Documentation

Table of contents

[Summary 2](#_Toc42563)

[SDWA Reporting Universe 2](#_Toc42564)

[Count of Systems by Primacy Agency 2](#_Toc42565)

[Systems by Type of Service Area 4](#_Toc42566)

[State Data 4](#_Toc42567)

[Place Matched Service Area Boundaries 5](#_Toc42568)

[Mobile Homes - OSM 5](#_Toc42569)

[Mobile Homes - Parcels 5](#_Toc42570)

[Modeled Service Area Boundaries 6](#_Toc42571)

[Binary Water Use Model 6](#_Toc42572)

[1:1 Match 8](#_Toc42573)

[Random Forest 8](#_Toc42574)

[Results 13](#_Toc42575)

[State Boundary Data 14](#_Toc42576)

[Arizona 14](#_Toc42577)

[Arkansas 15](#_Toc42578)

[California 16](#_Toc42579)

[Colorado 17](#_Toc42580)

[Connecticut 17](#_Toc42581)

[Florida 18](#_Toc42582)

[Kansas 18](#_Toc42583)

[Mississippi 18](#_Toc42584)

[New Hampshire 18](#_Toc42585)

[New Jersey 19](#_Toc42586)

[New Mexico 19](#_Toc42587)

[New York 19](#_Toc42588)

[North Carolina 19](#_Toc42589)

[Pennsylvania 20](#_Toc42590)

[Washington 20](#_Toc42591)

[West Virginia 20](#_Toc42592)

[Bibliography 20](#_Toc42593)

Version 1.0 SDWA Data Vintage: 2023 (Quarter 4)

The Office of Research and Development has released a publicly available dataset of community water system (CWS) service area boundaries (SAB). CWS are defined as systems that provide water for human consumption through pipes or other constructed conveyances to at least 15 service connections or serves an average of at least 25 people year-round [1]. Under the safe drinking water act (SDWA) and various drinking water rules (see EPA Drinking Water Regulations) public water systems must test and report on the quality of their drinking water. These boundaries enable the link between data reported to SDWA and geographic areas served by specific systems. The boundary data is a collection of publicly available data and modeled boundaries. This document describes how the data was collected and modeled and details the modeling techniques used to generate the full dataset.

# Summary

The boundaries presented in this dataset are an aggregate of several methods used to present the most accurate representation possible with available data. These methods range in source and accuracy. Each method is described in detail.

# SDWA Reporting Universe

The current ORD dataset (Version 1.0), uses the 2023 (Q4) release of SDWA data. As of this release, there were a total of 49,396 active community water systems, which primarily serve residential areas (Figure 2). There are many systems, however, which primarily wholesale water to other systems and therefore do not have a service area boundary. The universe of systems in the ORD dataset is limited to currently active community water systems which serve at least 25 people and have at least 15 service connections. The total number of systems that fit these criteria for 2023 (Q4) is: 49,396.

# Count of Systems by Primacy Agency

Community water systems each fall under a primacy agency, which is the agency with primary responsibility for implementing SDWA, this is typically the state or territory. The exceptions, where EPA is the primacy agency are Wyoming, the District of Columbia, and federally recognized tribes excluding Navajo Nation. Navajo Nation is currently the only tribe with primacy agency status to enforce SDWA.

