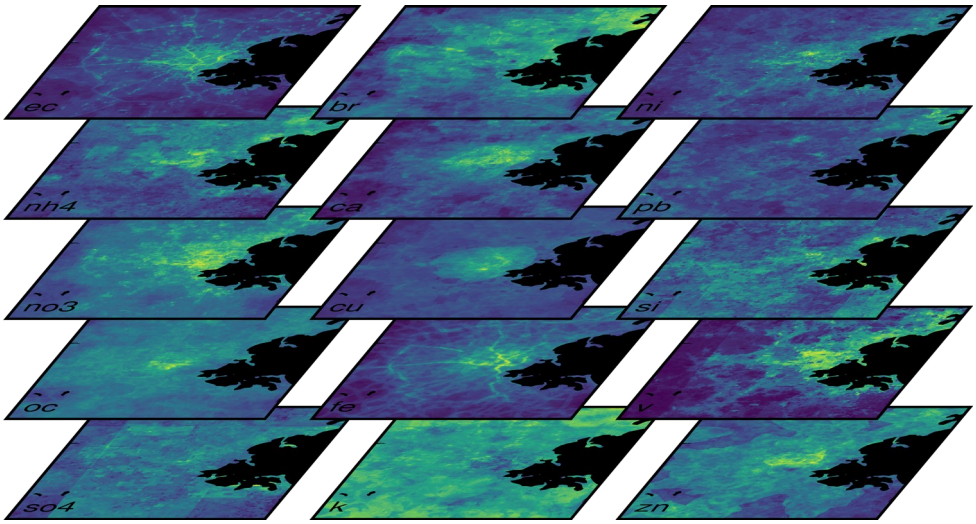
HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH
Department of Environmental Health

Annual $PM_{2.5}$ Component Concentrations for the Contiguous United States, 50-m Urban Grid and 1-km Non-Urban Grid (2000-2019)

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



Overview

This dataset contains predicted annual average ambient $PM_{2.5}$ component concentrations across the contiguous U.S. on a 50-m grid within urban areas[1] and on a 1-km grid outside of those areas. The predictions for each component were produced by separate super-learning ensemble models. As part of this process, the predictions from a variety of machine learners were used as the inputs to train a second set of machine learners (the "super-learners"), of which the super-learner with the highest R^2 was chosen for that component. This process was repeated separately for the urban and non-urban grids. Overall 10-fold cross-validated R^2 values for each component and location are shown in Table 1 at the end of this file.

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

> Amini, H., Danesh-Yazdi, M., Di, Q., Requia, W., Wei, Y., Abu-Awad, Y., Shi, L., Franklin, M., Kang, C.-M., Wolfson, J., James, P., Habre, R., Zhu, Q., Apte, J., Andersen, Z., Kloog, I., Dominici, F., Koutrakis, P., & Schwartz, J. (2022). Hyperlocal super-learned PM2.5 components across the contiguous US. Research Square Platform LLC. https://doi.org/10.21203/rs.3.rs-1745433/v1

1 Defined as being within a 1-km buffer of the urban areas delimited by the U.S. Census Bureau's 2010 TIGER/Line Urban Areas Shapefile. This area is provided as a GeoPackage file in the *example-code* directory.

Amini, H., Danesh-Yazdi, M., Di, Q., Requia, W., Wei, Y., Abu-Awad, Y., Shi, L., Franklin, M., Kang, C.-M., Wolfson, J., James, P., Habre, R., Zhu, Q., Apte, J., Andersen, Z., Xing, X., Hultquist, C., Kloog, I., Dominici, F., … Schwartz, J. (2022). Hyperlocal US PM2.5 Trace Elements Super-learned. Research Square Platform LLC. https://doi.org/10.21203/rs.3.rs-2052258/v1

Dataset contents

Grid files

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

Each prediction grid is divided up by Census division[2], of which there are 9 in total, for a total of 18 distinct grid files. Alaska Hawaii are excluded from the Pacific division. Files are named as follows:

[Census division abbreviation]-[urban / non-urban].[csv / gpkg]

All Census divisions and their included states and abbreviations can be found in Table 2 at the end of this file.

