U.S. PRIVATE WELL ESTIMATES (2020)

Contents

> **!** Important
>
> This document is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the U.S. Environmental Protection Agency. It does not represent and should not be construed to represent any agency determination or policy.

> Web Application
>
> A draft version of the accompanying web application is available for review purposes here: LINK

Overview

This document details our efforts on estimating water use in the United States for 2020. This work is listed as EPA Office of Research and Development sub-product **SHC.404.2.1.1**: 'US Private Domestic Well Density - 2020 Update'. In our previous formula for estimating wells Murray et al. (2021), we noted limitations in areas of significant housing unit change or population growth. We are proposing a new method using machine learning, which helps to resolve these limitations and more accurately estimate domestic well use in 2020. An

added benefit of this work is that by knowing where people are using domestic wells, we can also determine the locations of public water users. Delineation of public water service boundaries is a priority of the Environmental Protection Agency, and this work was completed with future applications in mind.

Previous Estimates (2010)

We published "Methods for Estimating Locations of Housing Units Served by Private Domestic Wells in the United States Applied to 2010" in 2021, based on work done between 2015 and 2017. The 1990 long form Census was the last time a comprehensive national survey was done which asked whether people obtained their domestic water from a public source or a well. This provides a baseline for domestic water supply at the census block group level for 1990. We then obtained well drillers logs from 20 states which have required their reporting since at least 1990. This empirical well drilling data was used as a validation for well estimates using a second method based on housing unit growth.

The empirical method uses the following formula:

$$P_{pdw(est)} = P_{pdw(init)} + \Delta\frac{N_w}{A} - f_{pdw}\frac{N_{HU(lost)}}{A}$$

where $P_{pdw(est)}$ is the PDW density estimate, $P_{pdw(init)}$ is the initial PDW density, $\Delta\frac{N_w}{A}$ is the change in the number of housing units reliant on wells N_w, and A is the area for analysis (km2), f_{pdw} is the fraction of PDW use to total water supply, and $\frac{N_{HU(lost)}}{A}$ is the number of housing units lost per unit area. The initial PDW density and f_{pdw} are inferred from the 1990 census results. N_w is calculated from geolocated well drilling records. The quantity f_{pdw} is updated after each incremental calculation is made, allowing for changing spatial patterns of PDW use. Including the loss of housing units accounts in part for the loss of PDWs.

For states where well records do not go back to 1990, we relied on what we termed the 'Net Housing Unit (NHU) method', which is represented by the formula:

$$P_{pdw(est)} = P_{pdw(init)} + f_{pdw}\Delta\frac{N_{HU}}{A}$$

where $\Delta\frac{N_{HU}}{A}$ is the net change in housing units per unit area (km2). The fraction of private well use f_{pdw} is determined from the 1990 census results.

To put it simply, we found that census block groups that were reliant on private wells in 1990 were still reliant on private wells in 2010. The NHU method essentially assumes that the rate of well use remained constant between 1990 and 2010. In the paper we note that this is not always true. There are of course areas that experience rapid development, which in turn necessitates the expansion or creation of public water systems. The farther we move away from the baseline data from the 1990 Census, the more this occurs as the United States continues to grow. Both our paper and others, specifically Johnson & Belitz have noted that housing unit density can grow to the point where private well use declines and public water increases. In our 2020 update of this work, we use machine learning to identify census blocks reliant on private or public water, and estimate the number of housing units reliant on each for the fifty states and Washington D.C.

Figure 1 shows the correlations of using a housing unit based well estimation versus our formula with drillers logs as a function of housing unit density change between 1990 and 2010.

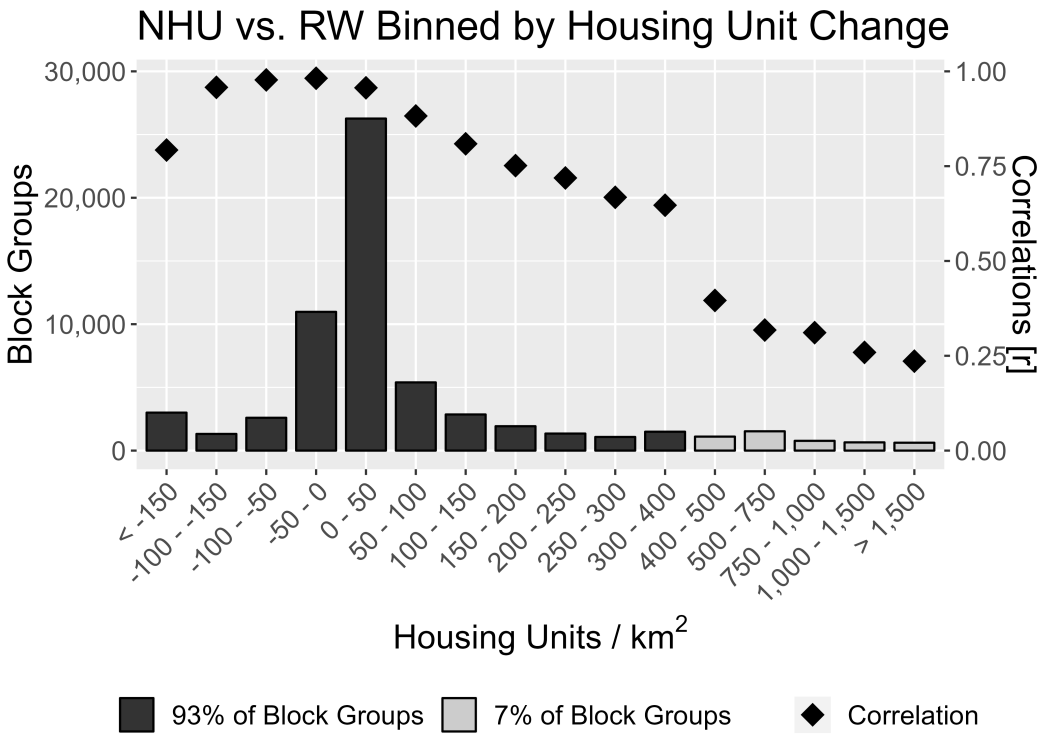

Figure 1: Plot showing that correlations between RW and NHU method drop significantly when housing units density is greater than 400 per sq. km.

Our predictive power starts to drop after increases of 50 housing units / km^2, with a sharp drop after 400. Our new method aims to more accurately estimate wells in these areas.

Characteristics of Areas Reliant on Private vs. Public Water

Several factors play a part in why a home may use private or public water. We have already identified housing unit density as a key indicator of type of water supply but have not previously identified any specific thresholds.

1990 Source of Water

The cumulative water sources as reported in the 1990 Census are shown in Figure 2. In 1990, roughly 84% of respondents reported using public water, 15% reported using either a dug or drilled well and just over 1% reported using some other source. Other sources include cisterns, springs, creeks and other unregulated sources. Any source other than a public water supply does not fall under the jurisdiction of the safe drinking water act and is therefore considered unregulated. The importance of this being that water quality testing is then the responsibility of the homeowner or tenant.

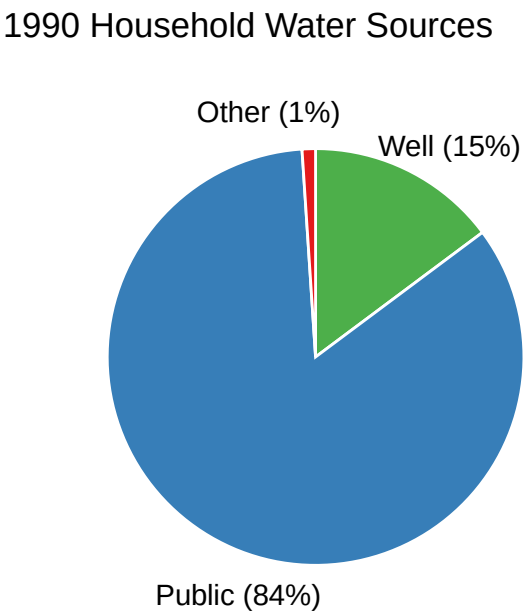

Figure 2: Pie chart showing 1990 Census source of water aggregate percentages. 84% reported using public water, 15% reported using a well and 1% reported using some other source.

