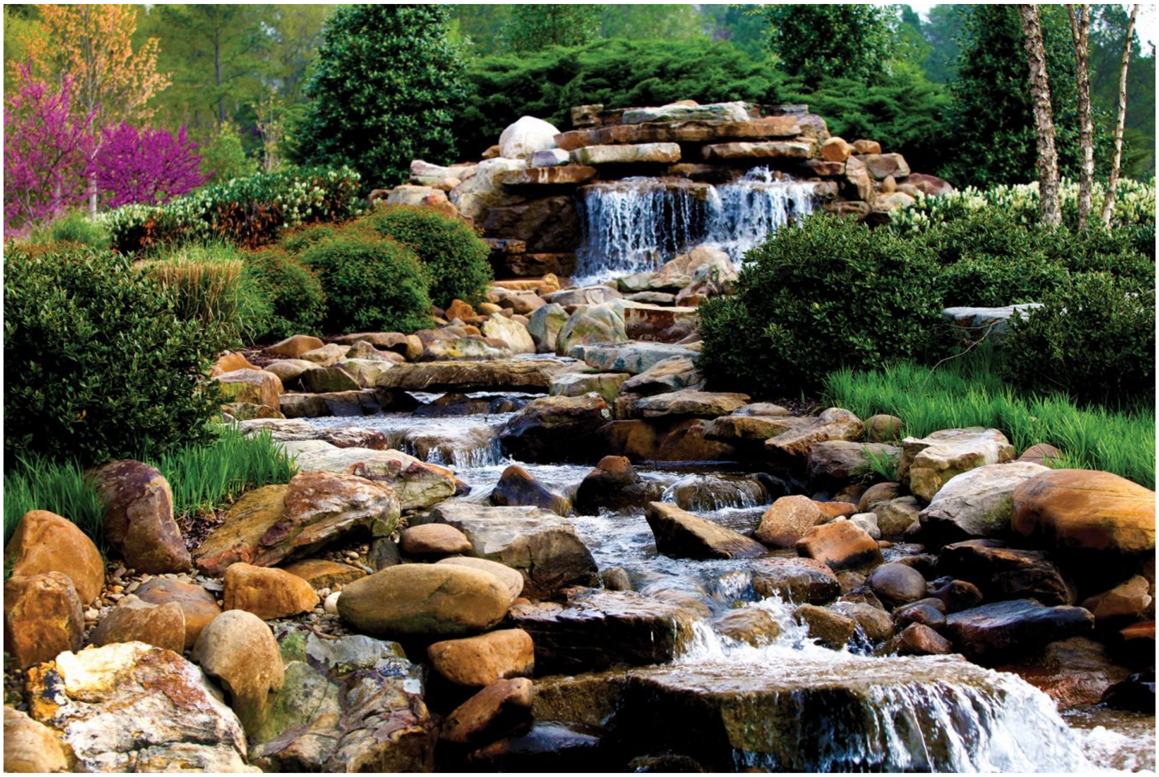
ORNL/SPR-2023/3207
PUB ID 205615

VA EDH Data Curation Documentation FY24-Q1



Hilda Klasky
Kevin Sparks
Alina Peluso
Aaron Myers
Jeremy Logan
Michael McGee
Alec Hamaker
Jonathan Zhang
Heidi Hanson
Rochelle Watson
Anuj Kapadia

December 2023

OAK RIDGE
National Laboratory

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

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/SPR-2023/3207
PUB ID 205615

Computational Sciences & Engineering Division

VA EDH DATA CURATION DOCUMENTATION FY24-Q1

Hilda B. Klasky
Kevin Sparks
Alina Peluso
Aaron Myers
Jeremy Logan
Michael McGee
Alec Hamaker
Jonathan Zhang
Heidi Hanson
Rochelle Watson
Anuj Kapadia

December 2023

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

CONTENTS

1. INTRODUCTION

The U.S. Department of Veterans Affairs (VA) places the health and well-being of our nation's veterans as its top priority. VA is dedicated to offering timely access to high-quality, evidence-based mental health care that meets the needs of veterans and supports their reintegration into society. One of our core missions is to prevent suicide among veterans through innovative approaches and resources.

Health outcomes, including suicide, are typically influenced by both genetics and environmental factors, such as air quality, transportation access, food availability, homelessness, and more. Mental health outcomes are associated with various stressors across socioeconomic, economic, and physical environments. Analyzing the connections between these stressors, covariates, and health outcomes relies on standardized data, which can be integrated into models like the VA's Recovery Engagement and Coordination for Health, Veterans Enhanced Treatment (REACH VET).

The World Health Organization (WHO) defines Environmental Determinants of Health (EDH) as factors like clean air, stable climate, water and sanitation, chemical safety, radiation protection, safe workplaces, sustainable agriculture, healthy urban environments, and nature preservation, all of which are crucial for good health.

1.1 BACKGROUND

With funding from the VA Office of Mental Health and Suicide Prevention (OMHSP), the EDH project has developed innovative datasets associated with specific health outcomes, a methodology for transforming spatiotemporal data from one spatial reference (e.g., a 1km grid) to another (e.g., US Census Tracts), and capabilities for modeling health outcomes. These datasets represent an enhancement of the Agency for Healthcare Research and Quality (AHRQ) Social Determinants of Health (SDoH) covariates, addressing key gaps by introducing finer spatial resolution (Census Tract) and additional environmental covariates.

The curation and standardization of these datasets is a complex task since they often originate from various sources and are measured at different spatial and temporal resolutions. For example, US Census data products typically use census blocks, block groups, or counties, while data like air pollutants from the US Environmental Protection Agency (EPA) and weather data are available on 1km grids. Some economic data may only be available at the zip code level. In this context, 'standardized' means that all datasets share the same spatial extent (e.g., US Census Tract and/or County), and 'curated' implies a repeatable process with data provenance and the use of appropriate methodologies for covariate conversion.

The EDH datasets draw from multiple sources, resulting in variables with varying degrees of availability, patterns of missing data, and methodological considerations across different sources, geographies, and years.

2. DOCUMENTATION OVERVIEW

This data source documentation report is designed to provide researchers with valuable insights into the structure, contents, and the data sources utilized to compile the datasets. It specifically covers the Fiscal Year 2024, First Quarter (FY24-Q1) dataset curation documentation for the Environmental Determinants of Health (EDH) project.