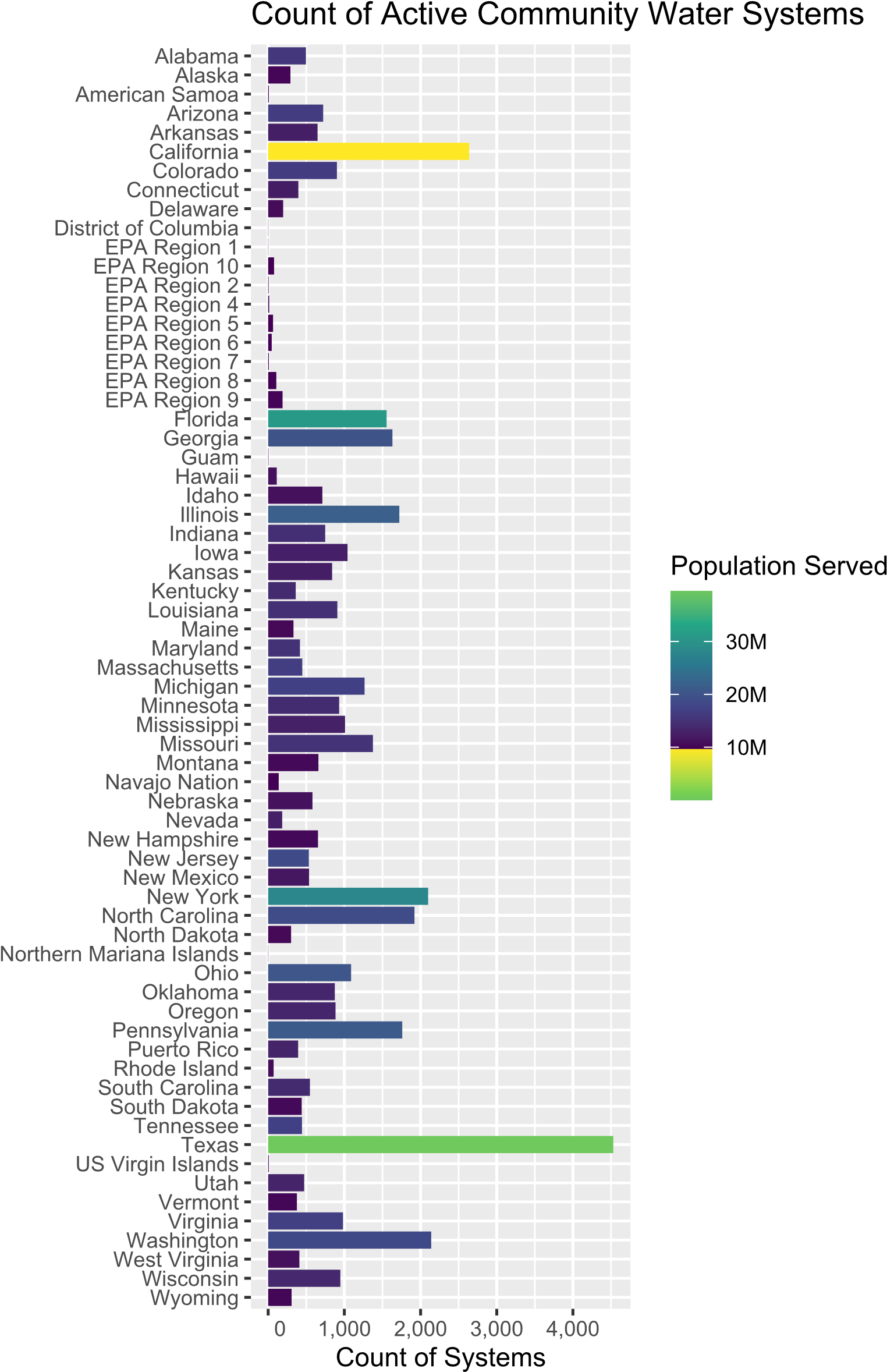


Figure 1: Bar plot showing the count of active community water systems by state, region or territory.