Housing Unit Density and Public Water Use

The 1990 long form Census was the last comprehensive survey that asked where people living in the United States obtained their household water from. We have published methods in the past, which estimated the number of domestic wells at the Census block level for 2010. We are now updating that data for 2020 along with attempting to also delineate public water systems. A major driver of whether or not a community is served by public water is the density of housing. When homes are more densely located, there is a cost benefit to supplying water from public sources as opposed to having a separate well from each home. Of course there is also significant overlap between the two extremes of communities of 100% self-supplying water and 100% obtaining water from a public supply. Figure 3 shows this distribution as it appears in 1990 Census block groups.

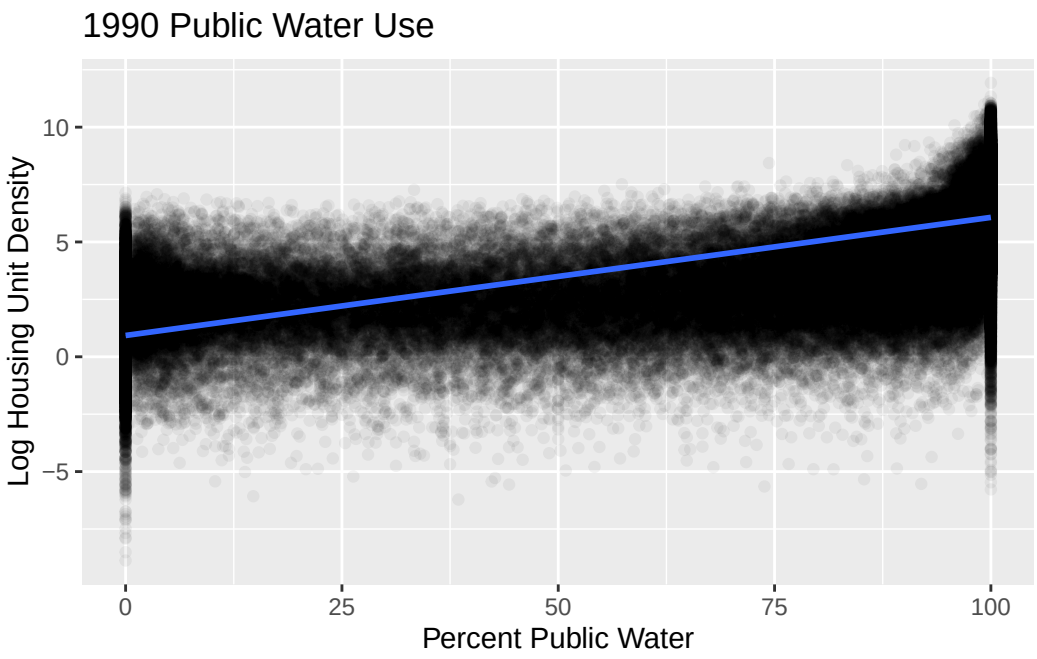

Figure 3: Plot illustrating that as housing unit density increases, so does the reliance on public water sources. Sourced from: 1990 Census.

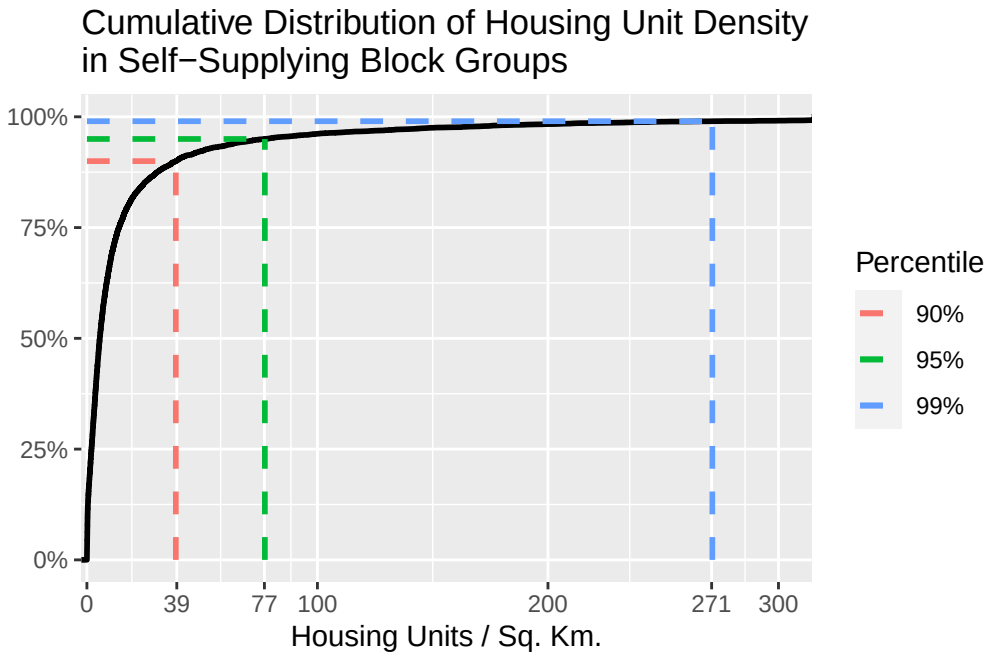

Figure 4: The cumulative distribution function shows that 99% of block groups that are 100% self-supplying water are in areas with less than 271 housing units per sq. Km.

Figure 4 illustrates the percentiles of housing unit densities which were completely self-supplying water in 1990. 95% of block groups that were completely self-supplying water had housing unit densities less than 77.2 Housing Units / km^2.

What's New in 2020

Using Census blocks from the start

The 2010 methods used a dasymetric mapping approach at the block group level to transform 1990 boundaries into 2010 boundaries. All analysis was done at the block group level. Once the analysis was completed, block groups were scaled to blocks using % housing units as weights. Our updated method scales to blocks in 1990 using an alternative weighting method and then crosswalks 1990 block boundaries to 2020 block boundaries. We believe that this reduces the error in transforming boundaries across time. Since census blocks are based on physical boundaries as opposed to block groups which are delineated based on population, they have less spatial change over time. We also believe the alternative weighting method improves the spatial allocation of wells to blocks. This is explained further in the methods section.

Machine Learning

The 2010 methods use a simple linear regression based approach. We show that this relationship works up to a point but is compromised in areas that experience high housing unit growth. This is especially apparent in areas of urban sprawl around large cities. To accommodate these complex factors, we have developed a method which uses a decision tree to first estimate where census blocks have access to public water systems, validated using known public water system boundaries. We then use a random forest model, validated by state well drillers logs to estimate the percent of housing units using wells at the edges of public water systems. These methods help to account for more complex relationships in communities like development and distance to a known public water system.

Data

Census Data

The 1990 long-form version of the Census was the last time a comprehensive national survey was taken on where people were obtaining household domestic water:

Do you get your water from –

- A public system such as a city water department, or private company?
- An individual drilled well?
- An individual dug well?
- Some other source such as a spring, creek, river, cistern, etc…?

The long-form version of the census was sent to 20% of the population and used to create sample-based estimates at the census block group level.