The datasets included in this documentation are as follows:

1. **ORNL Daily Surface Weather and Climatological Summaries - Daymet, 2017-2021, by county (new)**

2. **Veterans Service Organizations (VSO) 2010-2022, by state (new)**
3. **Veterans Service Organizations (VSO) 2010-2022, by county (new)**
4. **Veterans Service Organizations (VSO) 2010-2022, by zip code (new)**

2.1 RECOMMENDED CITATION FOR FY24-Q1 DATA CURATION DOCUMENTATION'S SPONSOR REPORT

The recommended citation for the FY24-Q1 data curation documentation's sponsor report follows:

> Klasky, H.B., Sparks, K., Peluso, A., K., Myers, A., Logan, J., McGee, M., Hamaker, A., Zhang, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY24-Q1, ORNL/SPR-2023/3207 PUB ID 205615, 2023.

This documentation provides a comprehensive understanding of the data and its sources for the specified period, supporting research and analysis within the EDH project.

2.2 PREVIOUS DOCUMENT RELEASES

Since the inception of the EDH project, we have delivered multiple releases of datasets along with data curation documentation sponsor reports. These resources are invaluable for researchers seeking to utilize the EDH data. Below is a list of the previous releases:

1. EDH Data Curation Documentation delivered in FY21 [1]

- Link to Documentation

2. EDH Data Curation Documentation delivered in FY22-Q1 [2]

- Link to Documentation
- Included Datasets:
  - Social Capital Index (*resolution:* county, 2019, *source:* ORNL)
  - Social Vulnerability Index (*resolution:* census tract, 2018, *source:* Centers for Disease Control, Agency for Toxic Substances and Disease Registry)
  - Area Deprivation Index (*resolution:* block group, 2019, *source:* Neighborhood Atlas, University of Wisconsin)
  - Low Food Access (*resolution:* custom geometry, 2017, *source:* Open Data DC)

3. EDH Data Curation Documentation delivered in FY22-Q2 [3]

- Link to Documentation
- Included Datasets:
  - Eviction Rates (*resolution:* county, 2000-2016, *source:* Eviction Lab)
  - Income Inequality (*resolution:* block group, 2019, *source:* American Community Survey)
  - Individual-Oriented Social Vulnerability Index (*alternate name:* IOSVI, *resolution:* block group, 2019, *source:* ORNL, Census Bureau)
  - National Instant Criminal Background Check System (*alternate name:* NICS, *resolution:* state, 2022, *source:* Federal Bureau of Investigation)

4. EDH Data Curation Documentation delivered in FY22-Q3 [4]

- Link to Documentation
- Included Datasets:
  - Veteran Population Status (*resolution:* county, 2020, *source:* American Community Survey)

  - Social Connectedness (*resolution:* county, 2021, *source:* Facebook)
  - Small Area Estimates of Housing Characteristics (*resolution:* block group, 2019, *source:* Census Bureau)
  - Internet Access Services (*resolution:* tract, 2019, *source:* Federal Communications Commission)
  - Medicare Part D Opioid Prescription Rates (*resolution:* county, 2019, *source:* Centers for Medicare & Medicaid Services)
  - High Intensity Drug Trafficking Areas (*alternate name:* HIDTA, *resolution:* county, 2018-21, *source:* Washington/Baltimore High Intensity Drug Trafficking Areas Program)

5. EDH Data Curation Documentation delivered in FY22-Q4 [5]

- Link to Documentation
- Included Datasets:
  - Occupational Employment and Wage Statistics (*alternate name:* Mental Health Care Professionals per capita, *resolution:* state, 2021, *source:* Bureau of Labor Statistics)
  - National Survey on Drug Use and Health (*alternate name:* NSDUH, *resolution:* state, 2019, *source:* Substance Abuse and Mental Health Services Administration)
  - National Mental Health Services Survey (*alternate name:* N-MHSS, *resolution:* state, 2018, *source:* Substance Abuse and Mental Health Data Archive)

6. EDH Data Curation Documentation delivered in FY23-Q1 [6]

- Link to Documentation
- Included Datasets:
  - State and Local Policies (Naloxone laws, *resolution:* state, 2017, *source:* Rand) (Good Samaritan laws, *resolution:* state, 2018, *source:* Rand)
  - Area Deprivation Index (*resolution:* block group, 2020, *source:* University of Wisconsin)
  - Opioid Mortality Rate (*resolution:* county, 2014-2018, *source:* OEPS, University of Chicago)
  - Opioid Prescribing Rate (*resolution:* county, 2019, *source:* OEPS, University of Chicago)

7. EDH Data Curation Documentation delivered in FY23-Q2 [7]

- Link to Documentation
- Included Datasets:
  - Total Household Income (*resolution:* county, 2016-2021, *source:* American Community Survey)
  - Medicare Part D Opioid Prescription Rates (update, *resolution:* county, 2013-2020, *source:* Centers for Medicare & Medicaid Services)
  - Poverty (*resolution:* county, 2016-2021, *source:* American Community Survey)
  - Rural Urban Continuum Codes (*resolution:* county, 2013, *source:* Census Bureau, Department of Agriculture)
  - Social Capital Atlas - Civil Engagement (*resolution:* county, 2022, *source:* Social Capital Atlas)
  - Social Capital Atlas - Cohesiveness (*resolution:* county, 2022, *source:* Social Capital Atlas)
  - Social Capital Atlas - Economic Connectedness (*resolution:* county, 2022, *source:* Social Capital Atlas)
  - Local Unemployment (*resolution:* county, 2018-2021, *source:* Bureau of Labor Statistics)

8. EDH Data Curation Documentation delivered in FY23-Q3 [8]

- Link to Documentation
- Included Datasets:
  - Population Weighted Average Elevation (*resolution:* county, 2020, *source:* United States Geological Survey, Jim VanDerslice)
  - Education Attainment (*resolution:* county, 2016-2021, *source:* US Census Bureau, American Community Survey)
  - Eviction Rates (update, *resolution:* county, 2016-2021, *source:* The Eviction Lab, Princeton University)
  - Food Insecurity (*resolution:* county, 2010-2021, *source:* Feeding America, US Hunger Relief Organization)

9. EDH Data Curation Documentation delivered in FY23-Q4 [9]

- Link to Documentation
- Included Datasets:
  - National Instant Criminal Background Check System (NICS, *resolution:* state, 2021-2023, *source:* US Federal Bureau of Investigation)
  - Internet Access Services (*resolution:* Census tract, 2021-2022, *source:* US Federal Communications Commission (FCC))

Please note that the URL for the FY24-Q1 documentation's URL will be provided next delivery.

