GeoLink R package

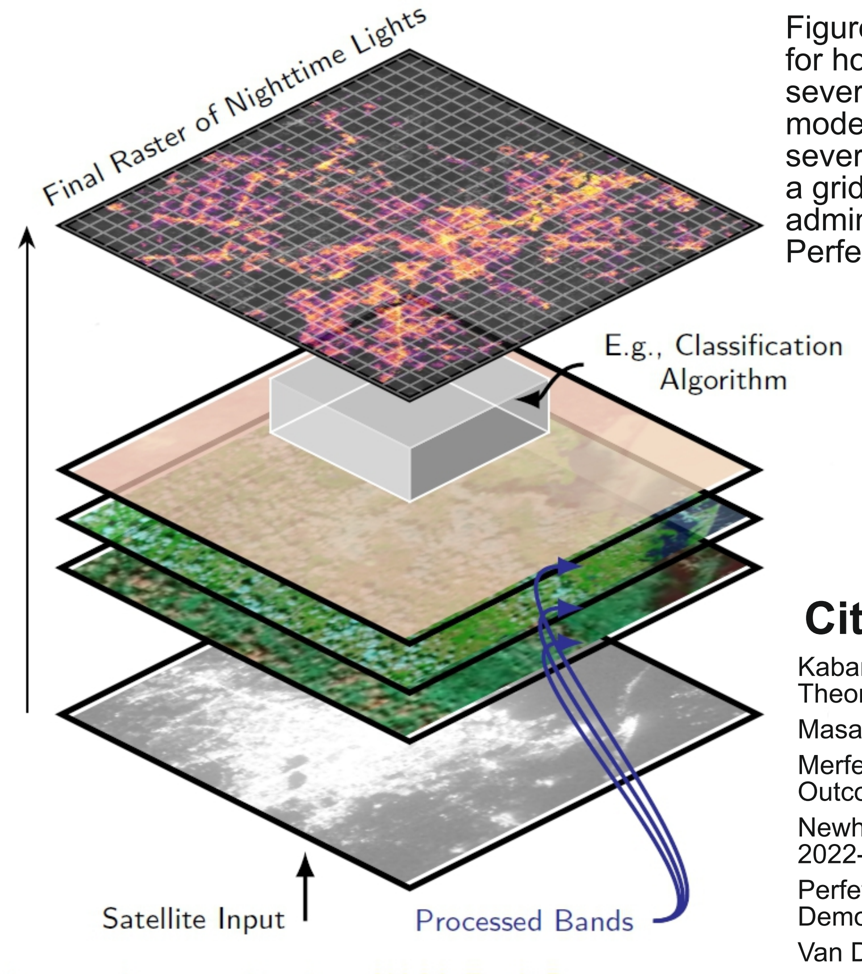
Christopher T Lloyd, Ifeanyi N Edochie, Diana E Jaganjac,
Daylan A.S Gomez, Nikos Tzavidis, and David Newhouse.

Background

GeoLink is an R package that assists users with merging publicly available geospatial indicators with georeferenced survey data. The georeferenced survey data can contain either latitude and longitude geocoordinates, or an administrative identifier with a corresponding shapefile. The procedure involves: Downloading geospatial indicator data, Shapefile tessellation, Computing Zonal statistics, and the spatial joining of geospatial data with administrative unit level data. The package, for example, can be used to link household characteristics measured in surveys with satellite-derived measures such as the average radiance of night-time lights. The package can also calculate indicator values for each pixel, covered by a tessellated grid, in which a household is located. Finally, the package can be used to calculate zonal statistics for a user-defined shapefile (at native resolution or tessellated) and the results linked to survey data. GeoLink complements the povmap and EMDI R packages to facilitate small area estimation with geospatial indicators. The latter two packages enable the Estimation and Mapping of regionally Disaggregated Indicators using small area estimation methods and includes tools for processing, assessing, and presenting the results.

Example Use Case: Applying Geospatial Indicators in Small Area Estimation modelling to estimate poverty

- Household surveys are considered the best source of auxiliary information on the living standards of a country's population nationally when census data are not available
- But the quality of estimates derived from household survey data often diminishes when disaggregated (split up) into local areas or population subgroups
- There are also confidentiality issues, which restrict access to household survey data
- Small Area Estimation (SAE) statistical methods provide more direct estimates when critical data are either not available or have insufficient sample size at lower (e.g. district or state) administrative levels
- Computational and methodological developments have led to the introduction of geospatial data as a valuable source of auxiliary information to be used in SAE models
- Geospatial data, while not substituting survey and census quality, serve as a viable proxy for population either between census surveys or when household survey data are not available
- Geospatial data are easily and freely accessible, frequently updated, and often have global coverage
- Use of geospatial data as a source of auxiliary information has been shown to be promising (Newhouse et al. 2022; Merfeld et al. 2022; Masaki et al. 2022; Kaban et al. 2022; Van Der Weide et al. 2023)
- Within an example SAE model that follows, geospatial data (indicators), potentially output from Geolink and input to povmap or EMDI, undergo several processes where each step attempts to obtain a layer of information
- The final result is an amalgamation of several transformations projected onto a grid
- The final grid cells better summarise the data at smaller administrative unit level (Figure 1)