Census data were obtained from IPUMS NHGIS, University of Minnesota and use the following tables:

- 1990 Census: STF 1 - 100% Data (1990 Housing Unit Counts)
- 1990 Census: STF 3 - Sample-Based Data (1990 sample counts of private & public water use)
- 2020 Census: P.L. 94-171 Redistricting Data Summary File (2020 Housing Unit Counts)

We also used the following block to block crosswalk files from NHGIS:

- nhgis_blk1990_blk2010_gj.csv

- nhgis_blk2010_blk2020_gj.csv

Drillers Logs

Well Drillers Logs were sought and obtained from as many states as possible, as in our previous 2010 methods. Reporting requirements, reporting dates, use definitions and data availability vary from state to state. Figure 5 shows the counts of domestic wells by state

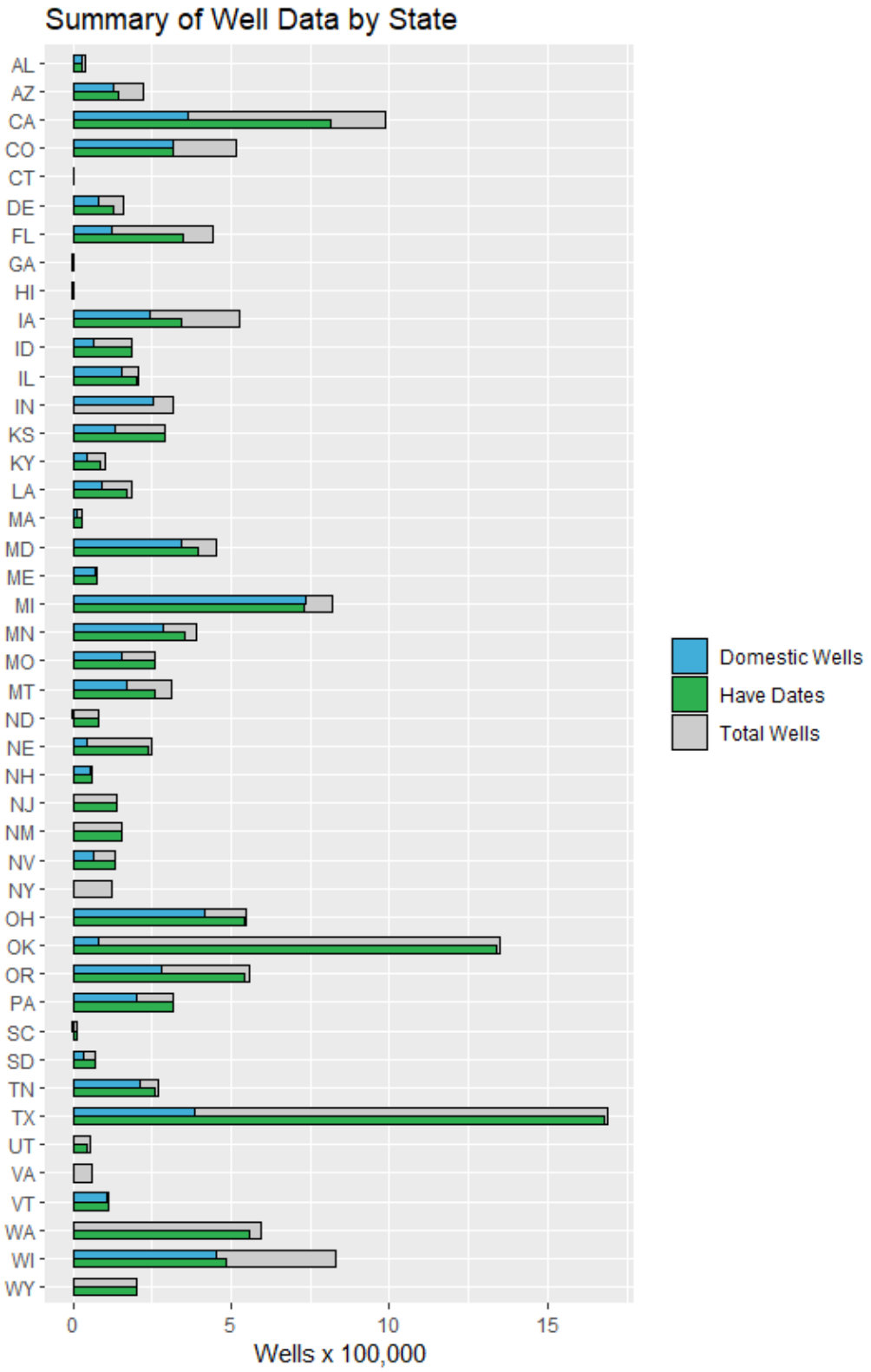

Figure 5: Plot showing number of obtained drillers logs by state, further broken down by the number that can be considered as domestic and whether they have an associated installation date.

A lot of investigation and background information is needed to assess whether state well records can be considered to be comprehensive over the time period between 1990 and 2020. It is well understood that reporting varies state-to-state based on several factors such:

- When reporting requirements were put into place (before or after 1990)
- How states collect their records
  - Some states delegate permitting to the county and they are then aggegrated to the state
  - Some states have multiple agencies permitting wells depending on well type
  - Many states began with paper records and have since switched to digital records with varying levels of completeness for older records.
- Under-reporting is known to frequently occur (permits are not properly filed)

Figure 6 shows the counts of domestic wells installed by year since 1990 for each state we were able to collect data for. These plots allow us to spot anomalies in reporting. For example, the spike in South Carolina may indicate that dates are not referencing installation but instead may be the date a drillers log was digitized. We can also see where states began to increase their reporting, as in Texas around 2003. Figure 7 shows the regression of housing units versus installed domestic wells by state. The red boxes/rings in both figures denote states that we believe are outliers, and that well records from those states are likely not comprehensive reflections of actual domestic wells installed.

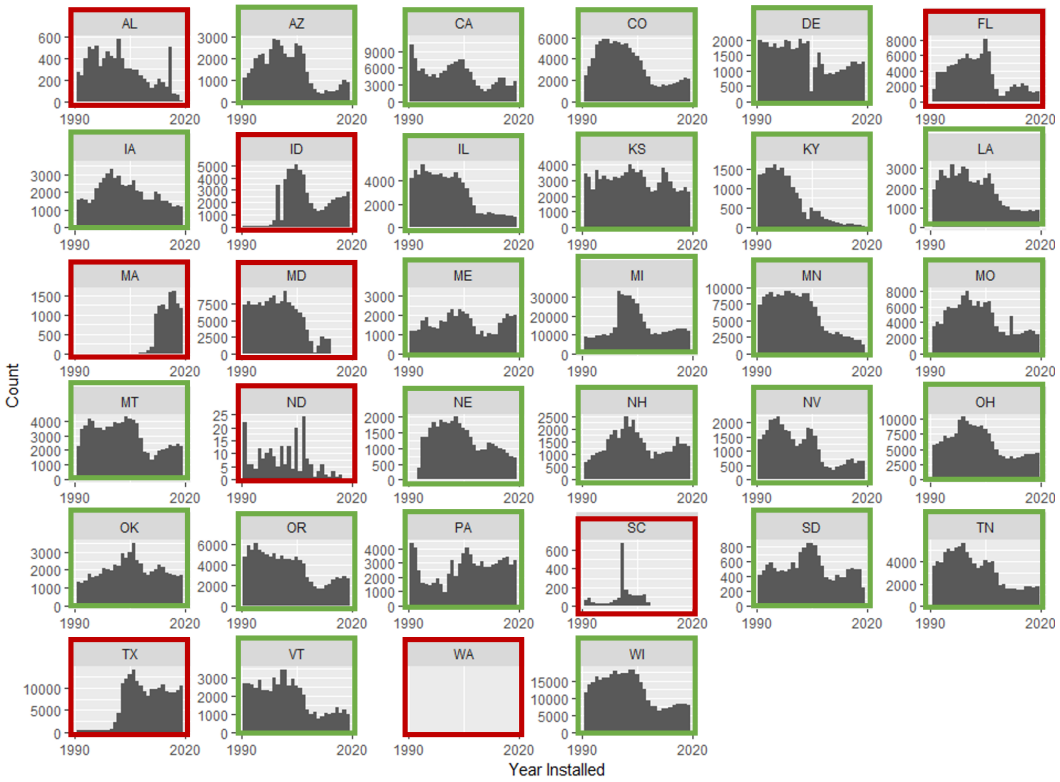

Figure 6: State plots showing the number of domestic wells installed by year since 1990. Red boxes are drawn around states with insufficient data.