This comprehensive list allows researchers to access previous releases for reference and analysis, enhancing the utility of the EDH project's data curation documentation.

3. CONTENTS AND STRUCTURE

3.1 DATASET CURATION DOCUMENTATION STANDARD FORMAT

Each data source description adheres to a standardized format with the following fields:

1. **Source**: The name of the organization that provided the raw data (e.g., Health Resources and Services Administration [HRSA] for the Area Health Resources Files [AHRF]). Note: Prior to the FY23Q4 release, we referred to the source organization as the "sponsor."
2. **Description**: A brief, general description of the data.
   - *Inclusion in the EDH datasets*: Lists the social or environmental determinants of health domains to which the data source has contributed variables. Includes additional information relevant to the EDH dataset.
3. **Resources**: Links to original data source documentation, data download sites, and other pertinent information.
4. **Update Frequency**: Indicates how often each dataset will be updated.
5. **Variable Definitions and Specifications (in tabular format)**:
   - *Variable name (column name)*
   - *Variable label (optional, if different from the variable or column name)*
   - *Source table (optional, if multiple data tables were available from the original data source)*
   - *Numerator (for derived variables; optional)*
   - *Denominator (for derived variables) or original variable (when renamed for the EDH dataset; optional)*

- *Total_rows*: Indicates the number of rows in each column within each dataset (Starting in FY23Q2).
- *Null_rows*: Specifies the count of null rows for each column in each dataset (Starting in FY23Q2).

6. **Variable Availability Across Years (in tabular format)**:
   - *Variable name (column name)*
   - *Data year availability (e.g., 2009 to 2018)*

This standardized format ensures consistency and ease of reference in the curation documentation for each data source.

3.2 DATASET CONVENTIONS

In terms of dataset versioning, we utilize the Microsoft SQL Server database system to provide these datasets. Each dataset is stored in a dedicated table within the database. The quarterly releases are organized under distinct schema names within the database, such as OMHSP_FY22Q1, OMHSP_FY22Q2, OMHSP_FY22Q3, OMHSP_FY22Q4, OMHSP_FY23Q1, and so forth. These schema names facilitate distinguishing between releases when we deliver the same dataset, albeit updated, from one release to the next.

The variables within the EDH dataset are derived from various data sources through one of two methods:

I. Direct extraction from the original data source: When the data was readily available from the source, we renamed the original variables to ensure clarity and consistency across years, aligning them with the naming conventions of the SEDH data files.
II. Derivation using data from the original data source: In certain cases, we needed to calculate percentages or rates for inclusion in the data files. We provide the numerators and denominators for these variables, along with their respective sources, in the data source descriptions.

To ensure the SEDH datasets serve as a consistent and user-friendly resource for researchers, we adhered to the following conventions:

- **Variable assignment to annual datasets:** Variables appear in the annual datasets corresponding to (1) the single year represented by the original data source (e.g., US Area Deprivation Index 2020) or (2) the final year in a period represented by the data (e.g., American Community Survey data aggregated over 2012 to 2016 is included in the 2016 dataset).
- **Variable availability:** Variable availability varies across data years. Following each data source description in this report, you will find a table that outlines the availability of each variable in the annual datasets. When a variable is not available, we indicate it with ‘NA’ (not available) or simply ‘-’.
- **Variable naming:** With the exception of geographic ID variables, all variable names begin with a data source acronym, followed by an underscore and a descriptive title.
- **Missing values:** In the datasets, we use a blank to denote missing values, with one exception being the provider ratio variables from the County Health Rankings (CHR) data. These have negative values for counties where the number of providers is zero, a detail further explained in the CHR data description.

For comprehensive information about each data source, please refer to the subsequent sections of this report.

3.3 METADATA TABLE

Starting from FY23Q1, the ORNL team provides an updated metadata table, known as SEDH_meta_table, located in the OMHSP schema. SEDH stands for the Social and Environmental Determinants of Health repository. This table contains the following columns:

- **schema**: Quarterly release schema names in the database (e.g., OMHSP_FY22Q4, OMHSP_FY23Q1, and so on).
- **table_name**: The table name as it appears in the MS SQL Server database.
- **table_name_description**: A description of the table name.
- **column_name**: Column names within each dataset as they appear in the MS SQL Server table.
- **column_name_description**: Descriptions of each column name.
- **availability_across_years**: The years for which data is available.
- **reference_report**: A reference to the ORNL report containing data curation documentation.
- **report_url**: URL link to the ORNL report.
- **column_type**: The column type in the MS SQL Server table.
- **column_length**: The column length in the MS SQL Server table.
- **total_rows**: The number of rows in each column in each dataset (starting in FY23Q2).
- **null_rows**: The number of null rows for each column in each dataset (starting in FY23Q2).
- **data_source**: The name of the source organization that provided the raw data (starting in FY23Q4).
- **data_source_description**: Description of the source organization (starting in FY23Q4).
- **data_source_url**: URL of the source organization (starting in FY23Q4).
- **data_categories**: General data categories, such as social, economic, educational, etc. (starting in FY23Q4).
- **spatial_resolution**: Spatial resolution or geography (e.g., state, county, block group, tract, etc.) (starting in FY23Q4).

With each new quarterly release, the metadata table will be updated with new information in the aforementioned columns for each delivered dataset.

Please note that the report_url column will be updated in the VA's CDW transmit database as soon as it becomes available on the Office of Scientific and Technical Information website (osti.gov) of the US Department of Energy, typically four weeks after each quarterly release.

3.4 FIPS AS GEOGRAPHIC IDENTIFIERS

At ORNL, we utilize the Federal Information Processing Standards (FIPS) as geographic identifiers and primary keys in each dataset or table for this project. FIPS codes are publicly recognized standards developed by the National Institute of Standards and Technology (NIST) for computer systems and non-military applications, particularly for standardizing codes of geographical areas. FIPS specifications encompass various geographical areas:

- FIPS 10-4 for country and region codes
- FIPS 5-2 for state codes
- FIPS 6-4 for county codes

These codes are unique within their respective geographic entities. For example, FIPS state codes are unique within a country, and FIPS county codes are unique within a state. Since counties nest within states, a complete county FIPS code combines the state and county identifiers. For instance, if multiple counties end with "001," the state FIPS code is added to make each county FIPS code distinct (e.g., 01001, 02001, 04001), where the first two digits indicate the state, and the last three digits represent the county.

