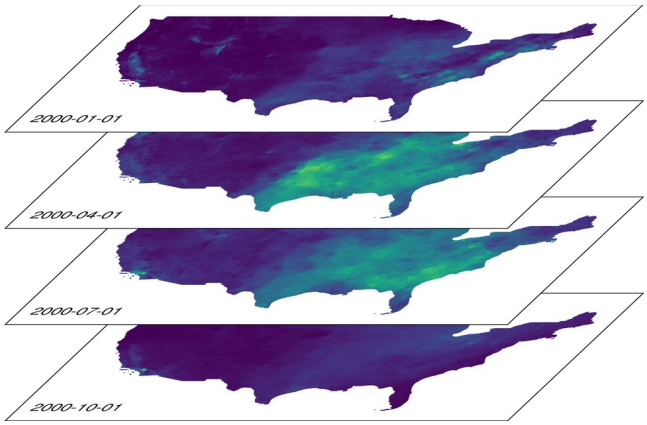
HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH
Department of Environmental Health

Daily, Monthly, and Annual PM2.5 Concentrations for the Contiguous United States, 1-km Grid (2000–2016)

This dataset is available at https://doi.org/10.7910/DVN/58C6HG.



Overview

This dataset contains predicted daily ambient $PM_{2.5}$ concentrations across the contiguous U.S. on a 1-kilometer grid and at monitoring locations, additionally aggregated by month and year. These predictions were produced by a geographically-weighted machine learning ensemble consisting of fitted neural network, random forest, and gradient boosting learners. Overall 10-fold cross-validated R^2 values were 0.86 for daily predictions and 0.89 for annual predictions.

When using this data, please be sure to also cite the following paper, which also details the data sources and processes which resulted in these predictions:

> Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., & Schwartz, J. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. In Environment International (Vol. 130, p. 104909). Elsevier BV. https://doi.org/10.1016/j.envint.2019.104909

Dataset contents

Grid files

The $PM_{2.5}$ prediction grid is provided in two formats: in CSV format (*grid.csv*), for use with statistical programs, and in GeoPackage format (*grid.gpkg*), for geospatial applications. Both files contain 11,218,022 records and include the variable *idx*, which is the ID / row number for that point. The CSV file additionally includes the longitude and latitude in the columns *lon* and *lat*, respectively.

Daily, monthly, and annual predictions

Predicted ambient $PM_{2.5}$ concentrations at daily, monthly, and annual resolutions are available in the *daily-**, *monthly-**, and *annual-** directories, respectively. These predictions are available in two formats: as plain text (in the *daily-dat*, *monthly-dat*, and *annual-dat* directories), and as RDS files (in the *daily-rds*, *monthly-rds*, and *annual-rds* directories). The plain text files are intended for general-purpose use while the RDS files are smaller and optimized for use with the base R programming language.

Each file is a vector of 11,218,022 numeric predictions describing the ambient $PM_{2.5}$ concentrations for the day / month / year indicated by the file name for each of the 11,218,022 points in *grid.csv* / *grid.gpkg*, and the n^{th} item of each prediction file corresponds to the n^{th} point. For example, the 100th record of the file *PM25-2000-01-01.dat* would correspond to the prediction at the 100th point in *grid.csv* / *grid.gpkg* on the 1st of January, 2000.

The units of these predictions are micrograms per cubic meter ($\mu g/m^3$).

Example code

Example code containing routines for reading the data and common merge operations are available in the *example-code* directory along with example spatial data. The file *massachusetts-counties.gpkg* is a subset of the nationwide counties file from the U.S. Census Bureau's TIGER/Line Shapefiles, version 2010[1].

Bulk downloads

To download in bulk, use the page and links generated by the Dataverse API directory index:

> https://dataverse.harvard.edu/api/datasets/:persistentId/dirindex?persistentId=doi:10.7910/DVN/58C6HG

For example, using wGet:

> *wget --recursive --execute robots=off --no-host-directories --span-hosts --content-disposition 'https://dataverse.harvard.edu/api/datasets/:persistentId/dirindex?persistentId=doi:10.7910/DVN/58C6HG'*

1 The original file is available for download at:
https://www.census.gov/geographies/mapping-files/2010/geo/tiger-line-file.html