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

Annual $PM_{2.5}$ Component Concentrations for the Contiguous United States, aggregated to common geographies (2000-2019)

This dataset is available at https://doi.org/10.7910/DVN/F6PQAA.

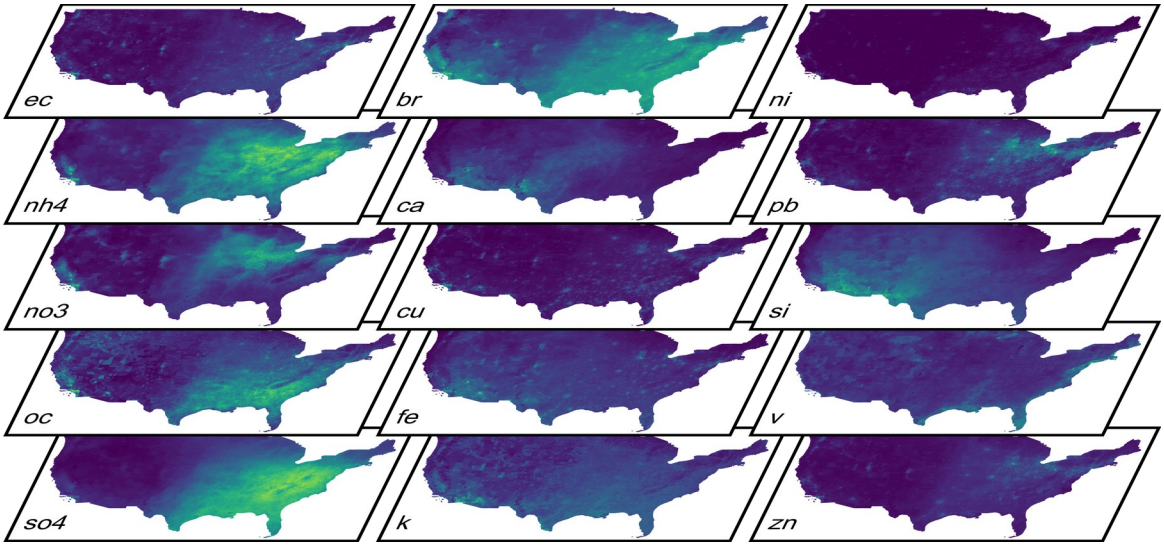

Table of Contents

Overview 2

Exposure assignment process 3

Block-level assignments 3

Higher-level aggregates 3

Dataset contents 4

Statistics 4

Predictions 5

A note on annual predictions and geographic boundary versions 5

Bulk downloads 6

Tables 7

Overview

This dataset contains predicted annual ambient $PM_{2.5}$ component concentrations across the contiguous U.S., aggregated to various commonly-used geographic levels. These aggregates were derived from predictions that were originally made on a 50-m grid within urban areas[1] and on a 1-km grid outside of those areas. For more details on how these predictions were produced, see the following two publications detailing the model data sources, training, and prediction processes:

> 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

> 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

When using this data, please also cite the above two papers where possible.

The raw outputs of the above models, from which the contents of this dataset were produced, are available on Dataverse:

> Amini, H., Castro, E., Danesh-Yazdi, M., Di, Q., Requia, W. J., Wei, Y., Abu Awad, Y., Shi, L., Franklin, M., Kang, C.-M., Wolfson, J. M., James, P., Habre, R., Zhu, Q., Apte, J. S., Andersen, Z. J., Xing, X., Hultquist, C., Kloog, I., … Schwartz, J. (2025). Annual PM2.5 Component Concentrations for the Contiguous United States, 50-m Urban Grid and 1-km Non-Urban Grid (2000-2019) [Dataset]. In Air Quality Data for Health-Related Applications (Version 2.0). Harvard Dataverse. https://doi.org/10.7910/DVN/3H7DNP

---

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.

Exposure assignment process

Block-level assignments

Before any higher-level aggregations, each Census block was first assigned one or more prediction points from the original 50-m urban and 1-km non-urban prediction grids. This assignment was performed as follows:

1. Perform a one-to-many point-in-polygon join from each Census block to the prediction points they encompass, if any;
2. Perform a one-to-one nearest-neighbor join from each Census block to the nearest prediction point, by Cartesian distance.

Then, annual $PM_{2.5}$ component concentration predictions were assigned to each Census block as follows:

1. For Census blocks that encompassed at least one prediction point: take the simple average of predictions from all encompassed points in each year;
2. For all other Census blocks: take the value of the nearest prediction point in each year.

Information on which assignment method was used for each Census block is available; see the *Statistics* section below.

This assignment was done separately for the years 2000 and 2010, since Census geographic boundaries change every 10 years.

Higher-level aggregates

For each geographic level higher than Census blocks, we first created a one-to-many crosswalk from polygons in that geographic level to Census blocks, either by truncating the Census block ID[2] or by using a point-in-polygon join from each polygon to the centroids of Census blocks they encompass. Aggregate values were then calculated as follows:

1. **For each polygon with non-zero population according to encompassed blocks**: take the weighted average of assigned predictions of those encompassed blocks, weighted by the block-level populations;

---

2 Used for block groups, tracts, counties, and states; see the table "GEOID Structure for Geographic Areas" from https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

2. **For all other polygons**: take the simple average of encompassed blocks.

Information on which assignment method was used for each polygon is available; see the *Statistics* section below.

