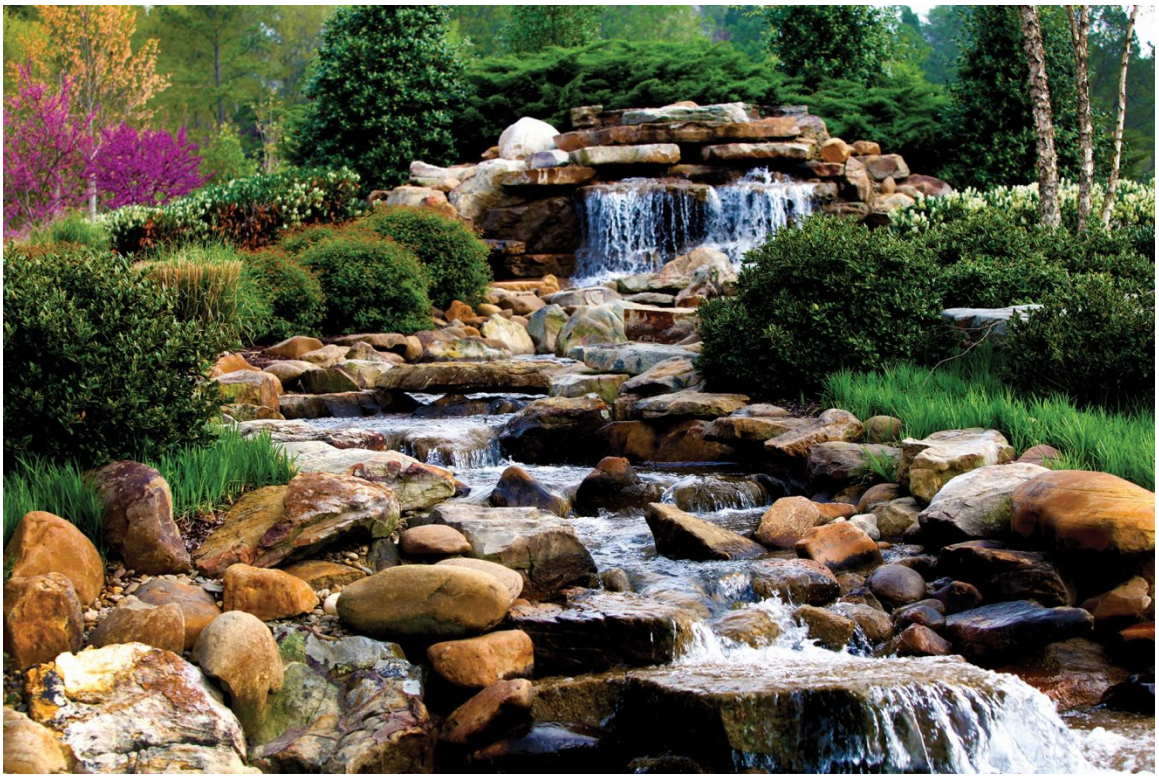
ORNL/SPR-2022/2316
PUBID 172755

VA EDH Data Curation Documentation FY22-Q1, Rev. 2.



Blair Christian
Hilda B Klasky
Kevin Sparks
Alina Peluso
Joe Tuccillo
Pravallika Devineni
Rochelle Watson

December 2021

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/SPR-2022/2316

Computational Sciences & Engineering Division

VA EDH DATA CURATION DOCUMENTATION FY22 - Q1, REV. 2.

Blair Christian
Hilda Klasky
Kevin Sparks
Alina Peluso
Joe Tuccillo
Pravallika Devineni
Rochelle Watson

December 2021

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CHANGE HISTORY

Rev No.	Change	Performed by:	Date
2	1. A co-author's last name has been corrected from 'Brandstetter' to 'Branstetter'.	Hilda B. Klasky	July, 2022

CONTENTS

TABLES

1. INTRODUCTION

The health and well-being of the Nation's men and women who have served in uniform is the highest priority for the U.S. Department of Veterans Affairs (VA). VA is committed to providing timely access to high-quality, recovery-oriented, evidence-based mental health care that anticipates and responds to Veterans' needs and supports the reintegration of returning Service members into their communities. VA is working to eliminate suicide among all Veterans by developing and implementing innovative suicide prevention approaches and resources.

