Data in brief 26 (2019) 104421

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A geospatial environmental concentrations database of Oklahoma, United States

Naci Dilekli [a, b], Sameer V. Gopalani [c, *], Janis E. Campbell [c], Amanda E. Janitz [c]

[a] Department of Mechanics and Mechatronics of the Institute of Space Technologies of Academy of Engineering, Peoples' Friendship University of Russia, Moscow, Russia
[b] Center for Spatial Analysis, University of Oklahoma, Norman, OK 73019, USA
[c] Department of Biostatistics and Epidemiology, Hudson College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, 73104, USA

ARTICLE INFO

Article history:
Received 4 June 2019
Received in revised form 29 July 2019
Accepted 13 August 2019
Available online 24 August 2019

Keywords:
Environmental concentrations
Geographic information systems
Oklahoma
Spatiotemporal distribution

ABSTRACT

Environmental factors can affect human health throughout the lifespan. Reliable and accurate data are needed to understand and establish relationships between environmental factors and health outcomes. In this article, spatiotemporal data (across time and space) on environmental concentrations were compiled in a database for the State of Oklahoma, United States. Data were collected from local, state, and federal government agencies, and organized into a metadata document, which includes spatial extent (information on the area covered), attributes (i.e., variables such as chemical concentration), and temporal extent (time period) of the dataset, among others. Data have been cataloged for concentrations found in water (n = 53 files), air (n = 15 files), land (n = 7 files), and industry (n = 3 files). Data also included physical characteristics (i.e., data on location, geology, and features of waterways, watersheds, and lakes, among others, n = 31 files) and administrative datasets (i.e., data on location and distribution of county boundaries and tribal statistical areas and reservations for federally recognized tribes in Oklahoma, n = 4 files). The main result is a collection of a wide range of spatially-resolved concentration data. This spatiotemporal database will assist in future epidemiologic investigations and assessment of the geographic

* Corresponding author.
E-mail address: sameer-gopalani@ouhsc.edu (S.V. Gopalani).

https://doi.org/10.1016/j.dib.2019.104421

and temporal distribution of environmental exposures in Oklahoma.

Specifications Table

Subject area	Environmental health.
More specific subject area	Environmental concentrations.
Type of data	Geographic Information System (GIS) database.
How data was acquired	Data on environmental concentrations were acquired from local, state, and federal government agencies in the US and imported to a geodatabase using ArcGIS for Desktop, version 10.5 (Environmental Systems Research Institute, Redlands, California)
Data format	Raw and spatialized.
Experimental factors	For the categories of water, air, land, and industry, concentration data were collected from national, state, and local sources.
Experimental features	Spatial data (e.g., shapefiles) were compiled and imported into a geodatabase. Non-spatial text files were spatialized using latitude and longitude information and transformed into feature classes when possible. All feature classes were projected to the same coordinate system, USA Contiguous Albers Equal Area Conic USGS version.
Data source location	Oklahoma, United States.
Data accessibility	Data are included with this data brief.

Value of the data

- The database combines data from multiple domains and provides an extensive picture of environmental concentrations in Oklahoma.
- The database is a valuable resource for investigators and health professionals interested in public health and geospatial research and practice.
- The database can be used for future analytical and surveillance studies to identify the geographic and temporal distribution of environmental concentrations in Oklahoma.

1. Data

Data have been cataloged for environmental concentrations found in water (n = 53 files), air (n = 15 files), land (n = 7 files), and industry (n = 3 files). Data also included physical characteristics (i.e., data on location, geology, and features of waterways, watersheds, and lakes, among others, n = 31 files) and administrative datasets (i.e., data on location and distribution of county boundaries and tribal statistical areas and reservations for federally recognized tribes in Oklahoma, n = 4 files). Table 1 demonstrates the breakdown of these data items by the data formats of a non-spatial table, point layer, line layer, polygon layer, and raster images. Additionally, Table 2 includes a summary and description of the metadata document that contains information on the data source, geometry type, attributes (i.e., variables, including concentrations), data processing, and temporal (time period) and spatial (area covered by the dataset) extents. The metadata table is provided as an Excel file (Microsoft Corporation, Redmond, Washington) in the supplementary material (Supplementary Table 1) and is available online as well [1].

