Daymet V4 aggregated to various TIGER/Line geographies

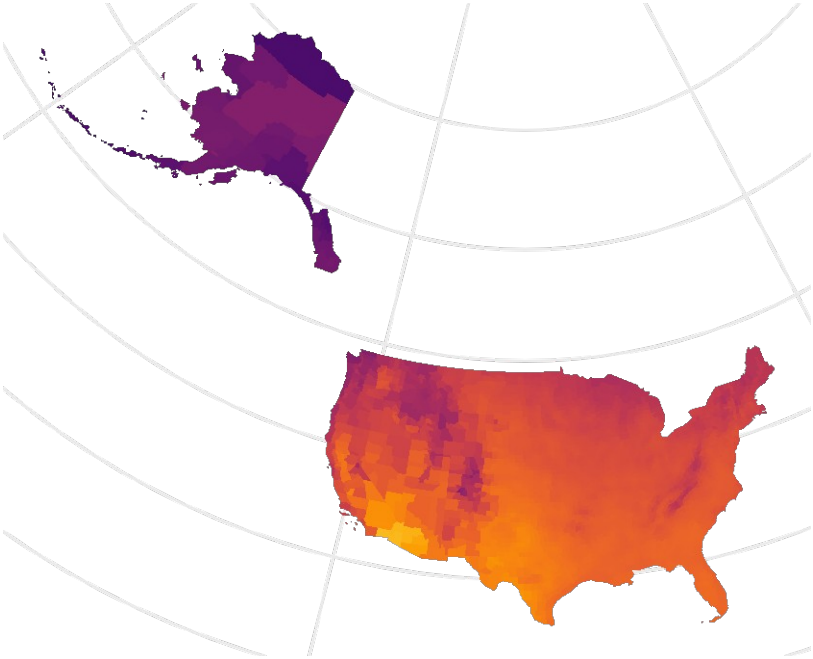

Table of Contents

Overview....2
Repository layout....2
- Geography + TIGER year....3
- Aggregation....3
- Daymet parameter....3
- Part number....4

File layout....4
Helpful code snippets....4
- Joining data....4
- Splitting the date column....4
- Zero-padding geographic identifiers....5

Downloaded from: Daymet v4 daily surface weather aggregated to TIGER/Line geographies

Castro, Edgar; Danesh Yazdi, Mahdieh; Schwartz, Joel, 2021, "Daymet v4 daily surface weather aggregated to TIGER/Line geographies", https://doi.org/10.7910/DVN/MRZGNQ, Harvard Dataverse.

Overview

The data in this repository comprise Daymet V4 1-kilometer gridded daily surface weather data from January 1st, 1980 to December 30th, 2020 aggregated to various U.S. Census Bureau TIGER/Line geographies. The purpose of this repository is to make Daymet data more readily available and accessible for individuals without large computational and storage resources, as raw Daymet data is very large (about 3 TB at the daily resolution).

Data sources and scope

Raw Daymet V4 data is available from the ORNL DAAC website:

> https://doi.org/10.3334/ORNLDAAC/1840

Aggregation has been performed using the exactextract command line utility produced by ISciences, LLC. Source code and extra details are available from the exactextract GitHub page:

> https://github.com/isciences/exactextract

Daymet V4 data has been aggregated to the following geographies, **for the continental U.S. only**:

- Census tracts
- Counties
- ZIP code tabulation areas

The following geographies will be available in a future release:

- Census-designated places
- Census block groups

All source code is available on the daymet-aggregation GitHub page:

> https://github.com/ercas/daymet-aggregation

Repository layout

The files in this repository have been organized as follows:

> **[geography]_[TIGER year]**/**[aggregation]_[Daymet parameter]**_part**[part number]**.csv.gz

For example:

zcta5_2010/mean_tmax_part01.csv.gz

The meaning of each component of the general file name template will be explained in the following subsections.