# Systems by Type of Service Area

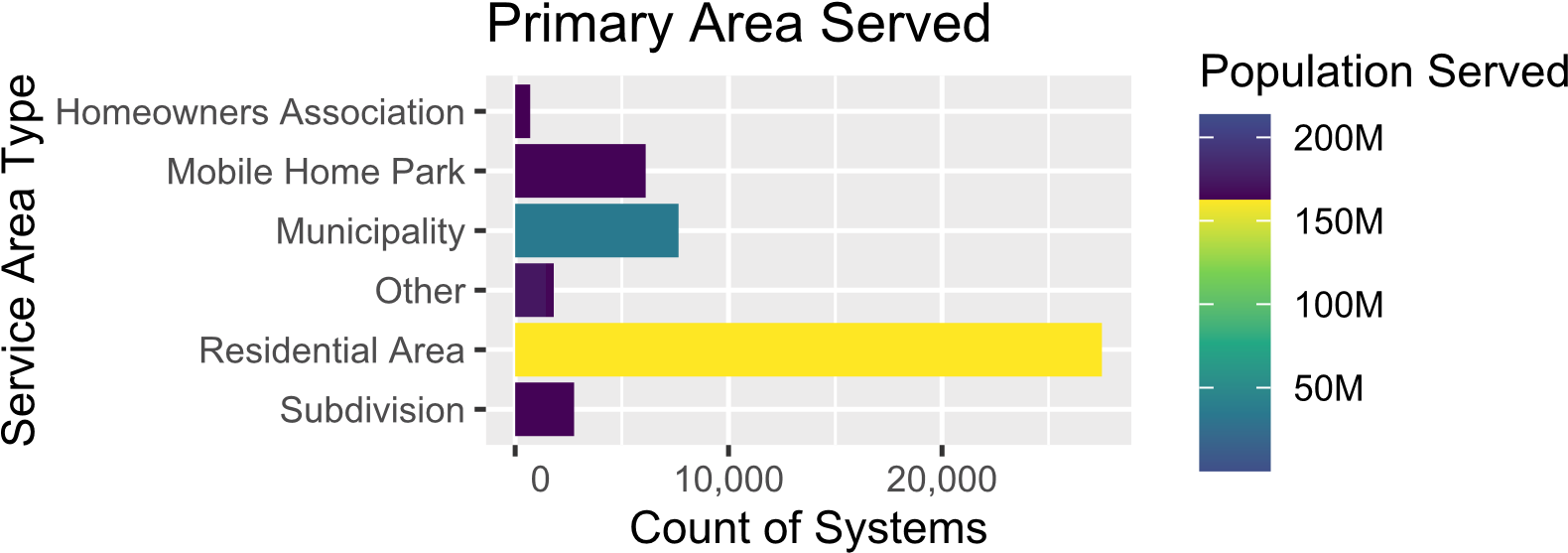


Figure 2: Types of Active Systems in SDWA Reporting

# State Data

When detailed water service area boundaries are publicly available, they are considered to be the highest quality spatial data possible. A comprehensive review was conducted of available boundary data and used to determine what would be included in the ORD national dataset versus what would be modeled. Detailed descriptions of state data are available in the state boundary appendix.

Table 1: Number of system boundaries from state sources included in ORD dataset

|  |  |
| --- | --- |
| State | Count of Systems |
| Arizona | 708 |
| Arkansas | 670 |
| California | 2,812 |
| Connecticut | 448 |
| Florida | 420 |
| Kansas | 781 |
| Kentucky | 367 |
| Mississippi | 361 |
| New Hampshire | 658 |
| New Jersey | 560 |
| New Mexico | 532 |
| New York | 308 |
| North Carolina | 482 |
| Pennsylvania | 1,770 |
| Tennessee | 431 |
| Texas | 4,538 |
| Utah | 484 |
| Washington | 1,747 |
| West Virginia | 235 |

# Place Matched Service Area Boundaries

# Mobile Homes - OSM

Areas delineated as mobile home parks were extracted from Open Street Map and name matched with community water system names reported under SDWA. To find open street map delineated areas, the point locations for mobile home parks from Homeland Infrastructure Foundation-Level Data (HIFLD) were intersected with areas tagged as ‘residental=trailer\_park’ in open street map. If a match was found, the given names of that mobile home park (from both sources) were matched with SDWA reported names.

# Mobile Homes - Parcels

Where open street map delineated areas were not present, point locations of mobile home parks from HIFLD were intersected with parcels. The name of the mobile home park was then matched with the SDWA reported public water system names.

# Modeled Service Area Boundaries

# Binary Water Use Model

A decision tree model was created to determine the probability that a census block is served by a public water system. The variables included are listed in Table 2. To validate the model, public water system boundaries from three states (New Jersey, Connecticut & California) were joined to census blocks to indicate the existence of a public water system. These three states were chosen based on the completeness and detail of their service area boundaries. For more details on this initial model, refer to the 2020 Water Source Model GitHub Site. The final decision tree is illustrated in Figure 3.

Table 2: Variables included in decision tree model.

|  |  |
| --- | --- |
| Variable | Description |
| Imperviousness | Urban Imperviousness in percent, derived from the National Land Cover Dataset |
| 2020 Housing Unit Density | Housing units per square kilometer |
| Area | Area in square kilometers |
| % Public Use in 1990 | % of housing units that reported public water use from  1990 Census |
| % Sewer Use in 1990 | % of housing units that reported public sewer use from  1990 Census |
| 1990 Housing Units | Count of housing units in block from 1990 Census |
| % Housing Unit Change 1990-2020 | % change in housing units between 1990 and 2020 Census |
| Distance to Public Water Intake | Distance in kilometers between block centroid and closest public water intake as derived from the Safe Drinking Water Information System |

