U.S. EPA National Sewershed Dataset & Model Documentation

Contents

This document serves as background material for the EPA National Sewershed Dataset. The dataset can be accessed via the EPA internal web application

1 Executive Summary

> "The Clean Watershed Needs Survey(CWNS) provides an assessment of the capital investments necessary for states, the District of Columbia, and U.S. Territories to meet the Clean Water Act's (CWA) water quality goals over the subsequent 20 years. These needs include projects and related infrastructure costs for wastewater publicly owned treatment works, stormwater treatment, nonpoint source control, and decentralized wastewater treatment. The U.S. Environmental Protection Agency (EPA) has prepared the 2022 CWNS Report to Congress in compliance with CWA section 516(b)(1)(B) (33 U.S Code §1375) as well as CWA section 609, which was added by the Infrastructure Investment and Jobs Act (IIJA), P.L. 117-58, November 15, 2021. This Report summarizes the results of EPA's 17th survey since the CWA was enacted in 1972."

As part of this survey, state-level teams submit technical information about publicly owned treatment works (POTWs). Some POTWs may discharge a portion of their wastewater downstream to other POTWs. When a treatment plant discharges less than 50% of its wastewater to another POTW, we refer to it as an 'endpoint', meaning that it is considered the final destination for wastewater prior to treatment and then discharge back into the environment, most frequently via direct discharge into a water body such as a river or ocean. The entire sewered area that contributes wastewater to a specific endpoint is referred to as a 'sewershed'. A sewershed may therefore include a single POTW and collection system or multiple POTWs and their collection systems. The relationship between POTWs and endpoints can be derived from CWNS reporting by accounting for reported discharge of wastewater between facilities.

Sewershed extents are not a part of reporting within the CWNS and are otherwise not statutorily required under the CWA. Understanding the sewered area related to an endpoint is critical to support the CWA, as well as multiple other agency and administration priorities. Sewershed boundaries can, for example, facilitate wastewater surveillance efforts and link community disease tracking to affected populations. The ability to map sewer systems can also be critical to infrastructure and urban planning, providing information on areas that can be developed or could benefit from resources to aid in economic development and information on infrastructure gaps.

EPA previously developed geospatial data for community water system service areas for the United States. The development of sewershed boundaries was built off of this experience, using

machine learning models which are trained on existing sewershed data and utilizing advanced geospatial techniques to account for the complexities of determining what areas of the United States are sewered and which treatment plants they are served by. The output of this dataset is a spatial dataset of polygons that are constructed from hexagons with a resolution of ~0.11 km^2. In total, 17,084 sewersheds are present in the data, with 3,192 obtained from a variety of public sources and 13,892 modeled using machine learning.

2 Introduction

The United States Environmental Protection Agency (EPA) has previously released a national dataset of community water system service areas [3]. Building on that work, EPA is now releasing a companion 'sewershed' dataset, which is the first national dataset of sewered service areas for the United States [4]. While there are many similarities between public water and wastewater infrastructure, there are several key differences. For example, there are more than twice as many community water systems in the United States as there are publicly owned treatment works (POTWs). In general, sewer systems represent a larger investment in infrastructure than public water service. Water supply lines are pressurized throughout the system and can more easily overcome changes in topography, whereas sewer largely relies on gravity for wastewater flow. Where gravity must be overcome within wastewater systems, pumping or lift stations must be installed. Sewer lines are also larger than water supply lines and come with increased construction costs. It is not uncommon for a suburban or rural home to have a public water connection for supply and a septic system for wastewater. However, having access to a sewer and no access to public water occurs far more infrequently. This suggests that sewer systems are more frequently located in more urbanized areas.

The previous model for community water systems utilized census blocks, supplemented with land parcels as the basic building block for service areas. We have changed the spatial unit for this model to H3 hexagons [6]. H3 hexagons are a global grid system that have several key advantages over census boundaries such as having uniform sizes, which allow for more consistent spatial resolution and more efficient computation.

This document provides detail on how public data was collected, how the model was developed and discusses the resulting national dataset of sewersheds and how it can be used to support EPA's mission.