Although NIST initiated the replacement of FIPS with the Geographical Name Information System (GNIS) Feature ID in 2002, many federal organizations in the United States, including the US Census Bureau, continued to use FIPS due to its broader coverage and precision in identifying geographic entities, especially smaller areas with uncertain natural boundaries. The US Census Bureau maintains a comprehensive hierarchy of census geographic entities for reference.

As the primary key in all datasets for this project, we consistently use the column "FIPS" to ensure unique data identification, regardless of the source FIPS granularity. We specify the FIPS granularity, such as region, state, county, census division, tracks, group blocks, etc., in the metadata table and reports' descriptions. Users are presumed to be familiar with joining datasets using FIPS columns at different geographic levels.

It's worth noting that only two datasets since the inception of this project do not include a FIPS column.

These exceptions are:

5. The National Mental Health Services Survey (table: national_mental_health_services_survey), received in FY22Q4.
6. The Veterans Service Organizations (VSO) 2010-2022, by zip code, delivered in FY24Q1.

These datasets were provided upon special request from the sponsor (table: va_vso_zipcode_2010_2022).

3.5 MAPPING ZIP CODES TO FIPS CODES FOR COUNTIES: OUR METHODS

When realigning spatial data to different boundaries that do not perfectly match or nest within the original spatial units, some data loss is inevitable. This occurs because the spatial distribution of data at higher resolutions than the native unit is often unknown. For example, certain zip code boundaries overlap with multiple county boundaries. When attempting to map zip code-level data to counties, there are situations where data must be reassigned to two or more counties with limited knowledge of how to allocate it accurately. Various methods exist to mitigate the degree of data loss, each with its strengths and weaknesses based on the data's nature. For social data, one effective approach is to allocate data based on population distribution or addresses within those boundaries to reduce misallocation.

Visual examples are provided below to illustrate this challenge:

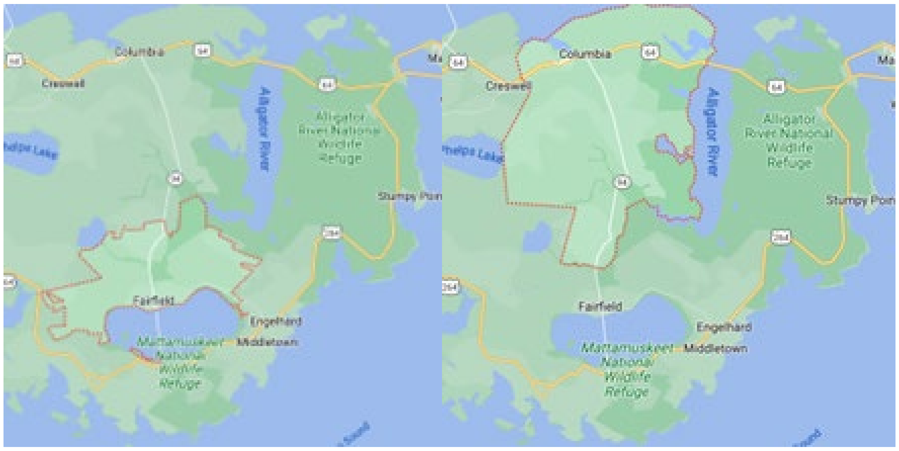


Figure 1 ***Example*****: -** ***Left*****: Zip Code 27826 -** ***Right*****: County FIPS 37177**

In the United States, approximately 40 percent of zip codes may encounter the described conflict. To address these concerns, we have made the following decisions:

a) **Veterans Service Organizations Datasets**: To tackle the issue of mapping zip code data to counties, we utilize crosswalk tables provided by the US Housing and Urban Development's Office of Policy Development and Research. These crosswalk tables consider factors such as residential and commercial address counts in allocating data. Given the dynamic nature of zip codes and county boundaries over time, we use yearly specific crosswalks to ensure accurate data mapping. For more detailed documentation, please refer to this link.

b) **Daymet Dataset**: As the Daymet dataset primarily comprises grid cells and deals with environmental, rather than social, data, we employ a different approach to address the conflict. We use the centroid of each grid cell to map data to different spatial units, such as counties.

We acknowledge that different perspectives may arise regarding the chosen approaches. However, it's important to note that these methods are continually evolving as we develop more precise algorithms for mapping zip codes to counties in the United States.

3.6 ERROR CHECKING

Beginning with the FY23Q1 release, the ORNL team will additionally give succinct information regarding error checking activities in order to provide formal evidence that the datasets supplied have been thoroughly error checked. Our data profiling process is described in our project's overview manuscript [10]:

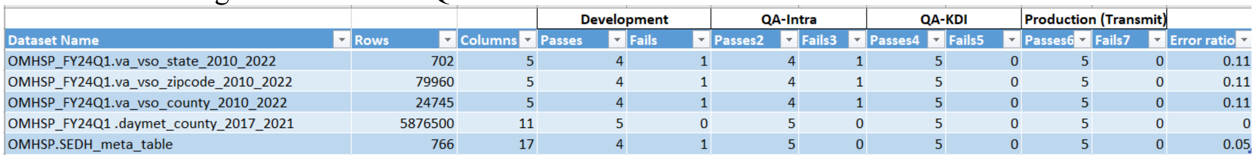
"Following standard data and software development methodologies, data profiling is performed in four different work environments: 1) a team-shared work environment for selection, extraction, and refinement of raw data (development); 2) an ORNL intranet work environment focused on quality assurance testing (QA-Intra); 3) an ORNL Knowledge Discovery Infrastructure (KDI) secure work environment that stores highly sensitive data and ensures its security (QA-KDI). And finally, 4) a production environment housed within the KDI environment and accessible to our VA sponsors, (Production). We carried out test iterations in each of the four work environments as the datasets moved through them to confirm data integrity and system compatibility.

All datasets were error-checked using a data profiling strategy that includes at least two reviewers and the following test groups:

1. evaluating missingness: i.e., determining the amount of missing data by randomly checking for them;
2. compiling descriptive statistics, such as the number of rows, columns, and types of variable data;
3. appending checksums to a subset of the columns on both the source and destination copies to ensure consistency;
4. consistently representing the social and physical environment using FIPS codes as geographic administrative boundaries and confirming that the FIPS codes correspond to the geographic administrative boundaries of the original data;
5. manually comparing the first, last, and five additional randomly selected rows for consistency between the source and target datasets.