Each set of aggregations is based on the assignments of the most recent set of Census blocks and populations of the most recent Decennial Census. For geographies released during a Decennial Census year (i.e. 2000 or 2010), the blocks and populations from that same year were used.

Dataset contents

Statistics

Each geographic level has an associated statistics file in the *statistics* directory detailing:

1. **For Census blocks**: the encompassing Census division (*division*), whether or not that block was considered urban (*urban*)[3], the population (*population*), the number of encompassed prediction points (*n_points_within*), the distance to the nearest prediction point in decimal degrees (*distance_nearest*), and whether or not associated predictions are based on the average of encompassed points (*used_within*; dichotomous 1/0).

2. **For all other polygons**: the encompassing Census division (*division*)[4], the number of encompassed Census blocks (*n_blocks*), the number of prediction points represented (*n_points*)[5], the number of encompassed blocks that were urban (*n_urban_blocks*), the population (*population*), the population contributed from blocks considered urban (*urban_population*), whether or not that polygon was considered urban (*urban*; dichotomous 1/0)[6], and whether or not the concentrations are the the result of population-weighted averages (*pop_weighted*; dichotomous 1/0; as opposed to simple averages).

Each file is named using the following convention:

> [*geographic level*]-[*year that boundaries were published*].[*csv/rds*]

---

3 Defined as whether or not that Census block's centroid was within urban areas (see footnote 1).

4 Only provided if all encompassed blocks are in the same Census division; NA otherwise.

5 Defined as the sum of the encompassed prediction points (n_*within*) and the number of cases where only the nearest prediction point was used, i.e.: *n_urban_blocks* = *n_within* + (1 - *used_within*).

6 Defined as if all encompassed Census blocks are urban.

Predictions

Predicted annual average concentrations of each component and location are available in two formats: as comma-separated values (CSV files), and as RDS files. The comma-separated 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 contains the identifier for each polygon in that geographic level (as the first column), the year of a given prediction (*year*), and the predicted aggregate concentrations, named after the components that they represent (e.g. *ec* for elemental carbon, *so4* for sulfate, etc.). 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.

Each file is named using the following convention:

> [*geographic level*]-[*year that boundaries were published*][-*[*component*] if block-level*]. [*csv/rds*]

For example, the file *aggregates/block-groups-2000.csv* contains annual predictions for the Census blocks boundaries file that was published in the year 2000. A list of all available geographic levels and their associated prefixes are shown in Table 2 at the end of this file.

Due to Dataverse size restrictions, the block-level files are split up by component and each file will only contain the component specified in the filename.

A note on annual predictions and geographic boundary versions

Note that each file of aggregated concentrations contains the full range of predictions to allow for temporal comparisons among spatially consistent geographies. In other words, regardless of what year is shown in the filename, all files will contain aggregates from 2000 to 2019.

For example, you may wish to plot how the predicted concentrations of elemental carbon at the Census block group level vary over time, using the year 2010 definition of block groups. In this case, you would load the file *aggregates-csv/block-groups-2010.csv* or *aggregates-rds/block-groups-2010.rds*, both of which contain annual predictions from 2000-2019. It would not be valid to pull pre-2010 predictions from the *aggregates-csv/block-groups-2000.csv* or

aggregates-rds/block-groups-2000.rds files, since these are two different sets of boundaries and block groups may have been added, removed, or changed.

Similarly, suppose that you are conducting an epidemiological study and wish to assess an individual's long-term exposure history, using home tract-level exposures, and the event of interest occurred in 2010. In this case, it would be best to load only *aggregates-csv/tracts-2010.csv* or *aggregates-rds/tracts-2010.rds* and use the predictions from that file alone. As with the previous example, it would not be valid to reassess what their Census tract in 2009 would have been and then pull the corresponding data from *aggregates/tracts-2000.csv* or *aggregates/tracts-2000.rds*, since their home tract in 2009 and their home tract in 2010 may not be comparable.

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/F6PQAA*

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/F6PQAA'*

Tables

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

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: Sources of geographic boundary files and the prefixes they appear as in the data set.

Geography	Source	Prefix
U.S. Census Bureau TIGER/Line blocks, versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Blocks	blocks-2000 blocks-2010
U.S. Census Bureau TIGER/Line block groups, versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Block+Groups	block-groups-2000 block-groups-2010
U.S. Census Bureau TIGER/Line tracts, versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Census+Tracts	tracts-2000 tracts-2010
U.S. Census Bureau TIGER/Line counties, versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Counties+%28and+equivalent%29	counties-2000 counties-2010
U.S. Census Bureau TIGER/Line states, versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=States+%28and+equivalent%29	states-2000 states-2010
U.S. Census Bureau ZIP code tabulation areas (ZCTAs), versions 2000 and 2010	https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=ZIP+Code+Tabulation+Areas	zctas-2000 zctas-2010
TomTom / ESRI ZIP code areas and PO boxes[7], yearly-updated[8]	https://www.arcgis.com/home/item.html?id=8d2012a2016e484dafaac0451f9aea24	zip-code-areas-20XX zip-code-po-boxes-20XX

7 TomTom / ESRI ZIP code areas are polygons and the PO boxes are points. PO box assignments are copied from the encompassing ZIP code areas.

8 The 2000 to 2009 versions of TomTom / ESRI ZIP code areas and PO boxes are based on the 2000 version of Census blocks, while the 2010 to 2019 versions are based on the 2010 version of Census blocks.