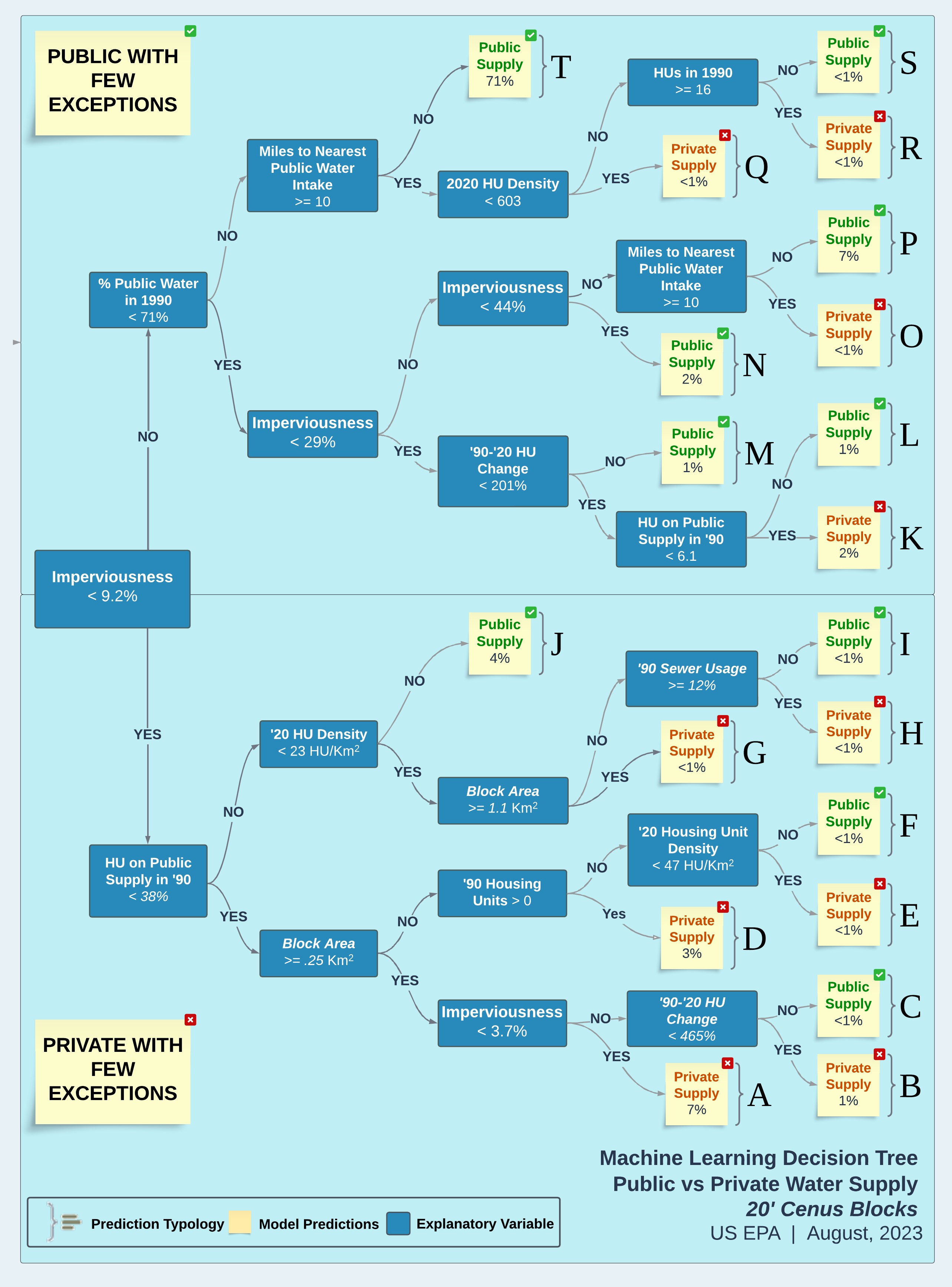


Figure 3: Visualization of decision tree to predict whether a census block is served by a public water system.

# 1:1 Match

Census blocks that are predicted to be served by public water by the binary water use model (Figure 3) are ‘aggregated’ or ‘dissolved’ spatially, meaning they are combined with any contiguous blocks that are also estimated to be served by public water systems into larger polygons. These aggregated polygons are then spatially joined with water system facility locations from SDWIS, which include intakes, wells and treatment plants. If a single aggregated area can only be associated with facilities from a single public water system, that area is assigned to the identified PWSID.

# Random Forest

Predictor Variables (factors)

The goal of the random forest model is to determine what type of community a block is within and based on the local public water infrastructure, what system it is most likely to be served by. For example, a block in a rural area that is using public water may be more likely to be very close to its source water intake or well if it is a very small system and more likely to be farther away if it a very large rural water district. If a system purchases all of its water from another system or systems, its facility locations may be farther away as well.

The random forest is set up to evaluate the relationship between a single census block and a single point associated with a public water system. The tabular data used to train and apply the random forest model has one row for the relationship between every block and every facility point within 25 miles. As an example, if there are seven different facilities within twenty-five miles of a census block, the table used for the random forest will have seven rows for that census block. Each facility is associated with a parent PWSID. The random forest model then predicts a probability that the parent PWSID is serving the census block in the same row of the table. Predictor variables can be thought of as belonging to one of two groups:

* Variables that that characterize the census block
* Variables that characterize the water system.

The random forest model then determines the correct interplay between the variables to determine a probability that a particular system serves a particular census block.

Variables that Describe the Census Block

Population

Variable Name: Population

Population is taken from the 2020 Census, obtained at the census block level.

Local Population Density

Variable Name: Pop\_km

The centroid of each block is buffered by 2 miles (3,218.69 meters). The population density is then calculated as an areal weighted population density of census blocks that intersect the buffer.

Area

Variable Name: Area\_Km

Area is calculated for each block in square kilometers using the Albers equal area projection (crs = 5070).