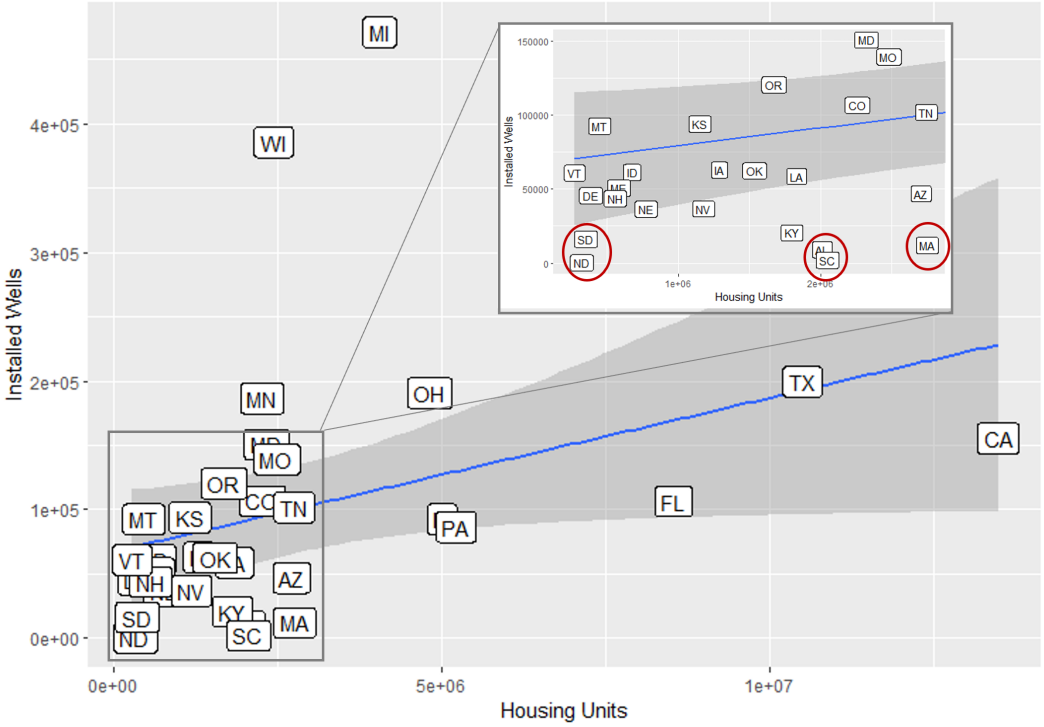

Figure 7: Regression plot showing housing units versus number of installed wells by state. Red rings are drawn around states with insufficient data

Data from states that we considered to have comprehensive domestic well records for the period between 1990 and 2020 are used to train a second decision tree model to estimate the percent of well use in census blocks we classify as having a combination of public and private water sources. This approach is explained further in the methods sections.

Public Water System Data

Our decision tree classification is trained and validated using public water system boundaries from states that we consider to have accurate and precise water service boundaries. This includes:

- California
- Connecticut
- New Jersey
- Washington

Data was sourced from respective state agencies and is available via the links in Table 1

State	Reference Year	link
California	2023	LINK
Connecticut	2022	LINK
New Jersey	2022	LINK
Washington	2021	LINK

Table 1: Water Service Boundary Sources

Urban Imperviousness

Urban imperviousness data was obtained from the 2021 release of the National Land Cover Database for the conterminous United States. For Alaska, the 2016 developed urban descriptor was used. For Hawaii, the 2001 percent developed imperviousness layer was used. These three layers represent the most recent releases for their respective areas. A visual inspection of Hawaii and Alaska showed very close alignment with recent satellite imagery.

Methods

Cascade Weighting

Prior work has used the 1990 census data as a baseline and performed projections at the block group level (e.g. Murray et al (2021), Johnson & Belitz (2019). Once estimates were made at the census block group level, various methods were used to increase the resolution to Census blocks. Our previous work used a universal housing unit weighting technique where the estimated number of wells in a block group were placed into their corresponding blocks based on the percent of housing units in each block. Johnson and Belitz simply removed blocks with zero population. While both methods accomplish some sort of increased resolution, both fall short of true block level estimates. Housing unit weighting can improperly place wells into dense blocks where wells are unlikely to be. Simply removing zero population blocks likewise does not eliminate densely populated blocks where wells are unlikely to be.

We propose a different weighting method and perform it before any further analysis which we believe offers several benefits. We refer to this method as a 'cascading density weight method' (CDW). One advantage of weighting 1990 block groups into 1990 blocks is that it simplifies and improves cross-walking the data to 2020 boundaries. Census blocks are drawn using physical boundaries as opposed to population based delineation (used for block groups and tracts) and as they are smaller are less prone to errors when cross-walking over time. The CDW method takes the blocks within a block group and ranks them based on housing unit density. We then calculate the estimated number of housing units in a block group that were using public water in 1990 by multiplying the number of housing units (100% count) by the sample-based percent of housing units using public water. We then assign public water users to blocks, starting with the most densely populated block and then cascading them into the next most dense blocks. Once all of the public water users have been assigned, the remaining housing units are assigned as 'self-supplying'.

Figure 8 shows an example block group that had a density of 77 housing units /km^2 in 1990 and had a mix of public and self-supplied water use. panel B shows the housing unit densities of the blocks that make up the block group.

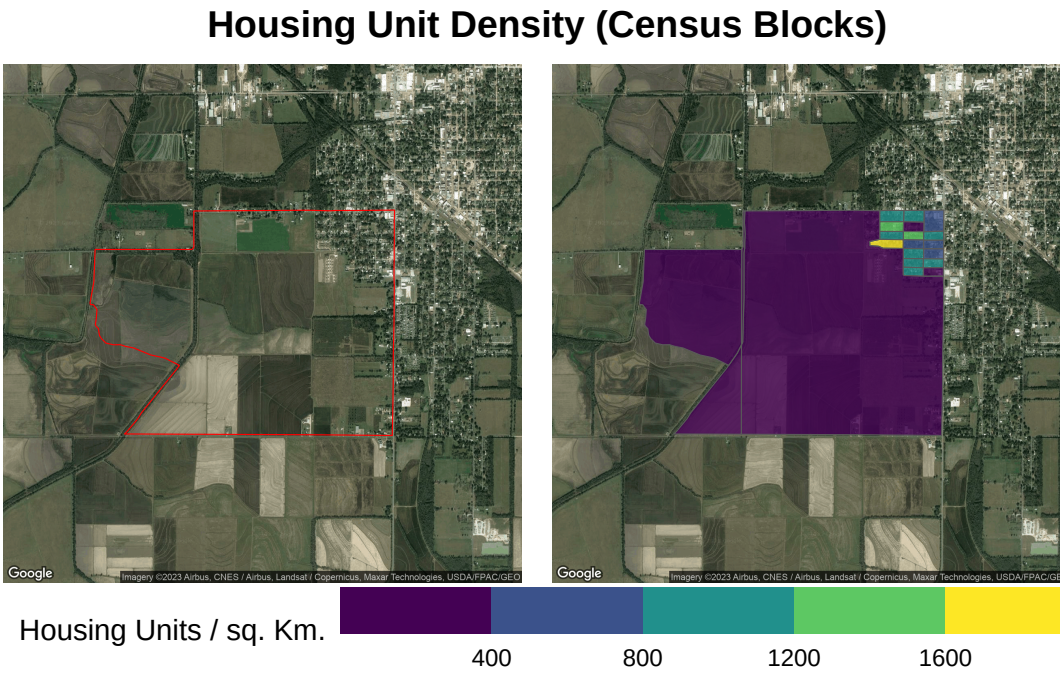

Figure 8: An example of a Block Group with 77 housing units per square kilometer and a mix of public and private water supply in 1990.

Visually, we can see where we would expect the public water system to be for this block group, in the northeast corner, near the town center. The CWD method reflects this in it's distribution of public and private water users to blocks. Blocks are ordered from most to least housing unit density (Table 2), then we iterate through the ordered blocks to assign 1990 water use. Table 3 shows the resulting table with estimated housing units using public water supplies allocated to census blocks.