Health outcomes, such as suicide are typically modeled as a function of genetics and environment, where environment refers to factors beyond medical, e.g., air quality, access to transportation and food, homelessness status, etc. Mental health outcomes for each individual are considered to be associated with multiple stressors that fall under a variety of categories – socioeconomic, economic, physical environment. Understanding the relationships between these stressors, covariates and health outcomes, requires curated, standardized data that can be input into the VA's Recovery Engagement and Coordination for Health – Veterans Enhanced Treatment (REACH VET) or other health outcomes model. Environmental Determinants of Health (EDH) as defined by the World Health Organization (WHO) is clean air, stable climate, adequate water, sanitation and hygiene, safe use of chemicals, protection from radiation, healthy and safe workplaces, sound agricultural practices, health-supportive cities and built environments, and a preserved nature are all prerequisites for good health.

1.1 BACKGROUND

With funding from the VA Office of Mental Health and Suicide Prevention (OMHSP), the EDH project has developed novel datasets associated with select health outcomes, a methodology for converting spatiotemporal data from one spatial reference (such as a 1km grid) to another (such as US Census Tracts), and health outcomes modeling capabilities. The datasets are an advancement to the AHRQ SDoH covariates as key gaps are addressed, a finer spatial resolution (Census Tract), and environmental covariates are included.

The process of curating and standardizing these datasets is non-trivial, as they are often measured at different spatial and temporal resolutions and have different spatial and temporal granularities. For example, the US Census data products typically use census blocks, block groups, or counties, whereas air pollutants from the EPA and weather data are available on 1km grids, and some economic data may be available only at a zip code level. In this context, standardized refers to the datasets all being at the same scale of spatial extent (e.g., US Census Tract and/or County), and curated refers both to a process that is repeatable, has data provenance, and which uses appropriate methodologies for converting covariates. The data contained in the EDH datasets are drawn from multiple sources, and variables may have differing degrees of availability, patterns of missing data, and methodological considerations across sources, geographies, and years.

2. DOCUMENTATION OVERVIEW

2.1 DATA SOURCE DOCUMENTATION REPORT

This data source documentation report contains information for researchers about the structure and contents of the datasets as well as descriptions for each of the data sources used to populate the data files.

This document contains dataset curation documentation updates for FY22 Q1. For the dataset curation documentation of FY21, please see the following source:

> Christian, J.B., Branstetter, M, Klasky, H.B., Tuccillo, J., Sparks, K., Rastogi, D., Watson, R., Yoon, H-J., Kim, Y., VA EDH Data Curation Documentation - FY 2021, ORNL/SPR-2021/2366 - Pub ID 170648. 2021.

2.2 RECOMMENDED CITATION FOR THIS DATASET PACKAGE

> Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Devineni, P., VA EDH Data Curation Documentation – FY22Q1, ORNL/SPR-2022/2316- Pub ID 172755. 2022.

2.3 CONTENTS OF THE EDH DATA CURATION DOCUMENTATION

The data presented in the EDH datasets were derived from four publicly available data sources. In this release the data sources included are presented in Table 1. In Table 1, the types are: ED = Economic Distress, SCC = Social Capital and Connectedness, LMA = Lethal Means Access, HQA = Healthcare Quality and Access.

Table 1. EDH FY22-Q1 Datasets.

Type	Dataset	Years	Source
SCC	Social Capital Index Dataset	2019 - Updated	US Census Bureau, MIT Election Lab, National center for Charitable Statistics
ED	Social Vulnerability Index Dataset	2018	US Centers for Decease Control/ Agency for Toxic Substances and Disease Registry
ED	Block Group Area Deprivation Index Dataset for Washington,	2019	Neighborhood Atlas, University of Wisconsin, Department of Medicine.
ED	Low Food Access Area Dataset for Washington, DC	2017	OPEN DATA DC

More detailed information about each data source is included in the following sections of this report.