Probability of Public Water

Variable Name: Prob\_Pub

The probability (0-1) of a block being served by public water as estimated from the ORD water use decision tree model (Figure 3).

Buildings

Variable Name: nBuildings

The number of buildings within a census block that are greater then 50 square meters in area. This is calculated using microsoft building footprints.

Percent Buildings

Variable Name: PctBldg

The percent area of the census block that is covered by buildings. This is calculated as the total area of microsoft building footprints divided by the area of the census block.

Mean Building Area

Variable Name: meanBldg\_m

The mean area in square meters of buildings within the census block.

Minimum Building Area

Variable Name: minBldg\_m

The area in square meters of the smallest building within the census block (Only buildings greater than or equal to 50 square meters included).

Maximum Building Area

Variable Name: maxBldg\_m

The area in square meters of the largest building within the census block (Only buildings greater than or equal to 50 square meters included).

Standard Deviation of Building Area

Variable Name: sdBldg\_m

The standard deviation of area of buildings in square meters within a census block.

Rural / Urban

Variable Name: PctRural The census defined urban / rural classifier for 2020 (census block level).

Mean Residential Acres

Variable Name: meanResAcres

The mean value in acres of parcels within the census block that are zoned for residential use.

Count of Parcels

Variable Name: nParcels

The total count of parcels within the census block.

Count of Mobile Homes

Variable Name: MH\_Count

The count of mobile home communities within a census block as derived from the (Homeland Infrastructure Foundation Level database)[https://hifld-geoplatform.opendata.arcgis.com/datasets/ mobile-home-parks/explore].

Mobile Home Size

Variable Name: MH\_Size

The cumulative size of mobile home parks within a census block representing the number of mobile home units. This variable is presented as a factor.

Possible Values:

|  |  |
| --- | --- |
| Value | Description |
| ‘50’ | <50 Mobile Homes |
| ‘75’ | 50-100 Mobile Homes |
| ‘100’ | >100 Mobile Homes |

Variables that Describe the Closest Systems

Distance

Variable Name: Facility\_Dist The distance in meters between the census block centroid and a single point from SDWIS for a system.

Distance Rank

Variable Name: Dist\_Rank

Describes the closeness rank of the particular facility being measured to. For example, if Dist\_Rank = 5, there would be 4 other facility points closer to the centroid of that census block.

Facility Type

Variable Name: Facility\_Type

The type of system point that was used in the distance calculation. Options include:

* “Well”
* “Treatment Plant”
* “Consecutive Connection”
* “Intake”
* “Other”

The ‘Other’ Category contains less frequent data and data used when there are no wells, intakes or treatment plants associated with a system. Some examples of less frequent locations are springs, reservoirs and infiltration gallery. This class also contains the office addresses of systems as reported to SDWIS which are typically within a service area but are also known to be unreliable depending on the public water system. Office address locations were geolocated.

Population Served

Variable Name: Population\_Served\_Count

The reported population that is served by the system in SDWA reporting.

Connections

Variable Name: Service\_Connections\_Count

The reported number of service connections within a system in SDWA reporting.

Distance to Center of System

Variable Name: Ctr\_Dist

If a system has more than one SDWIS point (intakes, wells treatment plants etc…) the mean center of all points is calculated and measured in meters from the centroid of the census block. If only one point exists within a system, this value will be identical to Facility\_Dist.

System Type

Variable Name: Service\_Area\_Type

The Primary type of area that is served by the public water system.

Possible Values:

* ‘Homeowners Association’
* ‘Mobile Home Park’
* ‘Multiple’
* ‘Municipality’
* ‘Other’
* ‘Residential Area’
* ‘Subdivision’

Sub-County Match

Variable Name: SubCounty\_Match

Reflects a classified Jaro-Winkler string distance between the census place the census block is within and the ‘City Served’ of the public water system associated with the point being measured to.

|  |  |
| --- | --- |
| Jaro-Winkler Distance | Classification |
|  |  |
| if no census sub county | “No SubCounty” |
| if no city served | “No City Served” |

Place Match

Variable Name: Place\_Match

Reflects a classified Jaro-Winkler string distance between the census place the census block is within and the ‘City Served’ of the public water system associated with the point being measured to.

|  |  |
| --- | --- |
| Jaro-Winkler Distance | Classification |
|  |  |
| if no census place | “No Place” |
| if no city served | “No City Served” |

County Match

Variable Name: County\_Match

A categorical value denoting whether the county that the census block is within matches a county reported to be served by the system being measured.