GISJOIN	HU_90	Area_Km	HU_Density
G22005309806208	46	0.027	1685.0
G22005309806206	27	0.017	1546.1
G22005309806204	30	0.024	1247.0
G22005309806217	18	0.017	1032.5
G22005309806215	19	0.019	1021.4
G22005309806214	18	0.019	967.5
G22005309806207	19	0.020	963.7
G22005309806213	20	0.021	950.6
G22005309806202	23	0.024	943.3
G22005309806203	25	0.028	893.3
G22005309806211	15	0.018	848.2
G22005309806210	14	0.018	766.9
G22005309806212	16	0.021	765.3
G22005309806209	14	0.018	761.2
G22005309806201	20	0.044	458.6
G22005309806216	7	0.018	393.9
G22005309806218A	137	4.287	32.0
G22005309806219B	1	1.030	1.0
G22005309806205	0	0.020	0.0
G22005309806218B	0	0.297	0.0
G22005309806219A	0	0.026	0.0

Table 2: A table showing the blocks ordered by descending housing unit density

Code to Assign Public Water Use to Blocks:

```
# Number of public water users in 1990 block group
pwu <- 465

# Set default to 0
order$Public_S_B <- 0

# Iterate through blocks
for(n in 1:(nrow(order))){
              # Add up to the number of housing units in a block
              if(pwu > order$HU_90[n]){
                order$Public_S_B[n] <- order$HU_90[n]
              } else(order$Public_S_B[n] <- pwu)
              pwu <- pwu - order$Public_S_B[n]
            }
```

GISJOIN (Census Block)	HU_90 (Housing Units)	Area_Km (Km^2)	HU_Density (HU/Km^2)	Public_S_B (Public Water)
G22005309806208	46	0.027	1685.0	46
G22005309806206	27	0.017	1546.1	27
G22005309806204	30	0.024	1247.0	30
G22005309806217	18	0.017	1032.5	18
G22005309806215	19	0.019	1021.4	19
G22005309806214	18	0.019	967.5	18
G22005309806207	19	0.020	963.7	19
G22005309806213	20	0.021	950.6	20
G22005309806202	23	0.024	943.3	23
G22005309806203	25	0.028	893.3	25
G22005309806211	15	0.018	848.2	15
G22005309806210	14	0.018	766.9	14
G22005309806212	16	0.021	765.3	16
G22005309806209	14	0.018	761.2	14
G22005309806201	20	0.044	458.6	20
G22005309806216	7	0.018	393.9	7
G22005309806218A	137	4.287	32.0	134
G22005309806219B	1	1.030	1.0	0
G22005309806205	0	0.020	0.0	0
G22005309806218B	0	0.297	0.0	0
G22005309806219A	0	0.026	0.0	0

Table 3: A table of estimated housing units using public water in 1990 allocated to census blocks.

Once we have allocated the public water users to the census blocks, we can classify the census blocks as being served by public water or not. For example, Figure 9 shows our example block group, with 465 housing units served by public water, broken down into blocks and classified by whether they are estimated to be completely using public water, completely self-supplying water or having a combination of sources.

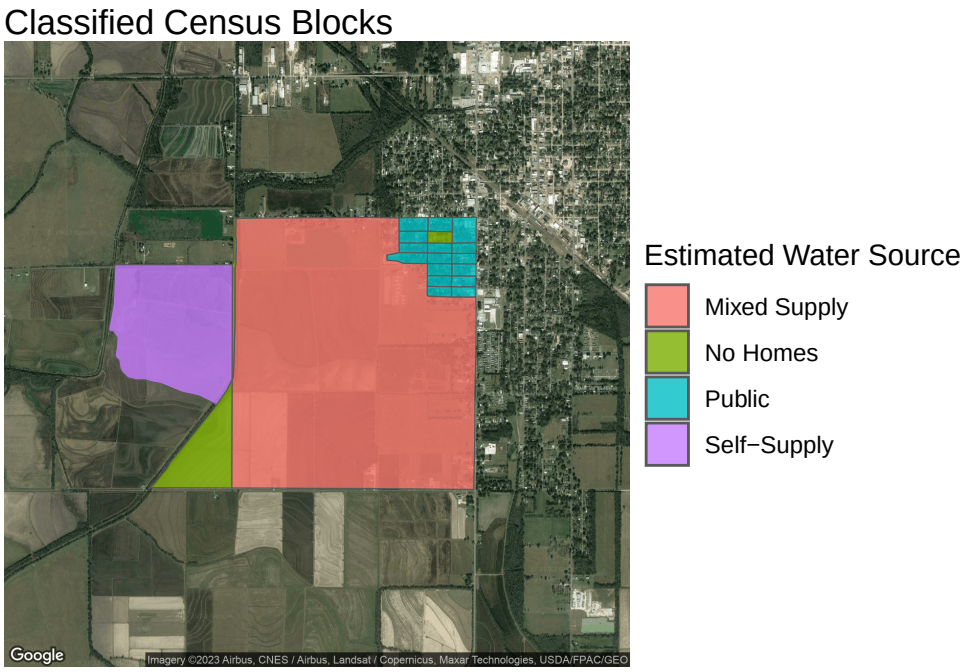

Figure 9: A map showing census blocks based on whether they are estimated to be served by public water, private supply, or a combination of both.

Cross-Walking Data

Once the 1990 data have been refined to census blocks, they are cross walked to 2020 boundaries using the geographic crosswalks from NHGIS.

Decision Tree (Public Water Access)

Using the tidymodels package, we trained a decision tree model (Figure 10) to predict if a census block was served by public water or not. To validate the tree, we use service area boundaries from California, Connecticut and New Jersey, which we deemed to be the most robust and precise of any publicly available statewide data sets, which were carefully reviewed. The decision tree was trained on a random sample of 75% (399,491) of the census blocks across the three states, stratified by whether or not their centroids fell within the boundary of a state-provided service area boundary. The 25% of blocks that were not selected were used to validate the decision tree after it was trained. A secondary validation was performed using Washington public service area boundaries, which were likewise deemed to be robust and precise. The value of a secondary validation is that the service areas come from a completely different source than anything used to train the model. The variables included in the training of the decision tree are listed in Table 4. The most important variable for predicting public water was percent imperviousness (Figure 11), reflecting that highly developed areas are most likely to be served by public water systems. The final decision tree model predicted the type of water supply correctly in 93.14% of blocks in the testing dataset which was removed prior to training and predicted 90% of blocks correctly in Washington (Figure 12).

Variable	Description	Source
Miles to Intake	Calculated Distance from Census Block centroid to closest public drinking water intake	EPA SDWIS

Variable	Description	Source
% Impervious	Mean Impervious value for census block as derived from 2021 NLCD Impervios layer	*NLCD (2021, 2016 & 2001)
HU (1990)	Count of housing units in 1990.	1990 Census
HU/Km (2020)	Housing unit density derived as 2020 housing units divided by land area (km)	2020 Census
Land Area (Km)	Land area in square kilometers.	2020 Census
% Public (1990)	% of housing units estimated to be using public water in 1990	1990 Census
% Sewer (1990)	% of housing units estimated to be using public sewers in 1990	1990 Census
% HU Change	% Housing unit change between 1990 and 2020	1990 & 2020 Census

Table 4: Table of input variables used to train final decision tree to predict public water use for census blocks in 2020.*The most recent imperviousness data is 2016 for Alaska & 2001 for Hawaii

