The Journal of Open Source Software

xagg: A Python package to aggregate gridded data onto polygons

Kevin Schwarzwald [1,2¶] **and Kerrie Geil** [3]

1 Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA **2** International Research Institute for Climate and Society, Palisades, NY, USA **3** Mississippi State University, Mississippi State, MS ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Scientific data is often stored on grids or rasters: gridded weather observations, interpolated pollution data, night-time lights, or other remote sensing products all approximate the continuous real world for ease of calculation, standardization, or technical limiations. However, living things don't live on grids, and rarely act or observe data on grids either. Instead, demographic or agricultural data is often collected on the county or city level, birds fly along complex migratory corridors, and rain- and watersheds follow valleys and mountains, in other words, along areas that can be described using geographic polygons.

When these raster and polygon worlds collide, as they often do in social or natural science research, data must often be aggregated between them (e.g., Auffhammer et al. (2013)). This aggregation must, however, be done with care. Consider a researcher who needs to aggregate temperature data from a gridded reanalysis product onto Los Angeles County, at which level they observe population or mortality statistics (Figure 1). The simplest way to aggregate data would be to average across every grid cell that partially overlaps with the county. However, given the complex topography of the region, a grid cell only slightly overlapping with the county, or only overlapping with the sparsely populated mountains of the county, would be unhelpful if studying the relationship between temperature and society.

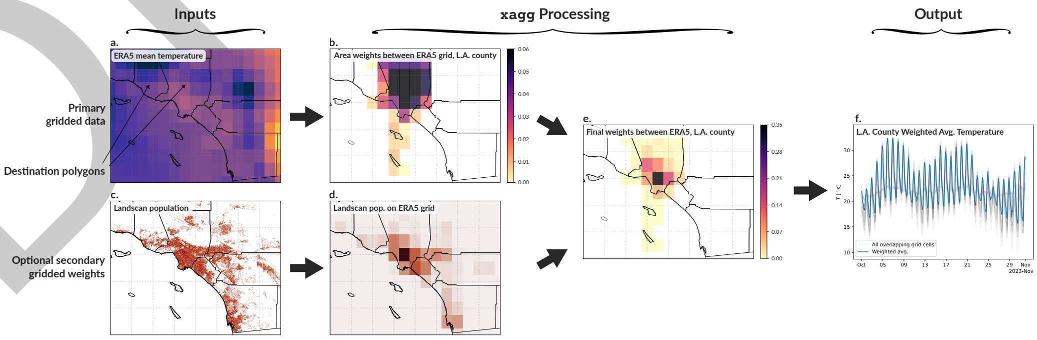

Figure 1: Illustration of xagg workflow. Variables stored on a geographic grid (in this case 2-meter daily temperature from ERA5 reanalysis; Hersbach et al. (2020)), a set of geographic polygons (in this case US county borders, focusing on Los Angeles County as an example), and an optional second weight on a geographic grid (in this case LandScan Day Population; Rose et al. (2017)) are inputted (panels a., c.). xagg calculates the relative overlap between each ERA5 grid cell and each county (panel b.). xagg regrids the population grid to the ERA5 grid (panel d.), and produces a set of final grid cell weights composed of both the area overlap and the population density (panel e.). For each county, these weights are used to calculate weighted averages of daily temperature (panel f.), which can be then be outputted in multiple formats for further analysis.

Therefore, an ideal aggregation would weight not only by the area overlap between grid cells and polygons, but also optionally by other densities of relevant variables - population, area planted, etc. (Auffhammer et al., 2013).

xagg fulfills this need, by providing a simple interface for aggregating raster data stored in `xarray` (Hoyer & Hamman, 2017) `Datasets` or `DataArrays` onto polygons stored in geopandas (Bossche et al., 2024) `geodataframes`, weighted by the fractional area overlap between the raster grid and the polygon, and optionally additionally weighted by a secondary gridded variable (see Figure 1 for a sample workflow). Fractional area weights are generated by constructing polygons for each grid cell and using `geopandas`' `gpd.overlay()` function to calculate the overlaps between input polygons and grid cells. Aggregated data is then returned as an `xarray` `Dataset`, a pandas `DataFrame`, or a geopandas `GeoDataFrame`, depending on the user's needs.

Statement of need

Aggregating gridded data onto polygons is a fundamental aspect of much social and natural science research (e.g., Auffhammer et al. (2013); Hsiang et al. (2017); Carleton et al. (2022); Mastrantonas et al. (2022)). Historically, this process has been conducted on an ad hoc basis by individual research groups, often using simplifications such as averaging over all grid cells that overlap with a county, regardless of the size of that overlap (e.g,. Schlenker & Roberts (2009)).

xagg fills a need for an easy, standardized, and accurate workflow for this aggregation. Working and outputting data in `xarray` and `*pandas` formats (including keeping by default relevant metadata and attributes from the inputted polygons) means xagg can be plugged into a wide array of existing workflows in natural and social sciences, and can easily export aggregated results in formats read by other languages often used in research, including R, QGIS, or STATA.