Possible Values:

* ‘Match’
* ‘No Match’
* ‘No County’

Place Name Present in System Name

Variable Name: Place\_in\_PWS

A measure of how much of the system name also appears in the census place name that the census block is within. The longest common sub-string (LCS) is calculated between the census place and the system name and is then using the formula:

(𝑝𝑤𝑠𝑁𝑎𝑚𝑒𝑙𝑒𝑛𝑔𝑡ℎ − 𝐿𝐶𝑆)/𝑃𝑙𝑎𝑐𝑒𝑙𝑒𝑛𝑔𝑡ℎ

where 𝑝𝑤𝑠𝑁𝑎𝑚𝑒𝑙𝑒𝑛𝑔𝑡ℎ is the length of the place name string in characters, 𝑃𝑙𝑎𝑐𝑒𝑙𝑒𝑛𝑔𝑡ℎ is the length of the public water system name in characters and 𝐿𝐶𝑆 is the length of the longest common sub-string between the two. An exact match would result in a value of 1. If either string is missing, a value of zero is assigned.

Sub-County Name Present in System Name

Variable Name: SC\_in\_PWS

A measure of how much of the system name also appears in the census place name that the census block is within. The longest common sub-string (LCS) is calculated between the census place and the system name and is then using the formula:

(𝑝𝑤𝑠𝑁𝑎𝑚𝑒𝑙𝑒𝑛𝑔𝑡ℎ − 𝐿𝐶𝑆)/𝑆𝑢𝑏𝐶𝑜𝑢𝑛𝑡𝑦𝑙𝑒𝑛𝑔𝑡ℎ

where 𝑝𝑤𝑠𝑁𝑎𝑚𝑒𝑙𝑒𝑛𝑔𝑡ℎ is the length of the place name string in characters, 𝑆𝑢𝑏𝐶𝑜𝑢𝑛𝑡𝑦𝑙𝑒𝑛𝑔𝑡ℎ is the length of the public water system name in characters and 𝐿𝐶𝑆 is the length of the longest common sub-string between the two. An exact match would result in a value of 1. If either string is missing, a value of zero is assigned.

Training & Validation

The random forest is applied to every census block and returns a probability for each facility location within 25 miles of each census block. That probability can be interpreted as the probability that the system PWSID associated with that facility is serving public water to that census block. A full comparison between the training and testing sets is shown in ?@tbl-train

The data used to build the random forest model was split randomly across Arizona, Arkansas, California, Connecticut, New Jersey and Texas into training and testing sets (?@tbl-train). This is done so the relationships constructed by the random forest model can be evaluated against data it has not been exposed to.

|  |  |  |
| --- | --- | --- |
|  | Training Set | Testing Set |
| # Rows | 3.1 Million | 47 Million |
| # Census Blocks | 760 Thousand | 1.1 Million |
| # Correct | 350 Thousand | “No Match” |
| # Incorrect | if no census place | “No Place” |
| # Systems | if no city served | “No City Served” |