When datasets are developed at ORNL, which we call 'derivative', ORNL will provide extra error-checking utilizing a combination of statistical methodologies based on each dataset's properties, in addition to the data profiling methodology described above." [10]

The error-checking results for FY24Q1 follows:



			Development		QA-Intra		QA-KDI		Production (Transmit)		
Dataset Name	Rows	Columns	Passes	Fails	Passes2	Fails3	Passes4	Fails5	Passes6	Fails7	Error ratio
OMHSP_FY24Q1.va_vso_state_2010_2022	702	5	4	1	4	1	5	0	5	0	0.11
OMHSP_FY24Q1.va_vso_zipcode_2010_2022	79960	5	4	1	4	1	5	0	5	0	0.11
OMHSP_FY24Q1.va_vso_county_2010_2022	24745	5	4	1	4	1	5	0	5	0	0.11
OMHSP_FY24Q1 .daymet_county_2017_2021	5876500	11	5	0	5	0	5	0	5	0	0
OMHSP.SEDH_meta_table	766	17	4	1	5	0	5	0	5	0	0.05

3.6.1 Daymet

In addition, we conducted visual error checks on the Daymet dataset. Given the dataset's extensive size (approximately 640,000 individual images), we employed a manageable subset of 50 random samples for this purpose.

The subsequent two images serve as illustrative examples. Each figure displays one variable over one state for a single day, providing a comparison between the original Daymet data and the derived county data.

The first image showcases the Daymet data for New Hampshire on day 5 of 2017, focusing on the variable 'vp,' which represents the water vapor pressure. In this case, our county-based aggregation method exhibits slight discrepancies compared to the original grid. While discernible differences are apparent, the aggregated data still offers a reasonably accurate approximation.

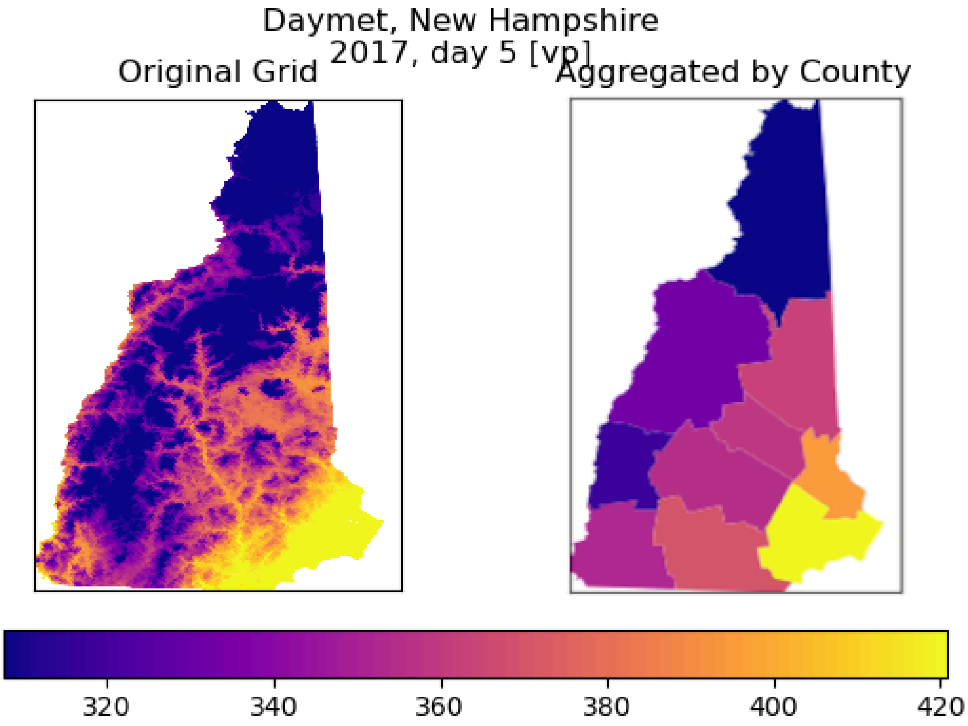


Figure 2. Daymet data for New Hampshire on day 5 of 2017, focusing on the variable 'vp,' which represents the water vapor pressure. The units are pascals (Pa)

In contrast, the second example displays Daymet data for North Carolina on day 264 of 2020, highlighting the 'srad' variable, which represent the radiation of the short wavelength. In this instance, we can observe a striking similarity between the county-based aggregation and the original grid data.

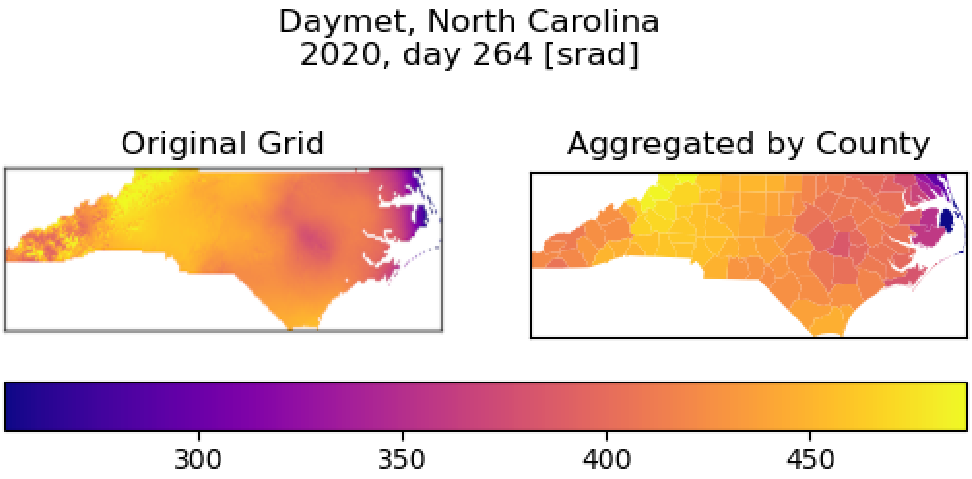


Figure 3. Daymet data for North Carolina on day 264 of 2020, highlighting the 'srad' variable, which represent the radiation of the short wavelength. The units are watts per square meter (W/m2).

3.6.2 VSO

Out of a total of 368,866 records received, a thorough review revealed that 10,780 records, equivalent to 3%, were identified as invalid. The table below provides descriptive statistics related to the number of invalidated records and the types of errors encountered during the data preparation process for the VSO raw data received.