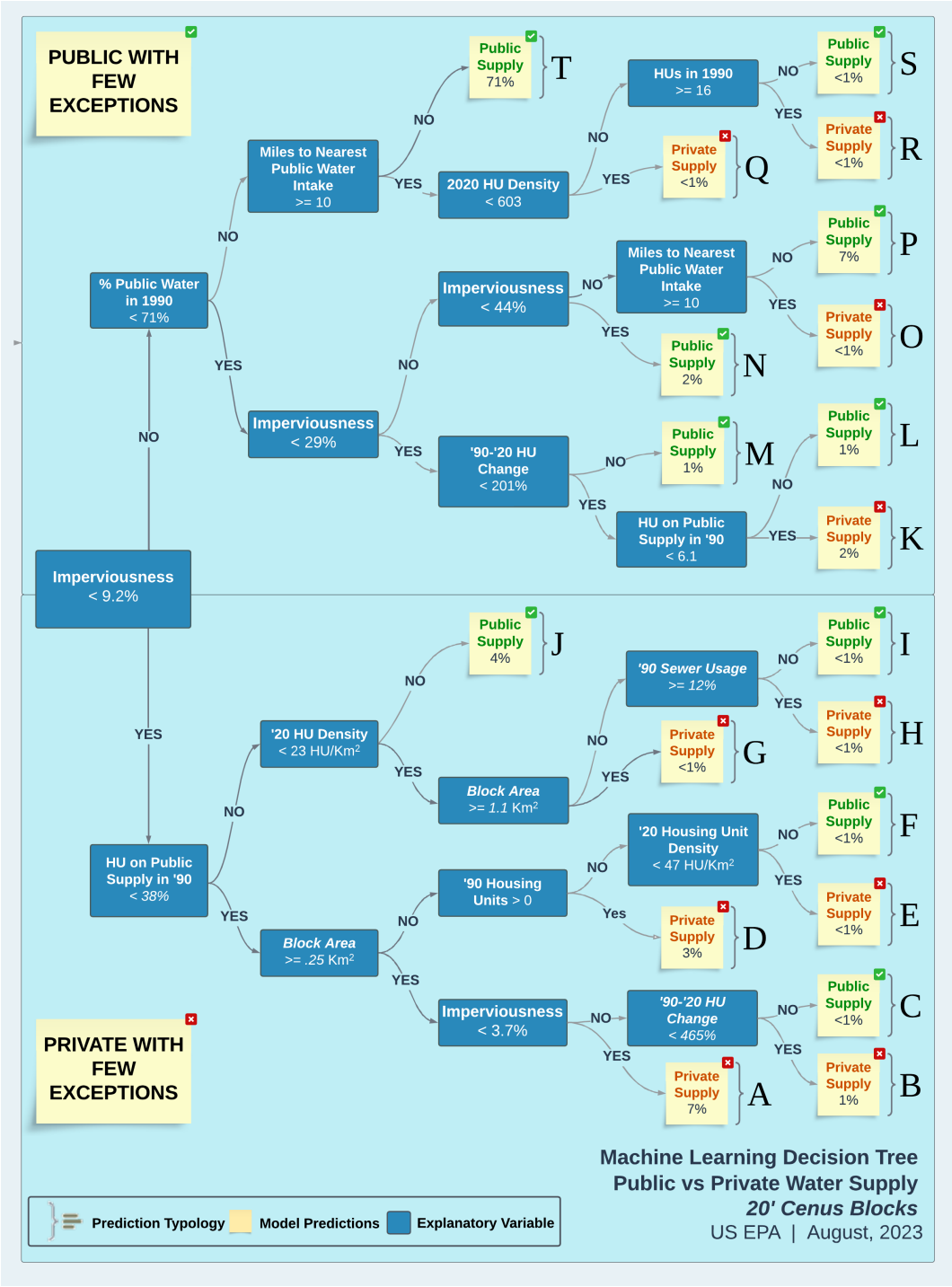

Figure 10: Decision Tree Result

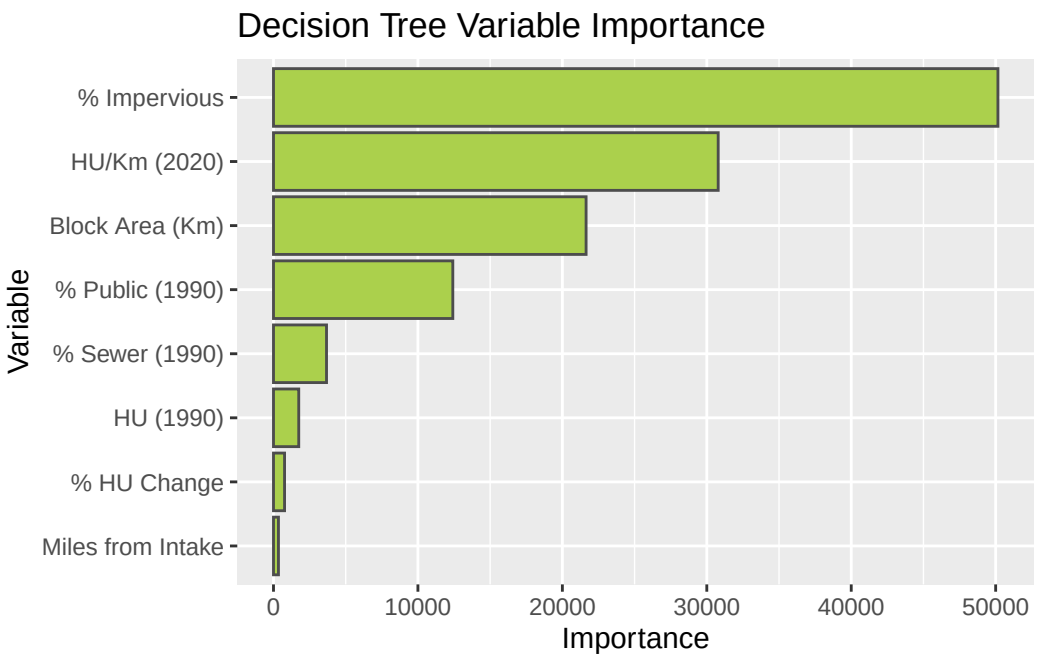

Figure 11: Plot of relative importance of input variables in decision tree

		OOB Testing		Washington	
		Truth			
		Private	Public	Private	Public
Predicted	Private	15432	5436	20787	6287
	Public	3699	108598	4967	80200

Figure 12: Table presenting performance results for decision tree when validated against testing data and seperate dataset from Washington. Green cells indicate blocks that were predicted correctly. Red blocks indicate blocks that were predicted incorrectly.

Reclassification

Once predictions are returned for blocks, we run them through two spatial reclassification operations. These operations are designed to identify certain blocks where we believe we can better classify water use based on their spatial properties relative to neighboring blocks as opposed to the decision tree result which looks at each block individually.

Reclassification Operation 1: Islands We identify blocks as 'islands' if they have an opposite classification from every other block that it touches as shown in Figure 13. Upon inspection of blocks identified as islands, the most common occurrence related to a census block in an urban area that was misclassified as being on private water. This can happen when a block only has a few homes in it, or if it contains urban green space. For example, a census block that encompasses a park within a city, but also contains a couple of homes

may appear to the model as a rural area. When islands are identified by the reclassification, their classes are flipped to match their neighbors.

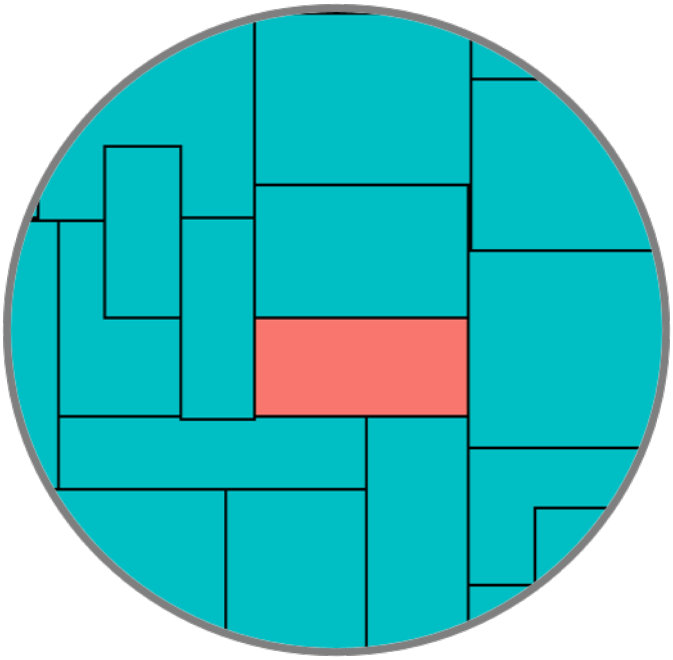

Figure 13: Example of a block identified as an island

Reclassification Operation 2: Urban Centers The second reclassification operation focuses on areas that could not be predicted by the decision tree. These are typically blocks with no housing units. In large urban areas, we typically see business districts where there are no permanent residents but are covered by a public water system. While this step is not essential for identifying areas with private domestic wells, it is necesarry to identify complete public water systems, which applies to future work. This reclassification is done by identifying blocks with NA values that border blocks with a known classification as in Figure 14. Starting on the edge where 'NA' blocks touch classified blocks, we spiral into the center to reclassify these blocks to match the surrounding area.