2. Experimental design, materials, and methods

2.1. Data categorization, exploration, and identification

Based on a literature review, the main categories of data collected for the database include water, air, land, and industry. We additionally included data for physical characteristics and administrative datasets (described above). Data on physical characteristics and administrative datasets are included to provide a complete dataset, inclusive of common auxiliary data needed in typical epidemiological

Table 1
Data items by theme and format in the database, Oklahoma, United States.

Theme	Table	Point	Line	Polygon	Raster	Total
Administrative	0	1	0	3	0	4
Air	0	7	1	7	0	15
Industrial	0	3	0	0	0	3
Land	0	3	0	4	0	7
Physical Characteristics	0	3	4	21	3	31
Water	11	27	4	11	0	53
Total	11	44	9	46	3	113

Table 2
Attributes of the metadata document of the database, Oklahoma, United States.

Column	Name	Description
A	Theme/Feature Dataset	Information on dataset domain (water, air, land, industry, physical characteristics, administrative).
B	Data Group Name	Group name for different sources of data.
C	Data Item Description	Purpose of including the data.
D	Original Data Item Name	Original file name from source.
E	Processed Table/Layer Name	Revised table or layer name after importing into ArcMap.
F	Geometry Type	Information on spatial representation, such as line, point, polygon, raster, and table.
G	Attributes	Attribute information, such as chemical concentration, days active, etc.
H	Original Spatialization Status	Information on whether the data were spatialized.
I	Spatial Extent	Information on the area covered in the dataset for Oklahoma.
J	Temporal Extent	Information on the time period of the dataset.
K	Variables	List of variables included in each dataset
L	Publication Date	Information on when the dataset were published.
M	Processing Description	Information on how the data were spatialized.
N	Notes	Additional information on the dataset.
O	Source	Information on the source of the dataset, such as the Environmental Protection Agency, etc.
P	URL or Contact Point	URL or contact person for the dataset.

research in Oklahoma. Additional information on the purpose of these datasets is available in column C of the supplementary material (Supplementary Table 1) and online reference [1]. We conducted permutations of web searches under each category with secondary terms such as concentration or pollutant and tertiary terms such as the US, national, or Oklahoma (Table 3). Accordingly, we used all combinations of these search terms (e.g., freshwater pollution Oklahoma) to find the most relevant data. For each category, we compiled data sources and their metadata as shown in Table 2, and acquired the data layers (known as shapefiles or feature classes) and tables under each category. We searched for data sources online and also contacted national, state, and local institutions, such as the Oklahoma Department of Environmental Quality to request and obtain relevant data.

Table 3
Search terms used for identifying data for the database, Oklahoma, United States.

Category	Search Terms		Secondary Terms		Tertiary Terms
Water	Freshwater Groundwater	+	Concentration Pollutant	+	USA US
Air	Traffic Exhaust		Toxin Carcinogen		National Oklahoma
Land	Pesticide Herbicide Minerals Mining		Hazard		
Industry	Chemicals Waste				

Table 4
Select compounds and datasets with spatial, temporal, and variable information in the database.