2.4 DATASET OVERVIEWS

Variables in the EDH Dataset were created from these four data sources in one of two ways:

1. Drawn directly from the original data source. When the data were available from the data source as needed, we renamed the original variables for clarity and consistency across years, and to fit the naming conventions of the SDOH beta data files.
2. Derived using data from the original data source. For some data sources, it was necessary to calculate percentages or rates for inclusion in the beta data files. The numerators and denominators for the variables and their sources are shown following each data source description.

The following conventions were followed in constructing the EDH Datasets to provide researchers with a consistent and easy-to-use resource:

- **Variable assignment to annual datasets.** Variables appear in the annual datasets that correspond with (1) the single year represented by the original data source (e.g., Nursing Home Compare data for facilities in 2016 appears in the 2016 county dataset), or (2) the last year in a period represented by the data (e.g., American Community Survey data aggregated over 2012 to 2016 is in the 2016 dataset).
- **Variable availability.** The availability of each variable changes across data years. Following each data source description in this report is a table showing the availability of each variable in the annual datasets.
- **Variable naming.** Except for the geographic ID variables, all variable names begin with a data source acronym followed by an underscore and a descriptive title.
- **Missing values.** The datasets use a blank to denote a missing value, almost exclusively. The one exception is the provider ratio variables from the County Health Rankings (CHR) data, which have negative values for counties where the number of providers is zero. This is described further in the description of the CHR data.'

3. CONTENTS AND STRUCTURE

Each data source description follows a standard format with the following fields:

- Sponsor (name of the organization that provided the raw data, e.g., Health Resources and Services Administration [HRSA] for the Area Health Resources Files [AHRF])
- Description (brief, general description of the data)
- Inclusion in the EDH datasets
  - Lists the SDOH domains to which the data source has contributed variables
  - Includes additional information about the data source relevant to the EDH dataset
- Resources (links to original data source documentation, data download sites, and other relevant information)
- Update frequency: how often is each dataset going to be updated.
- Resources: the URLS where the data was collected.
- EDH variable definitions and specifications (in tabular format)
  - Variable name
  - Variable label
  - Source table, if multiple data tables were available from the original data source
  - Numerator (for derived variables)
  - Denominator (for derived variables) or original variable (when renamed for the EDH dataset)
- EDH variable availability across years (in tabular format)
  - Variable name
  - Variable label
  - Data year availability (e.g. 2009 to 2018)

4. SOCIAL CAPITAL INDEX

4.1 SPONSOR

United States Department of Veterans Affairs

4.2 DESCRIPTION

A Social Capital Index for 2019 based on [Rupasingha at el 2006 (https://www.sciencedirect.com/science/arti cle/abs/pii/S1053535705000971)] (https://www.sciencedirect.com/science/article/abs/pii/S1053535705000 971)), and an update to the Social Capital Index for the years 1997, 2005, 2009, and 2014. Social capital has had a powerful impact on the study of politics, policy, and social science at large. While the concept of social capital is valid universally, the measure of social capital varies by context. Much of what we know about the causes and effects of social capital, however, is limited by the nature of data used regularly by scholars working in this area. Principal Component Analysis is used to extract principal components from data variables and create a signal index that indicates the social capital. Data are used that represents relevant establishments, voter turnout, census response rates, and non-profit organizations. The implementation presented various challenges including missing and suppressed data, and changing county names.

4.3 INCLUSION

Geographic Unit: county level (county FIPS codes) for the 48 continental US states plus Alaska and Hawaii. Index value: A single social capital index value for each county.

Input Data: Four factors used for the computation of the 2019 index.

1) Establishments per 10,000 population,
2) Voter turnout,
3) Census response rate,
4) Non-profit organizations per 10,000 population.

