The Journal of Open Source Software

AgePopDenom: Geostatistical Modeling for Fine-Scale Population Age-Structures

Mohamed A. Yusuf [1,2¶], **Victor A. Alegana** [1], **Chibuzor Christopher Nnanatu** [3,4], **and Andrew Tatem** [3]

1 The World Health Organization Regional Office for Africa (WHO AFRO) **2** Applied Health Analytics for Delivery and Innovation, Nairobi, Kenya **3** WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, United Kingdom **4** Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 23 July 2025
Published: unpublished

Summary

`AgePopDenom` is an R package that generates high-resolution (5 km × 5 km) estimates of population by single year of age using publicly available datasets including Demographic and Health Surveys (DHS) (Fabic et al., 2012), WorldPop gridded population data (Andrew J. Tatem, 2017), and administrative boundaries. It is designed for settings where recent, detailed age data are missing or outdated, common in many low- and middle-income countries.

The package models age distributions using a Gamma distribution framework, well-suited for representing typical age structures. It predicts distribution parameters across space using bivariate geostatistical models with Gaussian processes, capturing spatial autocorrelation in demographic characteristics. Users can include covariates such as urban/rural status to guide predictions. This approach models age structure as a high-resolution surface, capturing spatial heterogeneity rather than assuming uniformity within large regions, producing spatially continuous age-disaggregated estimates even in unsampled locations (Figure 1).

`AgePopDenom` is, to our knowledge, the first R package to operationalize a parameter-based geostatistical approach for modeling full age structures at fine spatial resolution. Unlike existing tools that focus on demographic simulation or limited health indicators, it integrates survey microdata, gridded population rasters, administrative boundaries, and user-defined covariates into a single streamlined workflow. The spatial modeling uses Template Model Builder (TMB) (Kristensen et al., 2016) for efficient C++-based optimization, reducing computation time from 2–3 hours to 7–10 minutes per country.

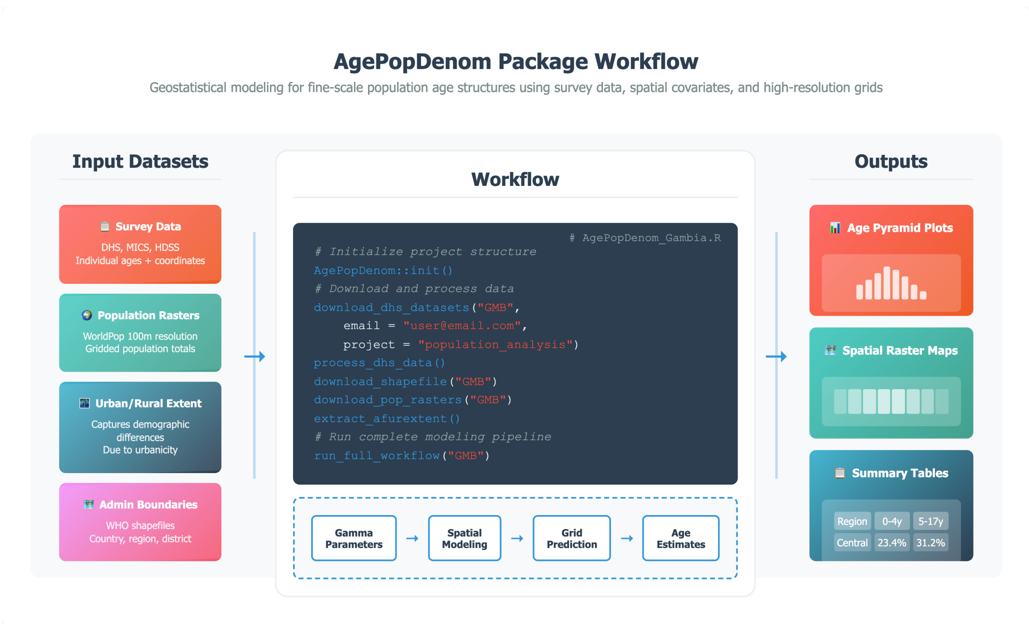

Figure 1: AgePopDenom workflow schematic showing the integration of input datasets, core processing pipeline, and resulting outputs.