Compound/Dataset	Spatial Extent	Temporal Extent	Select Variables
Carbon monoxide	State of Oklahoma	1980–2017	Daily max 8-h CO concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	1990	Pollutant, NAA status, class
Nitrogen dioxide	State of Oklahoma	1980–2017	Daily max 1-h NO_2 concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	1990	Pollutant, class
Ozone	State of Oklahoma	1980–2017	Daily 8-h ozone concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	2008	Pollutant, NAA status, class
Lead	State of Oklahoma	1980–1997 and 2005–2017	Daily mean Pb concentration, units, AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	2008	Pollutant, NAA status, class
Particulate matter ($PM_{2.5}$)	State of Oklahoma	1999–2017	Daily mean $PM_{2.5}$ concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	2006	Pollutant, NAA status, class
Particulate matter (PM_{10})	State of Oklahoma	1988–2017	Daily mean PM_{10} concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	1990	Pollutant, class
Sulfur dioxide	State of Oklahoma	1980–2017	Daily max 1-h SO_2 concentration, units, daily AQI value, AQS parameter description, daily observation count, POC
	Nonattainment areas in entire US	2010	Pollutant, NAA status, class
EPA Toxic Release Inventory	State of Oklahoma	1987–2016	Chemical, CAS #/compound ID, classification, metal, metal category, carcinogen, unit of measure
EPA Water System Violation Report—Copper	State of Oklahoma	07/01/1999–12/31/2016	Contaminant name, sample measure (mg/L), sampling start date, sampling end date
EPA Water System Violation Report—Lead	State of Oklahoma	07/01/1992–12/31/2016	Contaminant name, sample measure (mg/L), sampling start date, sampling end date
OK DEQ Water Sampling—Nitrate	State of Oklahoma	01/03/2011–12/15/2017	Analyte, concentration, sample ID
OK DEQ Water Sampling—Arsenic	State of Oklahoma	01/04/2017–09/18/2017	Analyte, concentration, sample ID
OK DEQ Water Sampling—Methylmercury	State of Oklahoma	10/2007–11/2016	Lake name, species name, length (mm), Hg in fish (mg/kg), weight (g), length (in)

Note: A detailed list of variables is provided in Supplementary Table 1 and online reference [1].
Abbreviations: AQI, air quality index; AQS, air quality system; CAS, Chemical Abstracts Service; CO, carbon monoxide; EPA, U.S. Environmental Protection Agency; Hg, Mercury; NAA, Nonattainment area status; NO_2, nitrogen dioxide; OK DEQ, Oklahoma Department of Environmental Quality; Pb, Lead; POC, parameter occurrence code.

2.2. Data compilation, processing, and development

Data for the environmental database originated from a variety of national, state, and local government departments and agencies. Natural units contained hydrology types, soil types, and other natural boundaries and administrative features contained within the county and tribal boundaries. Data available online were downloaded directly, and in some cases, we contacted necessary agencies and obtained the data that were not readily available online.

Most of the data from the government departments and agencies were in a spatial format (e.g., shapefiles) and imported to a geodatabase using ArcGIS for Desktop, version 10.5 (Environmental Systems Research Institute (ESRI), Redlands, California). When possible, extracted or provided data not available in this format (e.g., non-spatial text files) were spatialized using latitude and longitude information and transformed into feature classes. Non-spatial tables that could not be spatialized were still added to the geodatabase to achieve maximum completeness. All feature classes were projected to the same coordinate system, USA Contiguous Albers Equal Area Conic USGS version. Several data files were saved and separated by year of data collection. After importing and creating a feature class, a new column was created with the year of data collection, if it were a multiyear dataset. Multiple feature classes containing the same type of data were then merged and sorted by date. The spatial data in our database includes data of point (e.g., wells), line (e.g., motorways and waterways), and polygon (e.g., agricultural and superfund zones) based layers. Once all the relevant layers/feature classes were imported to the database with their respective feature datasets, each feature class's metadata was added manually if not already included previously. A list of select compounds and variables with spatial and temporal information is provided in Table 4, and a detailed list of variables in each dataset is available in column K of the supplementary material (Supplementary Table 1) and online reference [1]. Datasets from all domains include variables on compounds (such as type of sample, sample measure, and sample collection and date), geometry (such as shape), location (such as county and zip code), and coordinate system (such as latitude and longitude).

Acknowledgments

The authors are grateful for the efforts of national, state, and local departments and agencies for collecting and sharing data used in this database. The authors appreciate the invaluable effort and contribution from Ms. Cynthia Baker on developing the database.

All authors were supported by the Oklahoma Center for the Advancement of Science and Technology (OCAST) [grant number: HR16-048]. AEJ was also supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number R25MD011564. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104421.

Reference

[1] N. Dilekli, S.V. Gopalani, J.E. Campbell, A.E. Janitz. Oklahoma concentrations. https://docs.google.com/spreadsheets/d/1mAAd2xxRveKK94FMvTz-rtnWXCN41BMj7d4fL3CBzIE/edit?usp=sharing (accessed 22 July 2019).