3 Definitions

Term	Definition
CWNS	Clean Watershed Needs Survey

Term	Definition
Endpoint	The point of convergence for all wastewater collected within a 'sewershed'
Sewershed	The area that contributes wastewater to a specific endpoint
Utility	A public or private entity that provides wastewater treatment services
POTW	Publicly Owned Treatment Works
Sourced Sewershed	A sewershed that has been obtained from a publicly available source as opposed to having been modeled.

4 Existing Sewershed Data

The goal of this project was to publish a comprehensive dataset of sewershed boundaries in the United States that can be linked to the CWNS. To that end, we sought to include as many publicly available sewershed areas as possible. Ideally, every boundary would be sourced from the utility level to provide the most accurate representation of sewersheds possible. However, this data does not always exist, which necessitates the creation of modeling approaches to fill in the gaps.

EPA conducted a rigorous search for publicly available sewershed data through online searches and engagement with stakeholders including federal and state agencies, regional partners, NGOs and academics working in the wastewater field. In total, we obtained 3,280 sewersheds from public sources which could be linked to POTWs reported in the 2022 CWNS (Table 2).

5 Methods

Three types of tree-based models were considered for this project: a decision tree, a random forest and boosted trees. All three have advantages and disadvantages that must be considered. A single decision tree is the simplest and most easy to interpret method, but can be overly simplistic and less accurate than more complex models. A random forest is a collection of decision trees with randomized input variables, which can improve accuracy and protect from over-fitting but is less interpretable and requires more computational power. Boosted trees are a more advanced method that build decisions trees in succession. Each new decision tree attempts to solve for error found in the previous tree. This can lead to even better accuracy, but at the cost of increased complexity and reduced interpretability. Boosted trees are especially effective at leveraging variables that may otherwise reflect lower importance levels in other methods. We found that the higher accuracy returned by boosted tree models justified its selection over the other two methods. For brevity, only the boosted tree model is discussed in detail.

Table 2: Sourced Sewersheds by State

State	Sourced Population	# Sourced	% Sourced	% Sewered Population Sourced	State	Sourced Population	# Sourced	% Sourced	% Sewered Population Sourced
Arizona	824K	7	6%	16%	Nevada	50K	3	5%	2%
Arkansas	130K	1	0%	7%	New Hampshire	669K	61	76%	91%
California	1,286K	26	5%	3%	New Jersey	8,396K	150	94%	99%
Colorado	877K	72	18%	15%	New Mexico	143K	2	2%	10%
Connecticut	2,804K	80	87%	92%	New York	16,501K	595	96%	100%
Delaware	562K	7	41%	71%	North Carolina	3,467K	225	69%	51%
District of Columbia	2,016K	1	100%	100%	North Dakota	0K	1	0%	0%
Florida	8,424K	153	41%	49%	Ohio	5,173K	50	6%	56%
Georgia	188K	16	4%	3%	Oklahoma	128K	1	0%	4%
Hawaii	82K	5	23%	8%	Oregon	819K	5	2%	21%
Illinois	1,807K	37	5%	16%	Pennsylvania	431K	39	5%	4%
Iowa	750K	2	0%	28%	Rhode Island	732K	16	80%	94%
Kansas	228K	8	1%	10%	South Carolina	1,750K	29	18%	47%
Kentucky	2,042K	228	87%	70%	Texas	5,647K	470	33%	23%
Maryland	4,099K	148	90%	99%	Utah	3,319K	113	100%	100%
Massachusetts	5,597K	123	97%	100%	Vermont	323K	92	98%	100%
Michigan	2,844K	1	0%	39%	Virginia	1,819K	10	4%	31%
Minnesota	3,021K	9	2%	64%	Washington	4,835K	42	16%	76%
Mississippi	71K	13	5%	5%	West Virginia	797K	164	55%	66%
Missouri	5,490K	100	12%	78%	Wisconsin	2,108K	72	12%	48%
Montana	115K	10	6%	19%					

5.1 Defining the universe of sewersheds

Not all records within the 2022 CWNS represent endpoints for wastewater collection. The first step in our modeling effort is to establish the universe of systems. To do this, we use the following criteria to select endpoints:

- POTWs that treat human waste (as opposed to wet weather facilities)
- Total receiving population (the estimated number of people whose waste is treated directly by the POTW) > 0
- Facility type contains "treatment plant"
- Residential population (as opposed to non-residential, such as tourists or seasonal workers) > 0
- Percent of Discharge to another POTW $< 50\%$

The data compiled from the 2022 CWNS was reviewed for accuracy by EPA[1].