Method: The social capital index is created using principal component analysis using the above four factors. The four factors are standardized to have a mean of zero and a standard deviation of one, and the first principal component is considered as the index of social capital.

4.4 UPDATE FREQUENCY

Updates will be made once a year for the duration of the project or as requested by the sponsor.

4.5 RESOURCES

For more information on the Social Capital Index, the inspiration data products are archived here (https://aese.psu.edu/nercrd/community/social-capital-resources).

The primary data sources used in the calculations for 2019 are as follows:

1) **Establishments**: County Business Patterns (https://www.census.gov/programs-surveys/cbp.html)
2) **Population**: US Census, Population and Housing Unit Estimates (https://www.census.gov/programs- surveys/popest.html)
3) **Voter Turnout:** MIT Election Lab (https://electionlab.mit.edu/)
4) **Census response rate**: US Census 2020 (https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html)
5) **Non-profit**: National Center for Charitable Statistics (http://www.nccs.urban.org/)

Table 2. Social Capital Index (SOCAP)

variable name	variable label

FIPS	County FIPS code
sci	Social Capital index for 2019
civic	Number of establishments in civic and social associations
bowling	Number of establishments in bowling center
fitness	Number of establishments in fitness and recreational sports centers
golf	Number of establishments in golf courses and country clubs
religion	Number of establishments in religious organizations
sport	Number of establishments in sports teams and clubs
business	Number of establishments in business associations
political	Number of establishments in political organizations
professional	Number of establishments in professional organizations
labor	Number of establishments in labor organization
associations	Average of all 10 above variables divided by population per 10,000 (1st factor)
vote	Voter turnout (2nd factor)
response	Response rate (3rd factor)
nccs	Number of non-profit organizations divided by population per 10,000 (4th factor)
population	Population estimate
Putnam	Average of civic, bowling, fitness, golf, religion, and sport, divided by population per 10,000
Olson	Average of business, political, professional, and labor, divided by population per 10,000

Table 3 Variable availability across years, SOCAP

variable name	2019
FIPS	X
sci	X
civic	X
bowling	X
fitness	X
golf	X
religion	X
sport	X
business	X
political	X
professional	X
labor	X
associations	X
vote	X
response	X
nccs	X
population	X
Putnam	X

Olson	X

5. CDC/ATSDR SOCIAL VULNERABILITY INDEX DATA

5.1 SPONSOR

Centers for Decease Control / Agency for Toxic Substances and Disease Registry

5.2 DESCRIPTION

Social Vulnerability Index (SVI) Data

5.3 INCLUSION

- **Year**: 2018
- **Geographies**: Washington DC, Maryland, Pennsylvania, Virginia
- **Geography type**: census tract

5.4 UPDATE FREQUENCY

Updates will be made once a year for the duration of the project or as requested by the sponsor.

5.5 RESOURCES

For more information on the SVI data:
Source: https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
Documentation: https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html

Table 4. CDC/ATSDR Social Vulnerability Index Data (SVI)

variable name	variable label
FIPS	FIPS
RPL_THEME1	Percentile ranking for Socioeconomic theme summary
RPL_THEME2	Percentile ranking for Household Composition theme summary
RPL_THEME3	Percentile ranking for Minority Status/Language theme
RPL_THEME4	Percentile ranking for Housing Type
RPL_THEMES	Overall percentile ranking

Table 5. Variable availability across years, SVI

variable name	2018
FIPS	X
RPL_THEME1	X
RPL_THEME2	X
RPL_THEME3	X
RPL_THEME4	X
RPL_THEMES	X

6. WASHINGTON DC, 2019 BLOCK GROUP AREA DEPRIVATION INDEX FILES V. 3.0

6.1 SPONSOR

Neighborhood Atlas, University of Wisconsin, Department of Medicine.