Table: Descriptive Statistics for VSO Data (Years 2010 - 2022)

Summary	
Valid records	358086
Invalid records	10780
Issue type	**Number of records**
Blank rep name	9
Blank state name	0
Blank zip code	3746
Malformed zip code	1377
Unknown zip code	2345
Unknown state/territory abbreviation	138
Bad zip code for state	3071
Error finding county FIPS code	94

This table offers insights into the quality of the VSO data for the specified period, highlighting the extent of invalid records and the associated error types encountered during the data preparation process.

Appendix A presents descriptive statistics of error-checking results.

4. ORNL DAILY SURFACE WEATHER AND CLIMATOLOGICAL SUMMARIES - DAYMET

4.1 DATA SOURCE

Oak Ridge National Laboratory's Environmental Sciences Division in Oak Ridge, Tennessee.

4.2 DESCRIPTION

4.2.1 Daymet: High-Resolution Daily Weather and Climatology Data

Daymet is a research product developed by the Environmental Sciences Division at Oak Ridge National Laboratory in Oak Ridge, Tennessee. Funding for Daymet is provided by NASA through the Earth Science Data and Information System (ESDIS). The ongoing development of the Daymet algorithm and processing is carried out by the Office of Biological and Environmental Research within the U.S. Department of Energy's Office of Science.

4.2.2 What is Daymet?

Daymet leverages statistical modeling approaches to interpolate and extrapolate ground-based data, enabling the generation of long-term, continuous, gridded estimates for various daily weather and climatology variables. These datasets are indispensable for biogeochemical terrestrial modeling and find applications in Earth science, natural resource management, biodiversity studies, and agricultural research.

4.2.3 Key Weather Variables

Daymet provides data for a range of weather variables, including daily minimum and maximum temperature, precipitation, vapor pressure, shortwave radiation, snow water equivalent, and day length. These variables are available at a high spatial resolution covering continental North America, Hawaii, and Puerto Rico, with data extending from 1980 up to the end of the most recent full calendar year.

4.2.4 County-Level Aggregation

While the original dataset records data in 1km x 1km grid cells, it has been aggregated to county-level resolution for ease of use and broader applications.

4.2.5 Calculation of Shortwave Radiation (srad)

For those interested in calculating daily total radiation (MJ/m2/day) from the shortwave radiation variable (srad), the Daymet's formula is as follows:

$$((srad\ (W/m2)\ *\ day\ length\ (s/day))\ /\ 1{,}000{,}000).$$

MJ = Megajoules
m2 = square meter
day = day of the day
srad = shortwave radiation
W = Watts
s = seconds

Daymet's high-quality, high-resolution data is a valuable resource for a wide range of scientific and research endeavors.

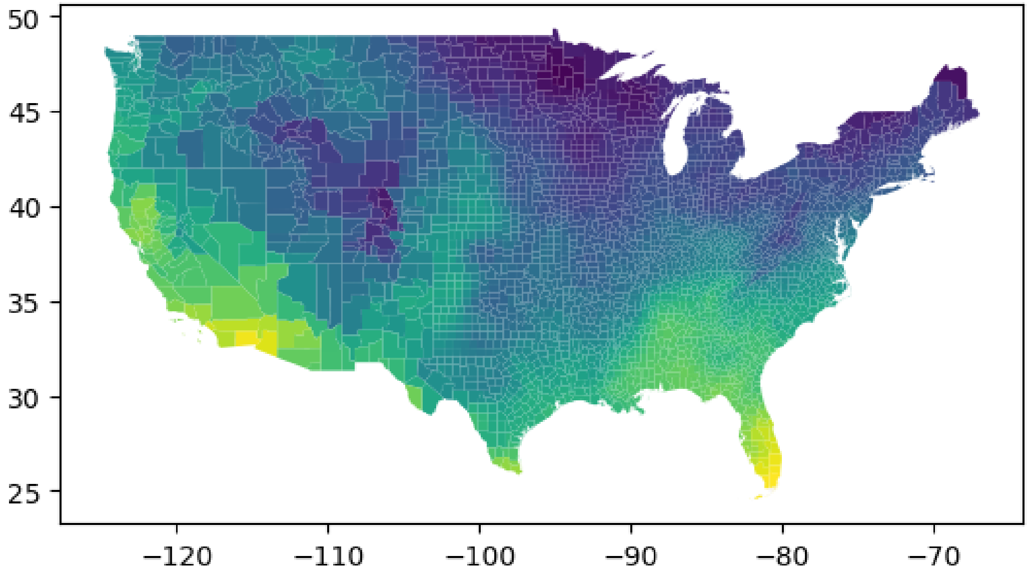


Figure 4. This graph shows the 'tmax' variable, which represents the maximum air temperature on day 30 of the year 2020. The x-axis represents longitude, while the y-axis represents latitude. Darker colors represent a minimum 'tmax' of -32°C, while lighter colors represent a maximum 'tmax' of 30.7°C.

4.3 INCLUSION

Year: 2017, 2018, 2019, 2020, and 2021.

Geographical unit: FIPS at county level, for the Continental US plus Alaska, Hawaii, and Puerto Rico.

4.4 RESOURCES

- Link to ORNL Daymet Home

4.5 UPDATE FREQUENCY

This dataset will be updated and distributed every fiscal year, or as requested by the sponsor. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 1 . ORNL Daily Surface Weather and Climatological Summaries - Daymet (DAYMET_COUNTY)

variable name	variable label
fips	Federal Information Processing Standards (FIPS) - at county level.
year	The year the data pertain to.
day	NA
dayl	The length of a day. The length of daylight in seconds per day. This calculation is based on the time of day when the sun is visible above a hypothetical flat horizon. The units are s/day.
prcp	Precipitation. Total daily precipitation in millimeters per day, totaled across all forms and converted to water-equivalent. The units are mm/day.
srad	Radiation of the short wavelength. Incident shortwave radiation flux density in watts per square meter, averaged over the day's daylight period. The units are watts per square meter (W/m2).
swe	Equivalent snow water in kilograms per square meter. The amount of water present in the snowpack. The units are kg/m2.
tmax	Maximum temperature of the air. Maximum 2-meter air temperature in degrees Celsius for the day. The units are degrees C.
tmin	Minimum temperature of the air. Minimum daily 2-meter air temperature in degrees Celsius. The units are degrees C.
tmean	Tmean is simply the average of the values for tmin and tmax. This is the average of the extreme temperatures, not the mean temperature. The units are degrees C.
vp	The pressure of water vapor. The pressure of water vapor in pascals. Water vapor partial pressure on a daily basis. The units are pascals (Pa).