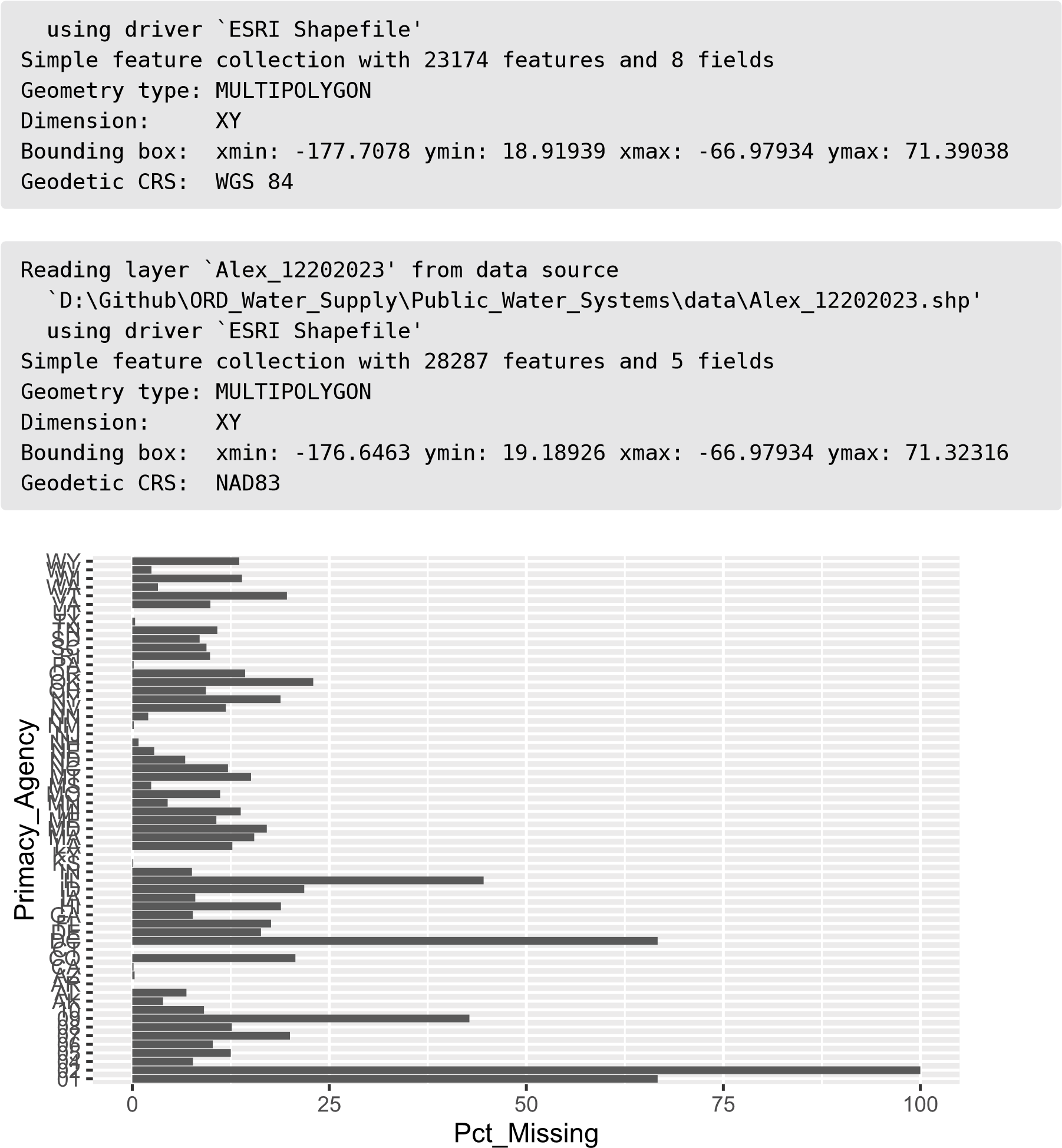
Application

# Results

Reading layer `Final\_Boundaries' from data source

`D:

\Github\ORD\_Water\_Supply\Public\_Water\_Systems\outputs\Final\_Boundaries.shp'



# State Boundary Data

# Arizona

CWS Service Area (Arizona Department of Water Resources)

Link: https://gisdata2016-11-18t150447874z-azwater.opendata.arcgis.com/datasets/cws-servicearea-1/about

Last Updated: March 24, 2022

Description

“The purpose of this feature class is to provide service area boundaries for community water systems regulated by the Arizona Department of Water Resources. This feature class contains service area polygons for each Community Water System (CWS). To determine the service area, ADWR utilized primary data provided directly from the water system (i.e. PDF, shapefile, verbal definition). If primary data is unavailable, secondary data was utilized to determine service area boundaries (i.e. Certificate of Convenience and Necessity (CCN), Census Designated Place shapefile from U.S Census Bureau.) New water systems are added, and contact information is updated for existing water systems on an annual basis. Service area maps are updated every 5 years. ADWR cannot verify the spatial accuracy of the information contained on this map.” Arizona also supplies a mapper (https://azwatermaps.azwater.gov/cws) with this layer and others (active/inactive well locations, groundwater basins) as a tool to allow the public to identify water providers in their area. It is also used by water systems to view the contact information and service area boundary that ADWR has on file for them. (They can submit corrections separately if anything is not right). The mapper will link to all annual reports and other files/correspondence for the CWS available to the public on the ADWR website. There is also a Community Water Systems Data Dashboard that includes visualizations from annual report data. There are visuals of population served, deliveries, demand, emergency water supply, etc.

Method

“ADWR utilized primary data provided directly from the water system (i.e., PDF, shapefile, verbal definition). If primary data is unavailable, secondary data was utilized to determine service area boundaries (i.e., Certificate of Convenience and Necessity (CCN), Census Designated Place shapefile from U.S Census Bureau.)” New systems are added annually by the AZ Department of Water Resources (ADWR). The service area maps are updated every 5 years. ADWR does not verify the shapes presented are accurate.

Coverage

* 886 PWSs in the file (887 records – one PWS is listed twice)
* Includes 717 of the 746 active CWS in SDWIS (96%)
* Missing 29 active CWS from SDWIS

# Arkansas

Link: https://gis.arkansas.gov/product/public-water-systems-polygon/

Last Updated: August 7, 2023

Description

“This dataset contains polygons which represent public water system boundaries in the State of

Arkansas. The compilation of this data is an effort of the Engineering Division of the Arkansas Department of Health (ADH) to build a comprehensive geographic database of water utilities and services in the public water system. A visual aid of water system boundaries overlaid on current digital aerial photography, associated road names, and landmarks, were verified by representatives of ADH to confirm the accuracy of the boundaries.”

Method

“A visual aid of water system boundaries overlaid on current digital aerial photography, associated road names, and landmarks, were verified by representatives of ADH to confirm the accuracy of the boundaries.” The data are compiled by the Engineering Division of the Arkansas Department of Health. Number of shapes (and likely coverage of shapes) appears to be being updated over time.