Statement of need

Accurate and spatially explicit age-structured population data are essential for public health planning, disease surveillance, and development efforts. Such data form the basis for estimates of disease burden, vaccination coverage, and intervention impact, aiding governments and organizations in allocating resources effectively (Diaz et al., 2021; Hay & Snow, 2006; A. J. Tatem et al., 2014). Yet granular age-disaggregated data remain scarce. Census data are infrequent and often lack subnational detail. National projections offer interim estimates, but their aggregate assumptions limit utility for local planning. Microdata from National Statistical Offices are often inaccessible due to restrictions (Linard et al., 2010; Stevens et al., 2015). Household surveys such as DHS and MICS provide rich demographic data but are not designed for subnational estimation (Boerma et al., 1993; Ye et al., 2012). This creates a persistent gap in small-area, age-structured population estimates.

Despite advances in fine-scale population modeling, particularly for total population and select age groups like under-fives (Alegana et al., 2015; Pezzulo et al., 2017; Stevens et al., 2015; Wardrop et al., 2018), methods for estimating full age distributions remain limited. Traditional techniques like spline interpolation (Fukuda, 2010; McNeil et al., 1977) lack spatial detail, top-down census disaggregation relies on outdated baselines, and bottom-up models typically exclude age (Wardrop et al., 2018). AgePopDenom addresses this gap by modeling continuous age distributions via Gamma parameterization with bivariate Gaussian processes, delivering high-resolution estimates with built-in uncertainty quantification.

State of the field

Several existing R packages address adjacent topics but none support empirical, geostatistical modeling of full age structures at fine spatial resolution. Simulation-oriented tools such as `IBMPopSim` (Giorgi et al., 2023), `demogR` (J. H. Jones, 2007), `demography` (Hyndman, 2024), `popdemo` (Stott et al., 2012), and `mpmsim` (O. R. Jones, 2025) focus on cohort or matrix-based demographic modeling rather than empirical spatial estimation. `SUMMER` (Li et al., 2025) supports small-area estimation of survey-based indicators but is limited to specific outcomes like under-five mortality. Similarly, `ungroup` (Pascariu, 2018), `sptotal` (Higham & Souza, 2019), and `APCtools` (Bauer, 2018) address age unbinning, spatial aggregation, and cohort analysis respectively, without modeling full age structures spatially. The `wpp2024` (United Nations Department of Economic and Social Affairs, Population Division, 2024) package, based on official UN model life table projections, provides standardized national age patterns but lacks spatial granularity and the ability to incorporate covariates.

`AgePopDenom` fills this gap by operationalizing a parameter-based geostatistical approach for modeling full age structures. Rather than contributing to existing packages, we built a standalone tool because the core requirement—joint spatial prediction of Gamma distribution parameters via bivariate Gaussian processes—is architecturally distinct from the single-outcome focus of packages like `SUMMER` or the simulation paradigms of demographic modeling tools. `AgePopDenom` integrates well-maintained packages (`rdhs` for survey microdata, `terra` for gridded population rasters) into a reproducible pipeline with modular functions for each workflow stage.

Software design

`AgePopDenom`'s architecture reflects three key design decisions: (1) a parameter-based approach to age modeling, (2) computational efficiency through compiled code, and (3) an end-to-end automated workflow.

Rather than storing and interpolating raw age counts, which would be memory-intensive and noisy, we model age distributions via Gamma parameters (shape and scale), reducing the problem to predicting two smooth spatial surfaces. This parameterization naturally captures typical age-structure shapes while enabling uncertainty quantification through the fitted distribution. For computational efficiency, spatial modeling is implemented using Template Model Builder (TMB) (Kristensen et al., 2016), which compiles model specifications to C++ and uses automatic differentiation for optimization, making the approach practical for operational use.