5.2 Matching sourced sewersheds to CWNS locations

Sourced sewersheds had to be matched with their corresponding CWNS endpoint locations as there is no common identifier between the two datasets. Matching was done through a combination of spatial intersections, distance measurements, text similarity between names and individual visual inspection.

Boundaries were spatially intersected with endpoints. For each match between a boundary and an endpoint, the total number of endpoints that intersect a boundary is divided by 1. For example, if only one endpoint exists within a boundary, the value would be 1. If 3 endpoints intersect a boundary, the value for all three matches (which are represented as three separate rows) would be 0.33. This score is labeled 'I_Score' or S_i

For boundaries that do not intersect any endpoints, a buffer is applied using a distance of 16 kilometers, which was found to be the 95% confidence level for maximum distance between an endpoint and a service area (See 'Analysis/Endpoint_Sewershed_Distance/ Endpoint_Sewershed_Distance.html' for more detailed information). Only endpoints that have not already been singularly intersected with boundaries are considered in this step. The distance of each endpoint from the edge of the boundary is divided by 16 and then subtracted from 1: $D_{score} = 1 - (d/16)$, where 'd' is the distance in kilometers. This yields a score such that the closest endpoint will have the highest score. This score is labeled 'D_Score' or S_d.

[1]The 2022 CWNS contained a separate list of "confirmed" POTWs which were treatment facilities that state teams could not confirm the technical data for, but did confirm their existence. After review, it was determined that it is **NOT** safe to assume all rows in 'FACILITIES_CONFIRMED.txt' are all end points. It is safe to assume all rows in 'POPULATION_WASTEWATER_CONFIRMED.txt.' are endpoints. rows identified as having 'CHANGE_TYPE' = 'New' should be removed.

For each pair of matches (CWNS <-> State Boundary), a fuzzy score was calculated to compare the given names of the treatment works. A score of zero indicates that the names are completely dissimilar whereas a score of 1 indicates identical names. This score is labeled 'F_Score' or S_f. The method we use is the 'Jaro-Winkler' distance:

"The Jaro-Winkler distance (method=jw, 0<p<=0.25) adds a correction term to the Jaro-distance. It is defined as d − l · p · d, where d is the Jaro-distance. Here, l is obtained by counting, from the start of the input strings, after how many characters the first character mismatch between the two strings occurs, with a maximum of four. The factor p is a 'prefix' factor, which in the work of Winkler is often chosen 0.1." Source

The Jaro-Winkler method uses 0 for identical and 1 for completely dissimilar. Therefore: $S_f = 1 - D_{JW}$ where D_{jw} = Jaro-Winkler Distance

- Scores are aggregated to yield a match score.
  - If 'I_Score' is > 0, then the formula we use is: $S_m = S_i + S_f$
  - If 'I_Score' $= 0$, we use the formula: $S_m = S_d + S_f$

This match score equally prioritizes the proximity of an endpoint with name matching between datasets. In the event that multiple endpoints intersect a boundary, greater weight is then given to name matching.

5.3 Spatial Resolution

Hexagons from the open source H3 geospatial indexing system were chosen as the spatial unit for the model for several reasons. Hexagons are roughly the same size across the United States, they allow for more consistent statistical calculations related to area and distance and density values do not need to be calculated. Hexagons also allow for more complex network calculations at higher computation rates because distance can be inferred as a constant between each neighboring pair of hexagons. Finally, hexagons offer an advantage to census blocks, which are delineated based on physical boundaries such as roads and rivers, which can include undeveloped areas, especially in rural settings. Using hexagons allows us to create more detailed sewersheds while reducing overall computation requirements. For this modeling effort, level 9 hexagons were used, which are ~0.11 km^2 in area.

