QGIS.

## Project Initialisation & Configuration

The user begins a new project within the “Main” tab of the plugin window, by specifying the address path that the working directory should be configured within. A new project structure including the output directory and a log file will be created at this address path location

The user can flexibly browse to and select input data for analysis via the “Input Data” tab. The plugin differentiates between datasets that are (i) required or (ii) optional to data analysis (Table 1). A raster file that defines zones using unique IDs, referred to here as a “mastergrid”, must be used to delineate census boundaries. Population count data must be supplied as a CSV file, with unique zone IDs that align with the mastergrid zone IDs. Optionally, additional attributes such as age-sex counts may be included in the CSV file. At least one covariate must be added to train the model. This is any geospatial variable that is related to human population distribution such as building location, infrastructure or elevation. Optional inputs allow users to refine the prediction areas to exclude uninhabited areas according to a mask (e.g. water bodies) or to constrain population disaggregation to specific areas (e.g. human settlement footprint).

Table 1. Summary of input data parameters.

Data file	Format	Dependency
Mastergrid raster defining analysis zones	GeoTIFF	Required
Census data with population counts	CSV	Required
Geospatial covariate rasters (e.g. landcover, infrastructure)	GeoTIFF	Required
Water mask for excluding water bodies	GeoTIFF	Optional
Constraint raster to specify areas (e.g. human settlement)	GeoTIFF	Optional
Age-sex population structure data	CSV	Optional

## Settings Configuration

Processing parameters and analysis options can be customised in the “Settings” tab. The plugin provides several options to improve computation performance when processing large datasets. Notably, computation can be parallelised across a user-specified number of CPU cores, whilst large datasets can be processed in smaller, adjustable blocks to improve memory-efficiency.

## Analysis and Interface

The plugin uses a Random Forest model (Stevens et al., 2015) to perform this analysis; a gridded population density weighting layer at the target resolution is created and implemented for dasymetric disaggregation of population counts from census zones into target grid cells as supplied by the mastergrid. The model is trained at administrative unit level using the set of covariates specified by the user.

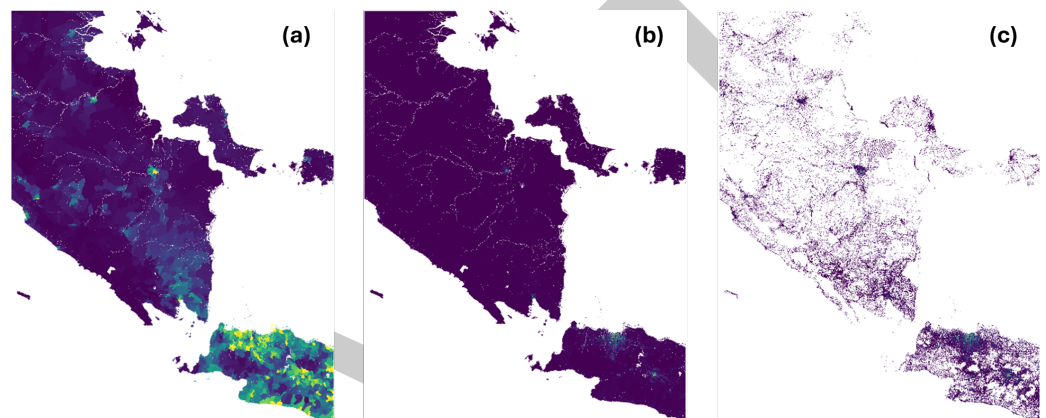
At its core, the plugin leverages the rasterio library (Gillies, 2019) for seamless raster input and output operations, ensuring efficient handling of geospatial covariates. For data computation in a vector format, pandas (team, 2020) and its spatial extension, geopandas (Jordahl, 2019), enable powerful data manipulation and analysis, particularly with vector data that might be associated with the raster information. To optimise performance and manage the workflow efficiently, joblib (developers, 2025) is integrated for pipelining and parallel processing, which significantly speeds up computation by utilizing multi-core processors. Finally, the foundational machine learning capabilities within pypopRF are powered by the extensive algorithms and tools provided by the scikit-learn (sklearn) library (Pedregosa et al., 2011), allowing for the construction and application of sophisticated models. Population can be redistributed more realistically by using masks and constraints. A mask, for example an inland water mask, can be used to prevent population prediction in areas identified as inland water. Moreover, when a constraint is provided, such as human settlement footprint, the model is instructed to

96 create a layer in which population estimation is only within areas mapped as containing built  
97 settlements, rather than across all land grid cells.

98 Analysis can be monitored in the interface by inspecting the percentage progress bar and  
99 the console area. The console also displays error messages, status updates and important  
100 notifications supporting the user with real-time feedback. This feedback can be applied via  
101 a Start / Stop button to control processing. A completion message will be displayed in the  
102 console area once the analysis has concluded.