Table 2 . Variable availability across years, (DAYMET_COUNTY)

variable name	2017	2018	2019	2020	2021	total rows	null rows
fips	X	X	X	X	X	5876500	0
year	X	X	X	X	X	5876500	0
dayl	X	X	X	X	X	5876500	0
day	X	X	X	X	X	5876500	0
prcp	X	X	X	X	X	5876500	0
srad	X	X	X	X	X	5876500	0
swe	X	X	X	X	X	5876500	0
tmax	X	X	X	X	X	5876500	0
tmin	X	X	X	X	X	5876500	0
vp	X	X	X	X	X	5876500	0
tmean	X	X	X	X	X	5876500	0

5. VETERANS SERVICE ORGANIZATIONS

5.1 DATA SOURCE

US Department of Veterans Affairs

5.2 DESCRIPTION

5.2.1 Veterans Service Organizations (VSOs) and Accredited Representatives

U.S. Veterans Service Organizations (VSOs) are vital in supporting and advocating for veterans. They offer a wide range of services to assist veterans in transitioning to civilian life, accessing healthcare and benefits, and addressing mental health concerns. VSOs also serve as platforms for veterans to connect and share their experiences.

VSOs are nonprofit organizations that provide assistance, support, and advocacy to military veterans and their families. They act as liaisons to help veterans access benefits, navigate civilian life challenges, and foster a sense of camaraderie among former service members.

5.2.2 Types of VA-Accredited Representatives

Accredited representatives assisting veterans are typically categorized into three groups:

1. **VSO Representatives:** These representatives, associated with organizations such as Veterans of Foreign Wars, the American Legion, and Injured American Veterans, provide free representation primarily for initial benefit claims. They can assist in gathering evidence, submitting claims on behalf of veterans, and communicating with the VA regarding claims.
2. **Attorneys:** Attorneys step in after the VA issues its initial decision. They handle the majority of representation and receive compensation. Attorneys assist veterans in navigating the VA appeals process.
3. **Claim Agents:** Claim Agents are independent legal practitioners performing tasks similar to attorneys but are not attorneys themselves. Becoming a claim agent requires passing a written test and undergoing a character and fitness evaluation.

5.2.3 Datasets on VA-Accredited Representatives

The data, especially the historical data spanning several years, was meticulously curated by Jonathan Zhang. Subsequently, we diligently processed the information to ensure it is available in the three distinct formats we provide. We offer three datasets containing information about Veterans Service Organization representatives, including attorneys, agents, and VSO representatives, categorized at the state, zip code, and county levels. These datasets exclusively include individuals accredited by the VA and organizations recognized by the VA as of December 2021. Please note that individuals with pending applications for accreditation or recognition with the VA Office of General Counsel are not included.

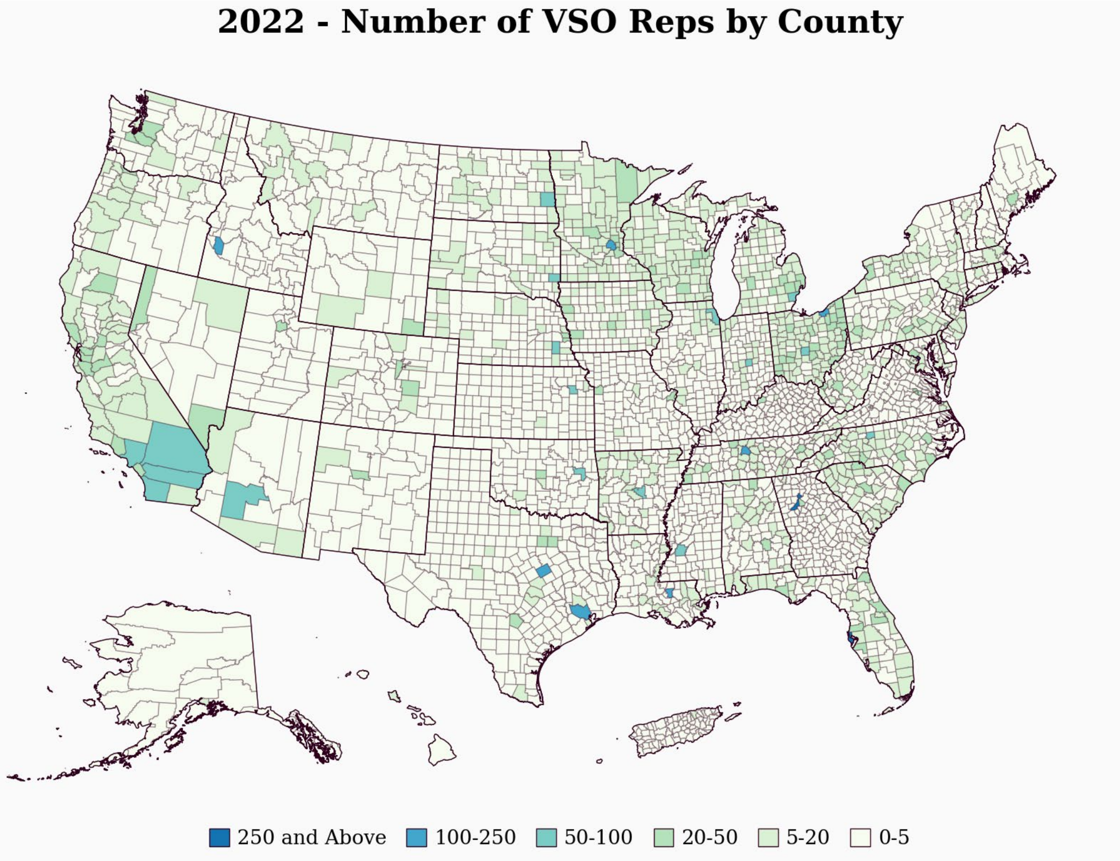


Figure 5. Representation of US Department of Veterans Affairs Representatives by State as of 2022. Among the counties in the United States, the highest number of Veterans Service Organization (VSO) representatives in a single county was recorded in Fulton County, Georgia, with a maximum of 800 representatives in 2022.

5.3 INCLUSION

Year: 2010 – 2022

Geographical unit: FIPS at state, zip code, and county level, for the Continental US plus the following incorporated organized territories: Guam, Puerto Rico, and Virgin Islands.

5.4 RESOURCES

- Link to VA VSO Home
- Link to VSO Accreditation Search
- Link to Jonathan Zhang, PhD, who provided the raw datasets

5.5 UPDATE FREQUENCY

This dataset will be updated and distributed every fiscal year, or as requested by the sponsor. Minimal quarterly updates may be necessary to correct minor data inaccuracies.