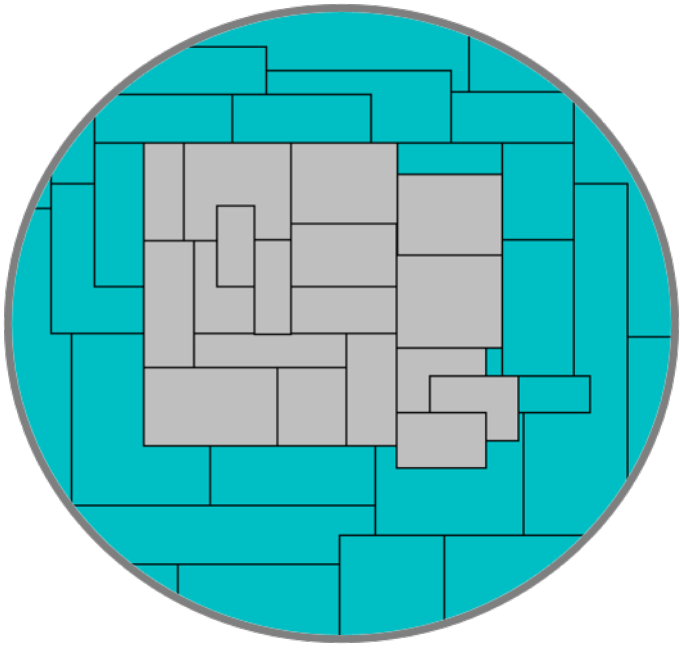

Figure 14: Example of a group of urban blocks with no housing units, likely to be on a public system.

The third reclassification method we use is a simple identification of blocks based on their 1990 estimated status. If we estimate that 100% of the housing units in a block were using public water in 1990 based on the census, we disregard their decision tree result and keep them at 100%. While there are examples of blocks changing from public to private supply, these are rare and public water systems in general have been expanding in the United States for decades. Once a community is connected to public water, it is unusual for it to switch back to self-supply. Figure 15 illustrates the percent of blocks in the validation data that reported 100% public use in 1990 but did not fall within a service boundary area.

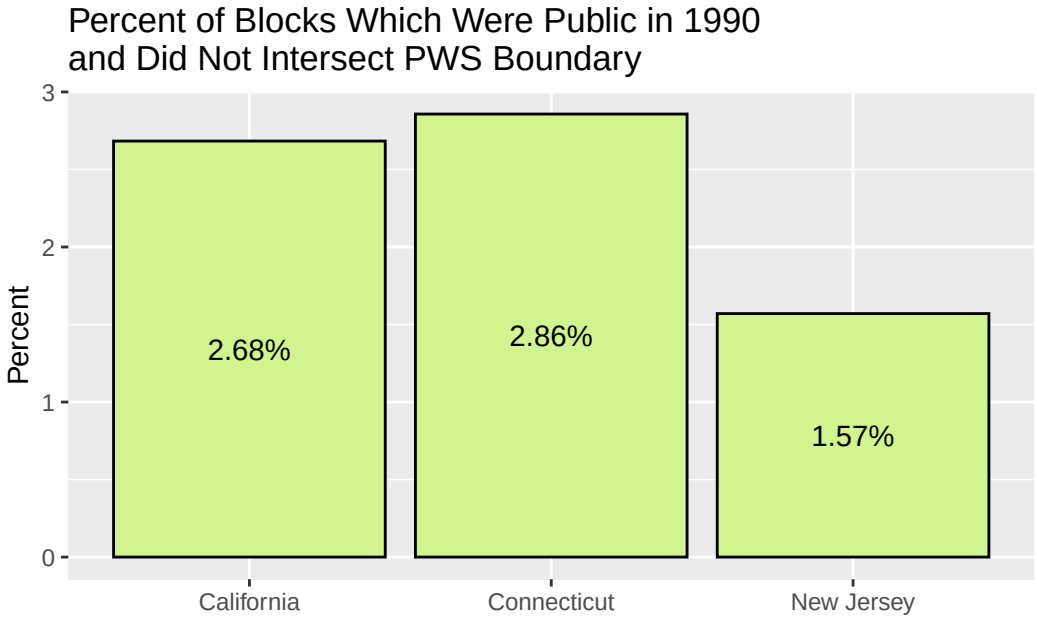

Figure 15: Bar plot showing the percent of census blocks estimated to be 100% on public water in 1990 but whose centroids did not fall within a service boundary in the validation data.

Random Forest (% Well Use)

Following the reclassification of the decision tree results, there are two classes which every block with housing units falls into:

- Private: 100% private use
- Public: At least some access to public

Almost all census blocks are either entirely on public water or entirely self-supplying water. Using our weighting method, we estimate only about 2% of census blocks have a mix of public and private supply (Figure 16). To better estimate wells within blocks that may have a combination of supply types, we identify block groups that have a combination of blocks classified as public and private. These block groups represent the public-to-private gradient that occurs as you move from urban to rural areas, where we observe the outer edges of public service boundaries. To estimate the number of wells in these blocks, we generate a random forest, identical in it's predictor variables to the decision tree model but with a different outcome. Whereas the outcome of the decision tree was whether or not a block was served by a public water system, the random forest solves for the percent of housing units using domestic wells. To validate this model, we use drillers logs from 25 states, as described in the data section. We execute the RW method as presented in Murray et. al (2021) by using the 1990 data as a baseline and geolocated wells from drillers logs. For drillers logs that were highly accurate (were located with GPS), a simple spatial intersection was performed on census blocks. For wells that were less accurate, such as those referenced to PLSS sections, we performed a spatial join to block groups and repeated the cascade weighting method to estimate block locations.

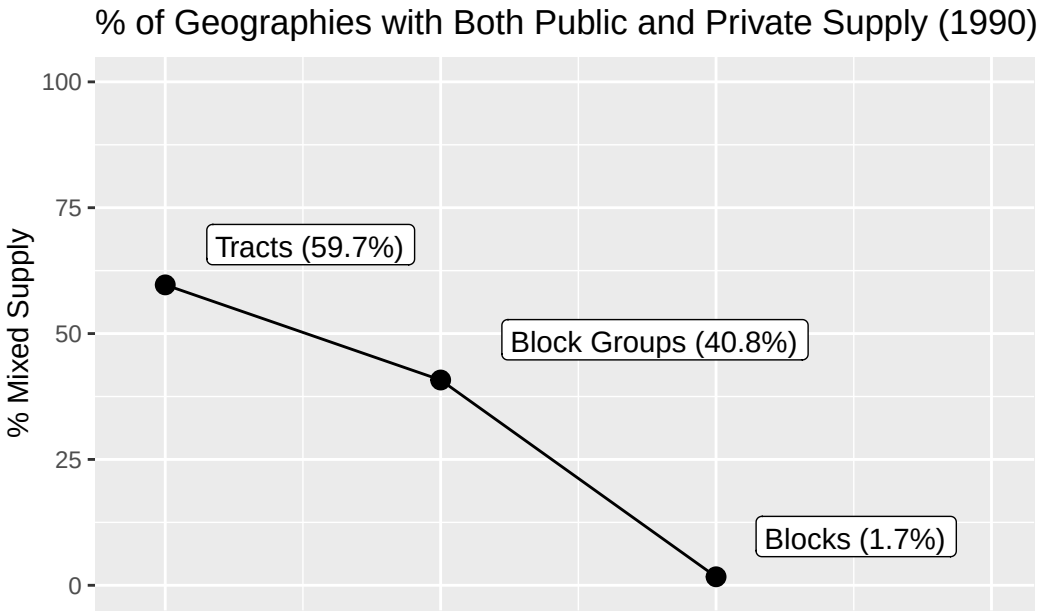

Figure 16: Plot of percent of Census tracts, block groups and blocks that had mixed supply sources in 1990. As geographies bet smaller, percent mixed source decreases.