Research impact statement

`AgePopDenom` addresses a critical data gap in low- and middle-income countries where detailed age-structured population estimates are essential for public health planning but rarely available at subnational scales. The package represents a novel capability: to our knowledge, it is the first R tool to operationalize geostatistical modeling of full age structures at fine spatial resolution. The primary outputs are age-disaggregated population denominators for downstream epidemiological and demographic analyses.

The package is distributed under an MIT license, includes CI/CD testing across macOS, Windows, and Ubuntu, and provides documented, runnable workflows. The package integrates with established data ecosystems including the DHS Program and WorldPop, facilitating immediate application in settings where these data sources are already used.

Collaboration with WHO AFRO and WorldPop (University of Southampton) makes evaluation and potential near-term use of `AgePopDenom` in operational settings possible. Target applications include vaccination coverage estimation, disease burden assessment, and resource allocation, all domains where age-disaggregated denominators directly influence policy decisions. The package's efficiency makes it practical for routine use by national statistical offices and health ministries, supporting evidence-based health planning in data-sparse settings. Future extensions will incorporate temporal dynamics, Bayesian uncertainty frameworks, and support for additional survey types.

Mathematics

Ages are modeled using a Gamma distribution for each survey cluster. Let A_{ij} represent the age of the j-th individual in cluster i. Each A_{ij} is assumed to follow a Gamma distribution with spatially varying shape $\alpha(x_i)$ and scale $\lambda(x_i)$, modeled as latent fields:

$$A_{ij} \sim \text{Gamma}(\alpha(x_i), \lambda(x_i))$$

The parameters are estimated via maximum likelihood as functions of a bivariate Gaussian Process over spatial locations x_i. The fitted Gaussian process generates predictions of Gamma parameters at unobserved locations, capturing spatial correlation and allowing smooth interpolation across space. The Gamma cumulative distribution function is then applied to compute proportions in any defined age group, supporting fine-scale age-disaggregated estimation.

Example

```r
# Note: Run within an RStudio Project for correct relative paths
install.packages("AgePopDenom")
AgePopDenom::init()
AgePopDenom::download_dhs_datasets("GMB", email = "email@example.org", project = "demo")
AgePopDenom::process_dhs_data()
AgePopDenom::download_shapefile("GMB")
AgePopDenom::download_pop_rasters("GMB")
AgePopDenom::extract_afurextent()
AgePopDenom::run_full_workflow("GMB")
```

AI usage disclosure

Generative AI tools were used during the development of this work. Specifically, AI assistance was used for debugging during software development and for editing and refining the manuscript text. All AI-generated suggestions were reviewed, validated, and modified by the authors. The corresponding author takes full responsibility for the final content.

Acknowledgements

We acknowledge the DHS Program for providing access to survey data and WorldPop for gridded population datasets. This work was supported by funding from the World Health Organization and the University of Southampton.

References

Alegana, V. A., Atkinson, P. M., Pezzulo, C., Sorichetta, A., Weiss, D., Bird, T., Erbach-Schoenberg, E., & Tatem, A. J. (2015). Fine resolution mapping of population age-structures for health and development applications. *Journal of The Royal Society Interface*, *12*(105), 20150073.

Bauer, A. (2018). APCtools: Descriptive and Model-Based Age-Period-Cohort Analysis. *Journal of Open Source Software*, *3*(22), 480. https://doi.org/10.21105/joss.00480

Boerma, J. T., Sommerfelt, A. E., & others. (1993). Demographic and health surveys (DHS): Contributions and limitations. *World Health Stat Q*, *46*(4), 222–226.