## 103 Outputs

104 Completion of the analysis generates several output rasters, capturing different levels of analysis  
105 [Figure 2](#).



**Figure 2:** Output rasters from the model. (a) normalised census adjusted values, (b), unconstrained population distribution (default), and (c) constrained population distribution (when a constraint layer is provided).

106 When age-sex population data has been supplied, additional outputs detailing population  
107 distribution of age-sex categories will be generated in the ".../agesex/" directory. Moreover,  
108 several files are created capturing different aspects of the analysis, including detailed processing  
109 logs, feature scalars and importance scores of predictor variables. The latter enables the user  
110 to assess the most influential covariates to the model.

## 111 Acknowledgements

112 This work was supported by funds from the Gates Foundation (INV-045237 and INV-088965)  
113 and Wellcome Trust (308679/Z/23/Z). This work forms part of the outputs of WorldPop  
114 ([www.worldpop.org](http://www.worldpop.org)). The funders had no role in study design, data collection and analysis,  
115 decision to publish, or preparation of the manuscript.

## 116 References

- 117 Bondarenko, M., Nieves, J., Stevens, F., Gaughan, A., Jochem, W., Kerr, D., & Sorichetta, A.  
118 (2021). *Poprf: Random forest-informed population disaggregation r package*. University of  
119 Southampton. <http://dx.doi.org/10.5258/SOTON/WP00715>
- 120 developers, T. joblib. (2025). Joblib: Running python functions as pipeline jobs. In *GitHub*  
121 *repository*. GitHub. <https://github.com/joblib/joblib>

- Gillies, S. (2019). Rasterio: Geospatial raster i/o for python programmers. In *GitHub repository*. GitHub. <https://github.com/rasterio/rasterio>
- Jordahl, K. (2019). Geopandas (version v0.8.1). In *Zenodo repository*. Zenodo. <http://doi.org/10.5281/zenodo.3946761>
- Leyk, S., Gaughan, A. E., Adamo, S. B., De Sherbinin, A., Balk, D., Freire, S., Rose, A., Stevens, F. R., Blankespoor, B., Frye, C., & others. (2019). The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, 11(3), 1385–1409. <https://doi.org/10.5194/essd-11-1385-2019>
- Maneepong, K., Yamanotera, R., Akiyama, Y., Miyazaki, H., Miyazawa, S., & Akiyama, C. M. (2025). Towards high-resolution population mapping: Leveraging open data, remote sensing, and AI for geospatial analysis in developing country cities—a case study of bangkok. *Remote Sensing*, 17(7), 1204. <https://doi.org/10.3390/rs17071204>
- Mobasheri, A., Pirotti, F., & Agugiaro, G. (2020). Open-source geospatial tools and technologies for urban and environmental studies. *Open Geospatial Data, Software and Standards*, 5(1), 5. <https://doi.org/10.1186/s40965-020-00078-2>
- OECD. (2024). *Society at a glance 2024: OECD social indicators*. <https://doi.org/10.1787/918d8db3-en>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12, 2825–2830. [http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post\\_page](http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page)
- Priyatikanto, R., Nosatiuk, B., Zhang, W., McKeen, T., Vataga, E., Tejedor-Garavito, N., & Bondarenko, M. (2025). pypopRF: Population prediction and dasymetric mapping tool. In *GitHub repository*. GitHub. <https://github.com/wpgp/pypopRF>
- QGIS.org. (2025). *QGIS geographic information system*. QGIS Association. <http://www.qgis.org/>
- Qiao, R., Gao, S., Liu, X., Xia, L., Zhang, G., Meng, X., Liu, Z., Wang, M., Zhou, S., & Wu, Z. (2024). Understanding the global subnational migration patterns driven by hydrological intrusion exposure. *Nature Communications*, 15(1), 6285. <https://doi.org/10.1038/s41467-024-49609-y>
- Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, 10(2), e0107042. <https://doi.org/10.1371/journal.pone.0107042>
- Sun, L., Chen, J., Li, Q., & Huang, D. (2020). Dramatic uneven urbanization of large cities throughout the world in recent decades. *Nature Communications*, 11(1), 5366. <https://doi.org/10.1038/s41467-020-19158-1>
- Tatem, A. J., Noor, A. M., Von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in east africa. *PloS One*, 2(12), e1298. <https://doi.org/10.1371/journal.pone.0001298>
- team, T. pandas development. (2020). *Pandas-dev/pandas: pandas*. Zenodo. <https://doi.org/10.5281/zenodo.3509134>
- 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. *Proceedings of the National Academy of Sciences*, 115(14), 3529–3537. <https://doi.org/10.1073/pnas.1715305115>