Though other `python` packages allow aggregation of raster data, to the authors' knowledge, none provide the same depth of functionality. `regionmask` (Hauser et al., 2023)'s `mask_3D_frac_approx` function also approximates relative overlaps between grid cells and regions, for example; this however only works for regular rectangular grids (while xagg works with any rectangular grid), and can be less accurate than xagg's . In addition, none allow easy weighting by a secondary raster variable (e.g., population density or yield), or keep polygon metadata intact.

xagg has already been used in peer-reviewed (e.g., Pulla et al. (2023); Mastrantonas et al. (2022); Schwarzwald & Lenssen (2022)) and upcoming (e.g., Sichone (2024); Peard & Hall (2023)]) scientific publications, has reached over 15,000 cumulative downloads across versions, and is a key component of a how-to guide for climate econometrics (Rising et al., 2024).

Acknowledgements

The authors would like to thank Ryan Abernathy, Julius Busecke, Tom Nicholas, and James Rising for help in getting this project across the ground, in addition to anyone who contributed to GitHub issues or the codebase over the years.

References

Auffhammer, M., Hsiang, S. M., Schlenker, W., & Sobel, A. (2013). Using Weather Data and Climate Model Output in Economic Analyses of Climate Change. *Review of Environmental Economics and Policy*, *7*(2), 181–198. https://doi.org/10.1093/reep/ret016

Bossche, J. V. den, Jordahl, K., Fleischmann, M., Richards, M., McBride, J., Wasserman, J., Badaracco, A. G., Snow, A. D., Ward, B., Tratner, J., Gerard, J., Perry, M., cjqf, Hjelle, G.

The Journal of Open Source Software

A., Taves, M., Hoeven, E. ter, Cochran, M., Bell, R., rraymondgh, … Gardiner, J. (2024). *Geopandas/geopandas: V1.0.1*. Zenodo. https://doi.org/10.5281/zenodo.12625316

Carleton, T., Jina, A., Delgado, M., Greenstone, M., Houser, T., Hsiang, S., Hultgren, A., Kopp, R. E., McCusker, K. E., Nath, I., Rising, J., Rode, A., Seo, H. K., Viaene, A., Yuan, J., & Zhang, A. T. (2022). Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits*. *The Quarterly Journal of Economics*, *137*(4), 2037–2105. https://doi.org/10.1093/qje/qjac020

Hauser, M., Spring, A., Busecke, J., Driel, M. van, Lorenz, R., & readthedocs-assistant. (2023). *Regionmask/regionmask: Version 0.11.0*. Zenodo. https://doi.org/10.5281/zenodo.8370810

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1). https://doi.org/10.5334/jors.148

Hsiang, S., Kopp, R., Jina, A., Rising, J., Delgado, M., Mohan, S., Rasmussen, D. J., Muir-Wood, R., Wilson, P., Oppenheimer, M., Larsen, K., & Houser, T. (2017). Estimating economic damage from climate change in the United States. *Science*, *356*(6345), 1362–1369. https://doi.org/10.1126/science.aal4369

Mastrantonas, N., Furnari, L., Magnusson, L., Senatore, A., Mendicino, G., Pappenberger, F., & Matschullat, J. (2022). Forecasting extreme precipitation in the central Mediterranean: Changes in predictors' strength with prediction lead time. *Meteorological Applications*, *29*(6), e2101. https://doi.org/10.1002/met.2101

Peard, A., & Hall, J. (2023). *Combining deep generative models with extreme value theory for synthetic hazard simulation: A multivariate and spatially coherent approach* (No. arXiv:2311.18521). arXiv. https://doi.org/10.48550/arXiv.2311.18521

Pulla, S. T., Yasarer, H., & Yarbrough, L. D. (2023). GRACE Downscaler: A Framework to Develop and Evaluate Downscaling Models for GRACE. *Remote Sensing*, *15*(9), 2247. https://doi.org/10.3390/rs15092247

Rising, J. A., Hussain, A., Schwarzwald, K., & Trisovic, A. (2024). A practical guide to climate econometrics: Navigating key decision points in weather and climate data analysis. *Journal of Open Source Education*, *7*(75), 90. https://doi.org/10.21105/jose.00090

Rose, A., Weber, E., Moehl, J., Laverdiere, M., Yang, H., Whitehead, M., Sims, K., Trombley, N., & Bhaduri, B. (2017). *LandScan USA 2016*. Oak Ridge National Laboratory. https://doi.org/10.48690/1523377

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. Crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*(37), 15594–15598. https://doi.org/10.1073/pnas.0906865106

Schwarzwald, K., & Lenssen, N. (2022). The importance of internal climate variability in climate impact projections. *Proceedings of the National Academy of Sciences*, *119*(42), e2208095119. https://doi.org/10.1073/pnas.2208095119

Sichone, J. (2024). *Assessment of Groundwater Storage Depletion using GRACE and Land Surface Models in Mzimba District, North Malawi* (No. 2024060149). Preprints. https://doi.org/10.20944/preprints202406.0149.v1