Geography + TIGER year

The **geography** and **TIGER year** components refer to the administrative level and Census year of boundary files that the Daymet data have been aggregated to. All boundary files have been downloaded from the U.S. Census TIGER/Line website, 2010 edition:

https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2010.html

Aggregation

The **aggregation** component describes how grid cells intersecting each polygon from the TIGER/Line boundary files has been aggregated using exactextract:

- **min** refers to the minimum value of any intersecting grid cell,
- **max** refers to the maximum value of any intersecting grid cell, and
- **mean** refers to the coverage fraction-weighted mean of intersecting grid cells, where the coverage fraction refers to the proportion of each grid cell that is within a given polygon.

Daymet parameter

The **Daymet parameter** component describes which parameter was aggregated. A full list of Daymet parameters with accompanying descriptions is available from the Daymet overview page:

https://daymet.ornl.gov/overview

The table has been partially reproduced below for convenience:

Daymet parameter	Definition	Units
dayl	Day length	s day^{-1}
prcp	Precipitation	mm day^{-1}
srad	Shortwave radiation	W m^{-2}
swe	Snow-water equivalent	kg m^{-2}
tmax	Maximum air temperature	°C
tmin	Minimum air temperature	°C
vp	Water vapor pressure	Pa

Part number

Due to Harvard Dataverse file size limitations, the files have been split into several parts. Parts have been labeled with an incrementing integer zero-padded to two digits. Please refer to the MANIFEST.txt file for more information about total file sizes and record counts when in doubt.

File layout

Each file is composed of three columns:

- **[A geographic identifier column]**: A value linking each record to a corresponding geometry in the TIGER/Line boundary files. This varies by file: all 2010 geographies use GEOID10, while 2000 counties use CNTYIDFP00, 2000 tracts use CTIDFP00, and 2000 zip code tabulation areas use ZCTACE00.
- **date**: The Daymet calendar day that the record corresponds to, formatted as an ISO 8601 date without delimiters (YYYYMMMDD).
    - Note that Daymet files all have 365 days and omit leap year days; see the Calendar section of the following page for more information: https://daymet.ornl.gov/overview
- **value**: The aggregated Daymet value that the record corresponds to.

Helpful code snippets

Due to peculiarities in the tools used and Dataverse file size limitations, some additional work may be required to reconstruct the original output files or merge with other data. This section provides code snippets in the R language that may be useful in preparing the data for certain joins and subsets.

Joining data

For reading and joining data, it is recommended to use the data.table library due to speed in reading tabular plaintext and in combining multiple data parts via rbindlist(). Reading can be done as follows:

```
library(data.table)
data <- rbindlist(lapply(Sys.glob("zcta5_2010/mean_tmax_part*.csv.gz", fread)
```

Splitting the date column

In some cases, the year, month, and day components of the date column must be split into separate columns for subsets or joins. While this can be done using date-aware functions from the base R library or packages like lubridate, this may not the most efficient method and inefficiencies can become

exacerbated due to the size of the data. Instead, it may be faster to extract these components using algebra and the floor() function:

```
library(data.table)
data <- fread("zcta5_2010/mean_tmax_part_01.csv.gz")
data[, year := floor(date / 1e4)]
data[, month := floor((date / 1e4 - year) * 1e2)]
data[, day := (date / 1e2 - floor(date / 1e2)) * 1e2]
```

Zero-padding geographic identifiers

In some cases, the geographic identifiers, which are normally zero-padded, may become reduced to simple integers. This can be quickly checked by calling the min() function on the geographic identifier column and seeing if there are any leading zeroes.

This issue makes it not possible to join the aggregate data to the original boundary files or other data that references them without reformatting the identifiers. This can be remedied by using sprintf():

```
library(data.table)
data <- fread("zcta5_2010/mean_tmax_part_01.csv.gz")
data[, ZCTA5CE00 := sprintf("%05.f", ZCTA5CE00)]
```

where "%05.f" is a C-style string formatting placeholder. See the table below for more details on which placeholder to use:

Geographic level	Variable name (2000)	Variable name (2010)	Length	Placeholder
County	0	0	5	`"%05.f"`
Tract	0	0	11	`"%011.f"`
ZIP code tabulation area	0	0	5	`"%05.f"`

The lengths of other geographic identifiers can be found in the GEOID Structure for Geographic Areas section of the following page:

https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html