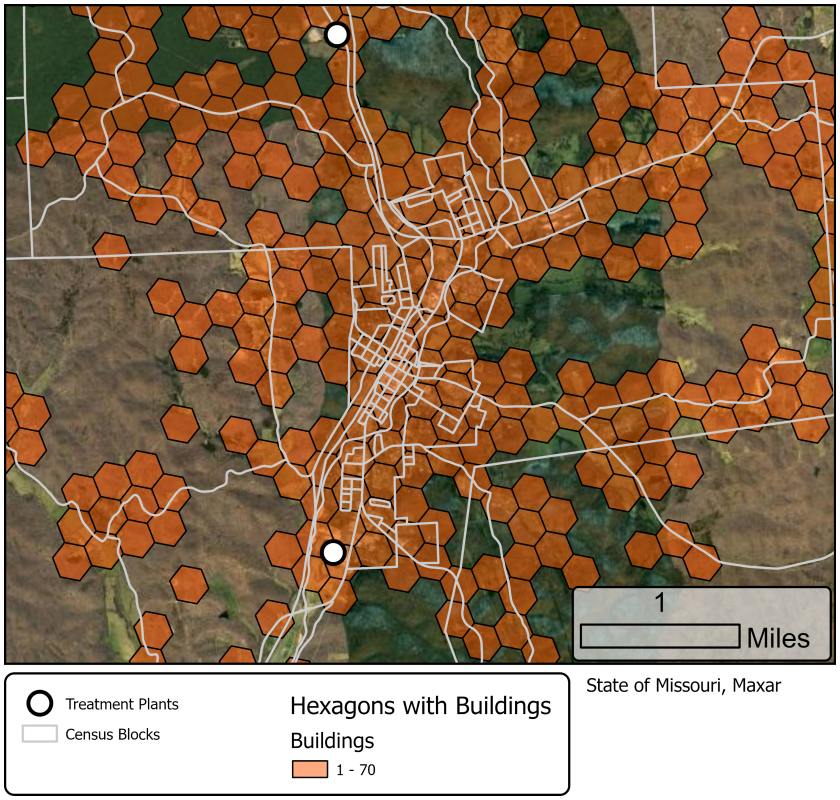

Figure 1: A satellite image of Piedmont, MO illustrating hexagons with buildings, overlaid by census block boundaries.

5.4 Model Development

The goal of the model is to predict whether a hexagon is sewered by a specific endpoint. Essentially the model is answering two questions. **1.** Is the hexagon connected to sewer? and **2.** If so, which endpoint is the most likely to serve that hexagon? The model is therefore a binary classification model, which returns a probability that a hexagon is part of the sewershed for a given endpoint location. The model considers each row of a table as a unique pairing between a specific hexagon and a specific endpoint. Additional data within each row are the 'features' which are predictor variables relating to physical, environmental and demographic variables unique to that pairing.

5.4.1 Defining Area of Consideration

To minimize computation time and increase efficiency, sourced sewersheds were used to test the maximum distance between an endpoint and the furthest possible hexagon that is served

by that endpoint. A cutoff of 32 km was used, which represents the 99th percentile of maximum distances (Figure 2). Hexagons outside this range for all endpoints were excluded from analysis.