6.2 DESCRIPTION

The Area Deprivation Index (ADI) is based on a measure created by the Health Resources & Services Administration (HRSA) over three decades ago, and has since been refined, adapted, and validated to the Census Block Group neighborhood level by Amy Kind, MD, PhD and her research team at the University of Wisconsin-Madison. It allows for rankings of neighborhoods by socioeconomic disadvantage in a region of interest (e.g. at the state or national level). It includes factors for the theoretical domains of income, education, employment, and housing quality. It can be used to inform health delivery and policy, especially for the most disadvantaged neighborhood groups. Note that neighborhood is defined as a Census Block Group.

6.3 INCLUSION

2019 data.

6.4 UPDATE FREQUENCY

Updates will be made once a year for the duration of the project or as requested by the sponsor.

6.5 RESOURCES

https://www.neighborhoodatlas.medicine.wisc.edu/

Table 6 Washington DC, 2019 Block Group Area Deprivation Index Files v 3.0 (ADI)

variable name	variable label
FIPS	The block group Census ID
ADI_NATRANK	National percentile of block group ADI score
ADI_STATERNK	State-specific decile of block group ADI score

Table 7 Variable availability across years, ADI

variable name	2019
FIPS	X
ADI_NATRANK	X
ADI_STATERNK	X

7. LOW FOOD ACCESS AREAS OF THE DISTRICT OF COLUMBIA, WASHINGTONDC

7.1 SPONSOR

OPEN DATA DC

7.2 DESCRIPTION

[Summary taken from Open Data DC] [https://opendata.dc.gov/datasets/DCGIS::low-food-access-areas/about]
Summary Low food access areas of the District of Columbia which are estimated to be more than a 10-minute walk from the nearest full-service grocery store.
Polygons in this layer represent low food access areas of the District of Columbia which are estimated to be more than a 10-minute walk from the nearest full-service grocery store. These have been merged with Census poverty data to estimate how much of the population within these areas is food insecure (below 185% of the federal poverty line in addition to living in a low food access area). Office of Planning GIS followed several steps to create this layer, including: transit analysis, to eliminate areas of the District within a 10-minute walk of a grocery store; non-residential analysis, to eliminate areas of the District which do not contain residents and cannot classify as low food access areas (such as parks and the National Mall); and Census tract division, to estimate population and poverty rates within the newly created polygon boundaries.

Note that the polygon representing Joint Base Anacostia-Bolling was removed from this analysis. While technically classifying as a low food access area based on the OP Grocery Stores layer (since the JBAB Commissary, which only serves military members, is not included in that layer), it is recognized that those who do live on the base have access to the commissary for grocery needs.

We intersected unique OPEN DATA DC polygons with census block group administrative boundaries for engineering and data consistency purposes. Currently, we are not delivering custom geometry data, and so linking the OPEN DATA DC custom geometry polygons to census administrative boundaries was a required step that still preserves the general structure of the data.

7.3 INCLUSION

Downloaded from Open Data DC in Dec 2021. Last updated November 2017

7.4 UPDATE FREQUENCY

Updates will be made once a year for the duration of the project or as requested by the sponsor.

7.5 RESOURCES

For more information on the Low food access areas of the District of Columbia:Data:
[https://opendata.dc.gov/datasets/DCGIS::low-food-access-areas/about]

Table 8 Low food access areas of the District of Columbia, Washington DC (LFAA)

variable name	variable label
FIPS	FIPS
WARD	WARD
PARTPOP2	The total population estimated to live within the low food access area polygon.
PRTOVR185	The portion of PartPop2 which is estimated to have household income above 185% of the federal poverty line (the food secure population)
PRTUND185	The portion of PartPop2 which is estimated to have household income below 185% of the federal poverty line (the food insecure population)
PERCENTUND185	A calculated field showing PrtUnd185 as a percent of PartPop2. This is the percent of the population in the polygon which is food insecure.

Table 9 Variable availability across years, LFAA

variable name	2017
FIPS	X
WARD	X
PARTPOP2	X
PRTOVR185	X
PRTUND185	X
PERCENTUND185	X