Coverage

* 787 PWSs in the file (788 shapes – one PWS is listed twice)
* Includes 657 of the 672 active CWS in SDWIS (98%)
* Missing 15 active CWS from SDWIS, although with fuzzy name matching they are likely in the file

Notes

* Does not overlap with Census Place, county, Zip Code, data.
* One PWS is included twice/has two separate shapes
* Some shapes seem to follow roads/pipes
* Some shapes overlap with each other
* Does not include active/inactive indication
* No data fields include indication of method/quality
* Includes PWSID (objected = numeric portion of SWDIS PWSID)

# California

Link: https://gis.data.ca.gov/datasets/fbba842bf134497c9d611ad506ec48cc/explore

Last Updated: November 13, 2023

Description

“Service area boundaries of drinking water service providers, as verified by the Division of Drinking Water, State Water Resources Control Board. In order to provide an accurate data set of service area boundaries for California drinking water systems, the Division of Drinking Water of the California Water Resources Control Board (SWRCB DDW) has undertaken a project to vet and verify the data collected by the Tracking California’s Water Boundary Tool (WBT). SWRCB DDW downloaded a copy of the current water system service areas loaded in the WBT as of June 27, 2019. Additional attribute fields indicating verification status, verification staff and system type were appended to the data set. SWRCB DDW staff are reviewing and validating the displayed boundaries of each service area as well as contacting the service providers regarding necessary corrections. The verification status of any particular service area may be found in the Verification Status field.”

Method

Water systems and other qualified users used California’s Water Boundary Tool (WBT) to input or edit water system service area boundaries. The majority of these service area boundaries were adapted from paper maps or individual waster providers digitized service area boundaries.

Coverage

* 4,782 PWSs in the file (20 duplicate records)
* Includes 2,788 of the 2,842 active CWS in SDWIS (98%)

Notes

* Shapes have overlap with each other
* Does include active/inactive field
* Includes verification status and date for each shape
* Data are maintained and updated continuously
* Incudes the type of boundary (served area or jurisdictional) and the original source of the data
* Methods and details for the WBT are well documented

# Colorado

Link: https://data.openwaterfoundation.org/state/co/owf/municipal-water-providerboundaries/

Last Updated: Nov 14, 2023

Description

“Boundaries and district names aggregated for this dataset represent a first version of effort toward an authoritative Statewide Special Districts Dataset. Each was aggregated from thousands of local jurisdictions by the Colorado Department of Local Affairs Demography office. Many of the district boundaries were created from scanned drawings or digitized PDFs, and therefore no guarantee of accuracy can be made for the data.”

Method

* PWSID not included in the file, however, it is possible to link the LGID to PWSID using the data provided in a related spreadsheet (https://github.com/OpenWaterFoundation/owf-data-comunicipal-water-providers/blob/main/data/co-municipal-water-providers.xlsx)
* There are 254 systems in the Districts map, and 531 in the spreadsheet (949 active CWSs in SDWIS).

Notes

* Shapes do overlap with each other
* Metadata and method information is sparse.
* The “source” field does provide detailed information about where the data come from.
* Would need to link multiple sources to get all the data available, and coverage is still low (25% to 50% of active CWSs)

# Connecticut

Link: https://maps.ct.gov/portal/home/item.html?id=684908bf05a2430f8a60d58a96d640d6

Last Updated: March 2, 2020

Description

An approximation of public water system service areas in Connecticut

Method

Shapes are a buffered approximation based on service lines.

Coverage

* 531 PWSIDs in the file
* Includes 448 of the 477 active CWS in SDWIS (94%)

Notes

* Shapes have some overlap with each other
* Metadata is minimal
* Does not include active/inactive field
* No overlap with Census Places
* Includes PWSID to link to SDWIS

# Florida

Link:

Last Updated

Description

Method

Coverage

# Kansas

Link:

Last Updated

Description

Method

Coverage

# Mississippi

Link:

Last Updated

Description

Method

Coverage

# New Hampshire

Link:

Last Updated

Description

Method

Coverage

# New Jersey

Link:

Last Updated

Description

Method

Coverage

# New Mexico

Link:

Last Updated

Description

Method

Coverage

# New York

Link:

Last Updated

Description

Method

Coverage

# North Carolina

Link:

Last Updated

Description

Method

Coverage

# Pennsylvania

Link:

Last Updated

Description

Method

Coverage

# Washington

Link:

Last Updated

Description

Method

Coverage

# West Virginia

Link:

Last Updated

Description

Method

Coverage

# Bibliography

[1] S. D. W. Act, “Safe drinking water act”, in Enacted by the 93rd United States Congress. Effective, 1974.