Final Raster of Nighttime Lights
E.g., Classification Algorithm
Satellite Input
Processed Bands

Sources: Images taken from NASA Earth Data

Figure 1: Geospatial data, as a proxy for household information, undergo several processes within the SAE model, resulting in an amalgamation of several transformations projected onto a grid - summarised in smaller administrative units (Image from Perfetti Villa et al. 2024)

Geospatial Indicator Data

Because of the range of geospatial indicators that are incorporated into GeoLink, the package can be applied in broad economic/policy analyses targeted at specific settings and so has relevance across many different user applications. Indicators currently incorporated into the package include:

- WorldPop human population estimates for individual countries from the year 2000 to 2020
  - Also built settlement constrained and UN adjusted population estimates for the year 2020
  - And more recent population estimates using more accurate bespoke methods for specific Sub-Saharan African countries
- WorldPop Buildings pattern and metrics for 51 developing countries for the year 2020/1
- High Resolution Electricity Access (HREA) data for 115 developing countries for the year 2015
- World Bank OpenCellId Global Cellular Tower Map for the year 2020
- ESA WorldCover cropland dataset for the year 2020
- Shuttle Radar Topography Mission (SRTM) elevation data with near global coverage for year 2000
- Coupled Model Intercomparison Project Phase 6 (CMIP6) future climate scenarios global climate projection data for years 2015-2100
- WorldClim climate temperature, precipitation, solar radiation, wind speed, water vapour pressure data for years 1970-2000
- Terraclimate temperature, precipitation, and other climate and weather data for global terrestrial surfaces from 1958 onwards
- CHIRPS precipitation data from 1981 onwards
- VIIRS Night Time Lights data from 2012 onwards
- Impact Observatory (IO) annual Land Use Land Cover (LULC, 9-class) V2 data from 2017-2023
- OpenStreetMap (OSM) Point of Interest data

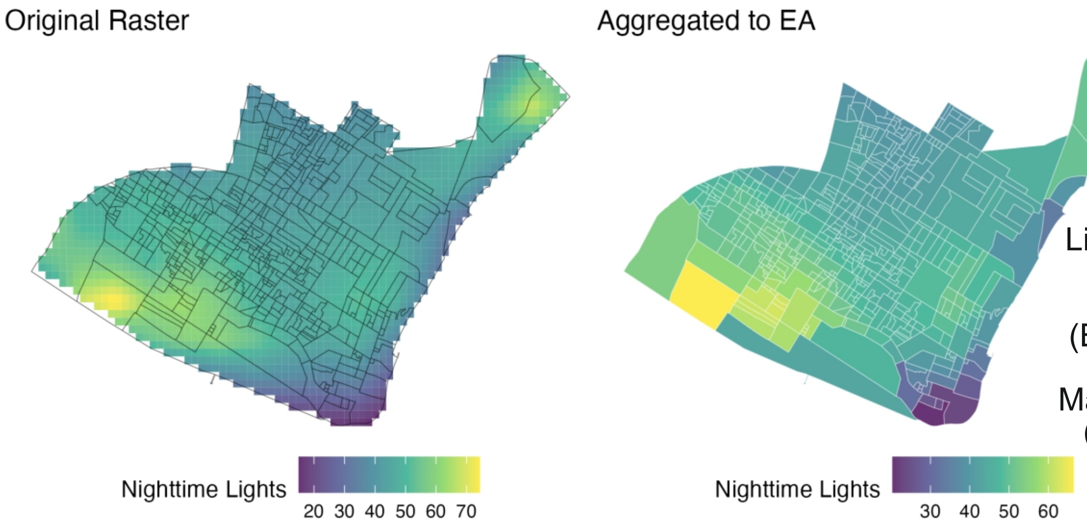
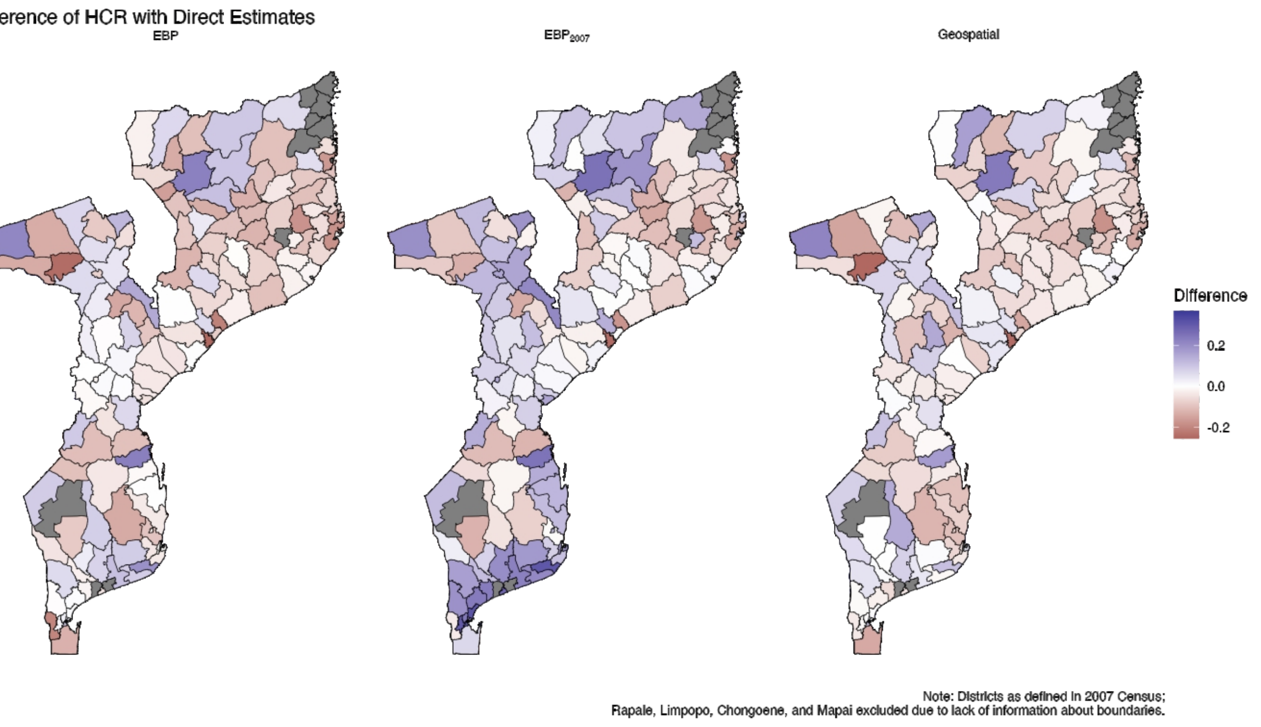