The Journal of Open Source Software

Diaz, T., Strong, K. L., Cao, B., Guthold, R., Moran, A. C., Moller, A.-B., Requejo, J., Sadana, R., Thiyagarajan, J. A., Adebayo, E., Akwara, E., Amouzou, A., Aponte Varon, J. J., Azzopardi, P. S., Boschi-Pinto, C., Carvajal, L., Chandra-Mouli, V., Crofts, S., Dastgiri, S., … Banerjee, A. (2021). A call for standardised age-disaggregated health data. *The Lancet Healthy Longevity*, *2*(7), e436–e443. https://doi.org/10.1016/s2666-7568(21)00115-x

Fabic, M. S., Choi, Y., & Bird, S. (2012). A systematic review of demographic and health surveys: Data availability and utilization for research. *Bulletin of the World Health Organization*, *90*, 604–612.

Fukuda, K. (2010). Interpolation and forecasting of population census data. *Journal of Population Research*, *27*, 1–13.

Giorgi, D., Kaakai, S., & Lemaire, V. (2023). Efficient simulation of individual-based population models: The r package IBMPopSim. *arXiv Preprint arXiv:2303.06183*.

Hay, S. I., & Snow, R. W. (2006). The Malaria Atlas Project: Developing Global Maps of Malaria Risk. *PLoS Medicine*, *3*(12), e473. https://doi.org/10.1371/journal.pmed.0030473

Higham, M., & Souza, L. de. (2019). sptotal: An R package for predicting totals and weighted sums from spatial data. *Journal of Open Source Software*, *4*(42), 1679. https://doi.org/10.21105/joss.01679

Hyndman, R. (2024). *Demography: Forecasting mortality, fertility, migration and population data*. https://pkg.robjhyndman.com/demography/

Jones, J. H. (2007). demogR: A package for the construction and analysis of age-structured demographic models in R. *Journal of Statistical Software*, *22*(10), 1–28. http://dx.doi.org/10.18637/jss.v022.i10

Jones, O. R. (2025). Mpmsim: An r package for simulating matrix population models. *Methods in Ecology and Evolution*, *16*(5), 904–911.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, *70*, 1–21.

Li, Z. R., Martin, B. D., Hsiao, Y., Godwin, J., Paige, J., Gao, P., Wakefield, J., Clark, S. J., Fuglstad, G.-A., & Riebler, A. (2025). *SUMMER: Spatio-temporal under-five mortality methods for estimation*.

Linard, C., Gilbert, M., & Tatem, A. J. (2010). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, *76*(5), 525–538. https://doi.org/10.1007/s10708-010-9364-8

McNeil, D. R., Trullell, T. J., & Turner, J. C. (1977). Spline interpolation of demographic oata. *Demography*, *14*, 245–252.

Pascariu, M. D. (2018). ungroup: An R package for efficient estimation of smooth distributions from coarsely binned data. *Journal of Open Source Software*, *3*(29), 937. https://doi.org/10.21105/joss.00937

Pezzulo, C., Hornby, G. M., Sorichetta, A., Gaughan, A. E., Linard, C., Bird, T. J., Kerr, D., Lloyd, C. T., & Tatem, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in africa and asia. *Scientific Data*, *4*(1), 1–15.

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

Stott, I., Hodgson, D. J., & Townley, S. (2012). Popdemo: An r package for population demography using projection matrix analysis. *Methods in Ecology & Evolution*, *3*(5).

Tatem, Andrew J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, *4*(1), 1–4.

Tatem, A. J., Gething, D. P., Pezzulo, C., Weiss, D., & Bhatt, S. (2014). *Development of pilot high-resolution gridded poverty surfaces: Methods working paper* [Technical Report]. University of Southampton.

United Nations Department of Economic and Social Affairs, Population Division. (2024). *wpp2024: An R package for accessing data from the 2024 revision of the world population prospects*. GitHub.

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.

Ye, Y., Wamukoya, M., Ezeh, A., Emina, J. B., & Sankoh, O. (2012). Health and demographic surveillance systems: A step towards full civil registration and vital statistics system in sub-sahara africa? *BMC Public Health*, *12*, 1–11.