Results

2020 Well Estimates

We estimate that a total of 22,538,261 housing units were self-supplying domestic water in 2020. This equates to roughly 22.3 million wells and 275,000 other types of supply (cistern, spring etc…). This equates to roughly 16% of housing units in the United States self-supplying water, which is a slight decrease from our 2010 estimates of 16.6% and an increase from 1990 (14.5%). While the percent of well users decreased from our 2010 estimates, the total number increased from 21.8 million to 22.5 million. This result was expected due to our suspected overestimation of well use in 2010 based on developing areas which we sought to correct in our updated methods.

Figure 17 shows the breakdown of percent well use by state for the 1990 Census, 2010 and 2020 estimates. The 2010 estimates are from our previous methods (NHU and RW).

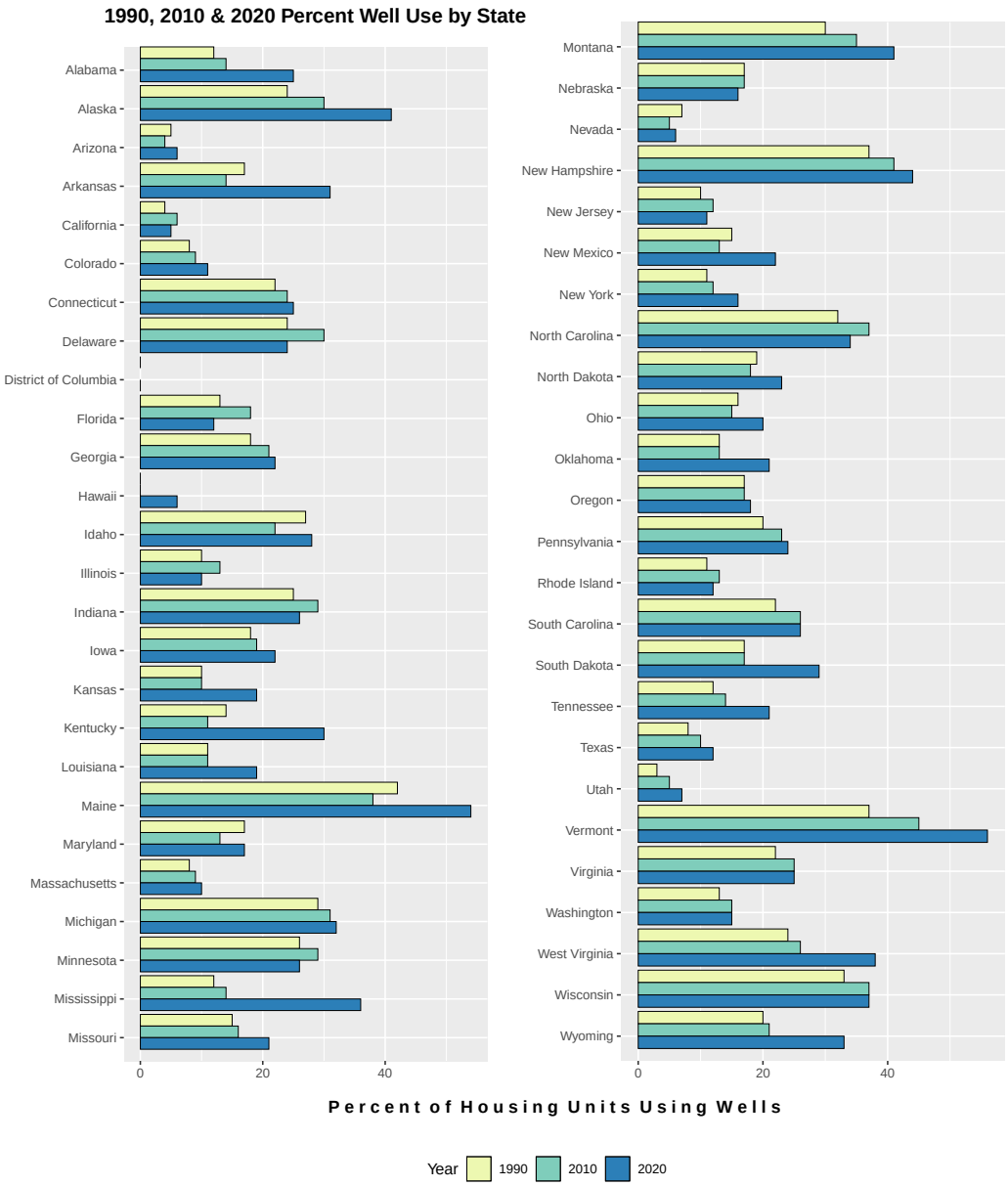

Figure 17: Bar plot showing state by state comparisons of 1990, 2010 and 2020 estimates of wells.

Spatial Improvements

As noted previously, we expect that increasing the resolution to blocks before building the decision tree will increase the spatial accuracy of block level estimates for 2020. Figure 18 shows an area in California with the original 2010 estimates using the NHU method and our new 2020 estimates overlaid by public service area boundaries from the California validation set.

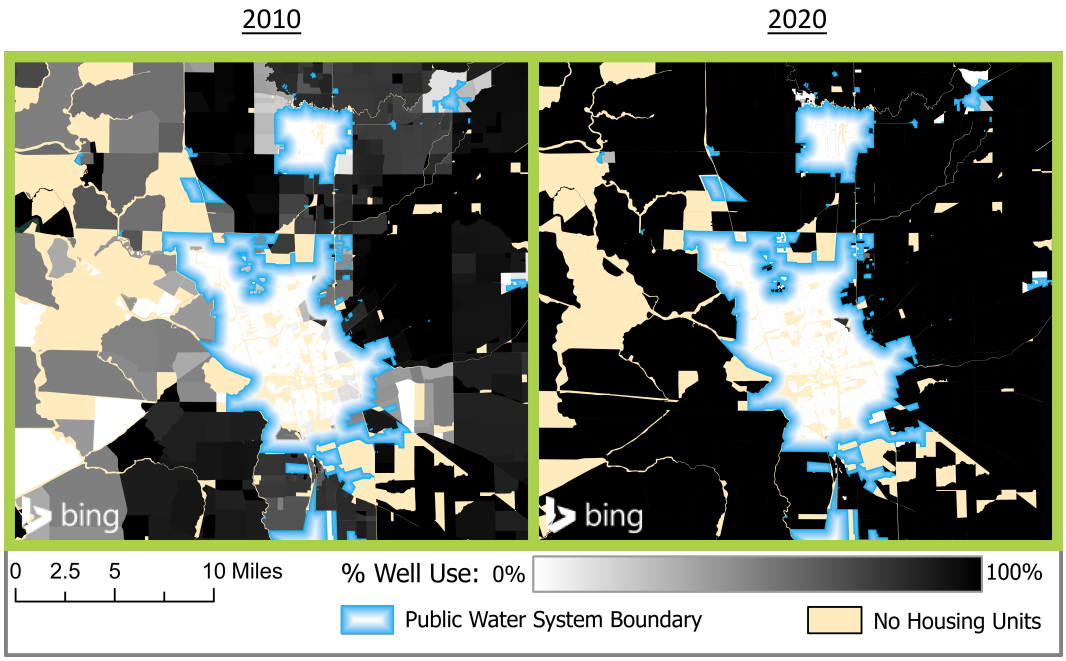

Figure 18: 2010 estimates using the NHU method and our new 2020 estimates overlaid by public service area boundaries from the California validation set.

Additional Resources

Code Repository

All code used to generate well estimates is available via the EPA Github repository. (Link to be added.)