Figure 2: Nighttime Lights data, and after aggregation to enumeration area (EA) in KaMpfumo, a municipal district in Maputo, Mozambique (Image from Perfetti Villa et al. 2024)

- In this example SAE modelling, geospatial data for Mozambique are based on Woods et al. (2024) and comprise 32 primary variables and 18 additional variables, varying in type and time of measurement. The main variables are:
  - Microsoft and Google Buildings, OSM distance to key locations, MERIT elevation and slope, Global Human Settlement Layers, and Nighttime lights
- The geospatial data are aggregated at enumeration area (EA) (e.g. Figure 2)
- SAE models perform well when geospatial data are employed, when compared to the directly modelled estimates (Figure 3)



Note: Districts as defined in 2007 Census;
Rapale, Limpopo, Chongoene, and Mapai excluded due to lack of information about boundaries.

Figure 3: Map of difference in Head Count Ratio (HCR, i.e. the proportion of population living below the poverty line) when geospatial data are employed in modelling, versus the direct Empirical Best Predictor model estimates (EBP) (Image from Perfetti Villa et al. 2024)

Citations

Kaban, Puspita Anggraini et al. (May 26, 2022). "Implementing night light data as auxiliary variable of small area estimation". In: Communications in Statistics - Theory and Methods 0.0. Publisher: Taylor & Francis eprint: https://doi.org/10.1080/03610926.2022.2077963, pp. 1–18.

Masaki, Takaaki et al. (Sept. 13, 2022). "Small area estimation of non-monetary poverty with geospatial data". In: Statistical Journal of the IAOS 38.3, pp. 1035–1051.

Merfeld, Joshua D. et al. (June 2022). Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes. Working Paper. Accepted: 2022-06-10T15:09:30Z. Washington, DC: World Bank.

Newhouse, David et al. (Sept. 2022). Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. Working Paper. Accepted: 2022-09-15T16:34:57Z. Washington, DC: World Bank.

Perfetti Villa, L., N. Tzavidis, and A. Luna Hernandez. (June 06, 2024). On the use of small area estimation with geospatial data. Department of Social Statistics & Demography, University of Southampton. SAE Conference.

Van Der Weide, Roy et al. (2023). How Accurate is a Poverty Map Based on Remote Sensing Data? An Application to Malawi.

Woods, Thea et al. (2024). Spatio-temporally harmonised datasets for Mozambique. Version 1.0.