Predictions

Predicted annual average concentrations of each component and location are available in two formats: as plain text (in the *urban-dat* and *non-urban-dat* directories), and as RDS files (in the *urban-rds* and *non-urban-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 numeric predictions describing the annual average ambient concentrations for the component, location, and year indicated by the file name for each point in the corresponding grid, and the n^{th} item of each prediction file corresponds to the n^{th} point. For example, the 100^{th} record of the file NE1-urban-ec-2000.*dat* would correspond to the prediction at the 100^{th} point in NE1-urban.*csv* / NE1-urban.*gpkg* for the year 2000.

The units of these predictions are in micrograms per cubic meter ($\mu g/m^3$) for the major components and nanograms per cubic meter (ng/m^3) for the trace elements. The classifications of components as major components vs. trace elements are shown in Table 1 at the end of this file.

Note that the contents of ZIP files may vary due to Dataverse size restrictions. For example, the non-urban predictions for the Mountain region in RDS format are split into separate ZIP files for the major components (~840 MB) and trace elements (~1.9 GB) because the size of a combined ZIP file would exceed what is allowed by Dataverse (2.5 GB). In these cases, the ZIP file names will include extra information about what subset is included in that ZIP file, for example “fe-major-components.zip” vs. “fe-trace-elements.zip” or “fe-2000-to-2009.zip” vs. “fe-2010-to-2019.zip”.

2 See: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf (“Census Regions and Divisions of the United States”)

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-block-groups.gpkg* is derived from the Massachusetts block groups file from the the U.S. Census Bureau's TIGER/Line Shapefiles, version 2010[3].

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/*3H7DNP*

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/3H7DNP'*

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

Tables

Table 1: Full names, classifications, and urban and non-urban R^2 values for predicted $PM_{2.5}$ components.

Component	Classification	R^2 for urban areas	R^2 for non-urban areas
Elemental carbon (EC)	Major component	0.910	0.913
Ammonium (NH_4^+)	Major component	0.910	0.918
Nitrate (NO_3^-)	Major component	0.879	0.878
Organic carbon (OC)	Major component	0.856	0.857
Sulfate (SO_4^{2-})	Major component	0.952	0.957
Bromium (Br)	Trace element	0.859	0.863
Calcium (Ca)	Trace element	0.865	0.875
Copper (Cu)	Trace element	0.797	0.787
Iron (Fe)	Trace element	0.875	0.875
Potassium (K)	Trace element	0.868	0.866
Nickel (Ni)	Trace element	0.848	0.850
Lead (Pb)	Trace element	0.833	0.843
Silicon (Si)	Trace element	0.851	0.857
Vanadium (V)	Trace element	0.838	0.841
Zinc (Zn)	Trace element	0.878	0.881

Table 2: Census divisions and their associated regions, states, abbreviations, and the number of grid cell points in the urban vs. non-urban prediction areas.

Census division	States included	Abbr.	Urban	Non-urban
New England	CT MA ME NH RI VT	NE1	10,750,894	176,590
Middle Atlantic	NJ NY PA	MA2	21,308,981	195,225
East North Central	IL IN MI OH WI	ENC3	29,712,644	529,891
West North Central	IA KS MN MO ND NE SD	WNC4	13,431,160	1,228,700
South Atlantic	DC DE FL GA MD NC SC VA WV	SA5	44,100,142	565,638
East South Central	AL KY MS TN	ESC6	15,769,846	407,153
West South Central	AR LA OK TX	WSC7	24,000,000	1,012,195
Mountain	AZ CO ID MT NM NV UT WY	M8	13,668,782	2,074,693
Pacific (CONUS)	CA OR WA	P9	20,021,752	738,363