Table 3 . Veterans Service Organizations (VSO)

variable name	variable label
fips_code	Federal Information Processing Standards (FIPS) - at state, or county level, depending on the dataset, either the va_vso_state_2010_2022 table or the va_vso_county_2010_2022 table
zip_code	Zip code (only in the va_vso_zipcode_2010_2022 table).
year	The year the data pertain to.
attorneys	Number of Veterans Administration attorneys in the state.
agents	Number of Veterans Administration claims agents in the state.
vso	Number of Veterans Administration veterans service organization representatives in the state.
dataset at state level	702 rows, 5 columns, and no null values.
dataset at zip code level	79960, rows, 5 columns, and no null values.
dataset at county level	24745 rows, 5 columns, and no null values.

Table 4 . Variable availability across years, (VSO)

variable name	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
fips_code	X	X	X	X	X	X	X	X	X	X	X	X	X
zip_code	X	X	X	X	X	X	X	X	X	X	X	X	X
year	X	X	X	X	X	X	X	X	X	X	X	X	X
attorneys	X	X	X	X	X	X	X	X	X	X	X	X	X
agents	X	X	X	X	X	X	X	X	X	X	X	X	X
vso	X	X	X	X	X	X	X	X	X	X	X	X	X

REFERENCES

[1] 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, Rev. 2, ORNL/SPR-2021/2366 - Pub ID 170648. 2021. https://www.osti.gov/biblio/1854468

[2] Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Devineni, P., and Watson, R. VA EDH Data Curation Documentation - FY22-Q1, Rev. 2, ORNL/SPR-2022/2316- Pub ID 172755. 2022. https://www.osti.gov/biblio/1854460

[3] Christian, J.B., Klasky, H.B., Sparks, K., Peluso, A., Tuccillo, J., Rastogi, D., Branstetter, M., Whitehead, M., Hamaker, A., and Watson, R., VA EDH Data Curation Documentation - FY22-Q2, Rev. 2, ORNL/SPR-2022/2391 - Pub ID 174092. 2022. https://www.osti.gov/biblio/1862127

[4] Klasky, H.B., Sparks, K., Logan, J., Tuccillo, J., Whitehead, M., Hamaker, A., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY22-Q3, Rev. 2. ORNL/SPR-2022/2487 - Pub ID 178645. 2022. https://www.osti.gov/biblio/1876283

[5] Klasky, H.B., Sparks, K., Logan, J., Hamaker, A., Whitehead, M., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY22-Q4, ORNL/SPR-2022/2587, PUB ID 183700. 2022. https://www.osti.gov/biblio/1892396

[6] Klasky, H.B., Sparks, K., Logan, J., Hamaker, A., Whitehead, M., Peluso, A., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q1, ORNL/SPR-2022/2694, PUB ID 187842. 2022. https://www.osti.gov/biblio/1909101

[7] Klasky, H.B., Sparks, K., Peluso, A., Whitehead, M., K., Logan, J., Hamaker, A., McGee, M., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q2, ORNL/SPR-2023/2857, PUB ID 19179. 2023. https://www.osti.gov/biblio/1971721

[8] Klasky, H.B., Sparks, K., Peluso, A., K., Logan, J., Hamaker, A., McGee, M., VanDerslice, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q3, ORNL/SPR-2023/2930 PUB ID 195499, 2023. https://www.osti.gov/biblio/1992724

[9] Klasky, H.B., Sparks, K., Peluso, A., K., Myers, A., Hamaker, A., McGee, M., Zhang, J., Logan, J., Hanson, H., Watson, R., and Kapadia, A., VA EDH Data Curation Documentation - FY23-Q4, ORNL/SPR-2023/3097 PUB ID 202517, 2023.

[10] Klasky, H.B., Hanson, H., Sparks, K., Whitehead, M., Blair, C., and Kapadia, A., “Dataset Repository for Investigating Suicide Risk Using Social and Environmental Determinants of Health”, ORNL/TM-2023/3027 Pub ID 183902. 2022. https://www.osti.gov/biblio/1997699

APPENDIX A. ERROR CHECKING

APPENDIX A. ERROR CHECKING

This section lists descriptive statistics of the datasets provided for the FY24Q1 delivery. The following statistics were run using Python’s describe() function.

Daymet

	year	day	dayl	prcp	srad	swe	tmax	tmean	tmin	vp
count	5876500	5876500	5876500	5876500	5876500	5876500	5876500	5876500	5876500	5876500
mean	2019	183	43200.11	3.213138	323.2987	4.614577	19.58305	13.4705	7.357941	1200.572
std	1.414214	105.366	6947.403	8.176923	119.2027	21.62769	11.20685	10.64977	10.57138	785.6372
min	2017	1	0	0	0	0	-36.7603	-39.527	-42.406	14.49863
25%	2018	92	37015.97	0	235.0028	0	11.59693	5.710615	-0.47	545.7857
50%	2019	183	43199.83	0	327.464	0	21.70979	14.74967	7.967053	975.1888
75%	2020	274	49384.31	2.377677	417.0071	0	28.87278	22.59026	16.40932	1812.508
max	2021	365	86400	424.5881	790.2066	993.0573	49.66783	39.58621	32.39859	4382.928

VSO

County:

	year	fips_code	attorneys	agents	vso
count	24745	24745	24745	24745	24745
unique		2338			
top		1001			
freq		13			
mean	2016.055		6.0107901	0.189048	8.271206
std	3.656183		28.802582	0.818465	31.04696
min	2010		0	0	0
25%	2013		0	0	1
50%	2016		1	0	3
75%	2019		3	0	7
max	2022		1107	42	1286

State:

	year	fips_code	attorneys	agents	vso
count	702	702	702	702	702
unique		54			
top		1			
freq		13			
mean	2016		211.87607	6.663818	291.5541
std	3.744325		305.9625	9.213637	369.8989

min	2010		0	0	0
25%	2013		35	1	44.25
50%	2016		91	4	176
75%	2019		265	8	379.5
max	2022		2032	65	2307

Zip code:

Column1	year	zip_code	attorneys	agents	vso
count	79960	79960	79960	79960	79960
unique		9471			
top		28779			
freq		13			
mean	2016.057		1.8601426	0.058504	2.559667
std	3.447715		5.6592801	0.35492	15.25115
min	2010		0	0	0
25%	2013		0	0	0
50%	2016		1	0	0
75%	2019		2	0	2
max	2022		247	37	1190