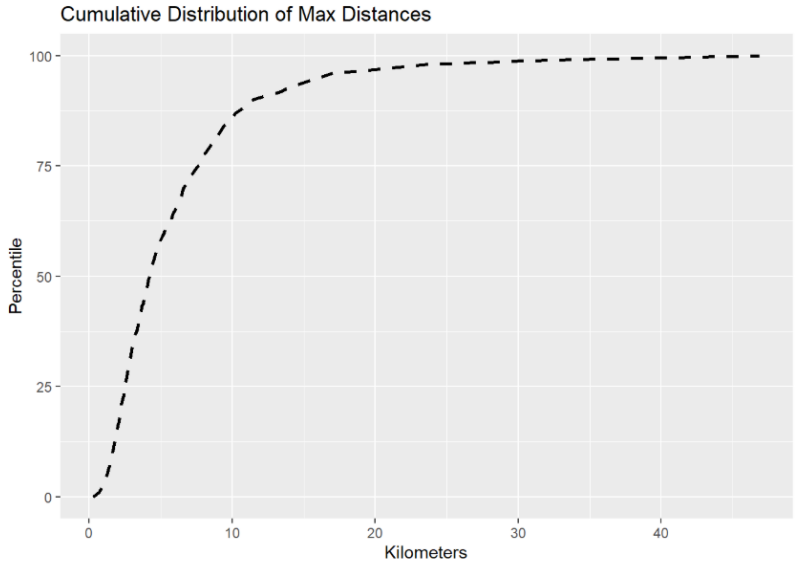

Figure 2: The cumulative distribution function illustrating the maxium distance a point can be from an endpoint and still discharge wasteweater to it.

5.4.2 Feature Engineering

In machine learning, 'features' refer to the input data that will act as predictors for the model. 'Feature engineering' therefore refers to the methods we use to prepare the data so that it contributes to a successful model. The primary source for the sewershed model is the 2022 CWNS, which includes the locations of POTWs that are considered 'endpoints', meaning that it is the final stop for wastewater before being discharged back into the environment. The sewershed is therefore the entire contributing area that discharges wastewater to each endpoint.

Sewered areas are typically developed, have relatively dense populations and are close to a treatment plant. Information fed into the model must provide relevant tools for relationships to be determined. For example, sewer systems, unlike public water systems, are generally not pressurized, meaning that elevation may be a factor. The distance between a hexagon and a treatment plant may vary depending on how many people the treatment plant serves. A hexagon with no population, but with several large buildings and very little green space may indicate an industrial zone.

The random forest model requires several input datasets, which will provide contextual information to help it determine if an area is served by a public wastewater collection system, and which collection system it is served by. These data must be prepared to conform to our output geospatial data, which are level 9 H3 hexagons. The hexagons are roughly 0.11 km2 in area. Here, we describe each input dataset and how it was developed to conform to the hexagon grid.

5.4.2.1 1990 Census Data

Data used from the 1990 census includes total population and housing units (1990 Census: STF 1 - 100% Data); estimated number of housing units on public sewer and estimated number housing units on public water (1990 Census: STF 3 - Sample-Based Data).

The 1990 Census data was cross-walked to 2020 boundaries using the cascade weighting method described in Murray & Hall (2024) and the block-to-block crosswalks published by the Minnesota Population Center [1]. Once cross-walked, the data was included with the 2020 Census data described in the next step.

5.4.2.2 2020 Census Data

2020 Census data included 100% counts at the census block level for population, housing units and count of urban/rural population. To convert census blocks into hexagons, a building weighted calculation was performed. Microsoft building footprints [2] larger than 40 square meters (about the size of a detached 2-car garage) were intersected with both hexagons and census blocks. The percent of buildings within each block were calculated for each hexagon and used to weight census data into hexagon parts. Hexagon parts, which were subdivided by census blocks were then re-aggregated to complete hexagons using the H3 Index for each. This yielded estimated Census counts at the hexagon level.

5.4.2.3 Building Footprints

When the previous step of weighting census data was performed and buildings were intersected with hexagons, a table was saved that included one row for every building, along with its associated hexagon, total area in km2 and its height as estimated by Microsoft [2]. Further predictive variables were derived from these data including:

- Mean building area
- Median building area
- Count of buildings
- Mean building height
- Median building height
- Maximum building height

For buildings where height was not available, we applied a value of 4.5 meters, which is roughly the height of a one-story building with a roof.

5.4.2.4 Land Cover / Land Use

The National Land Cover Database was used to derive the predominate land cover class and average percent imperviousness for each hexagon. Data was obtained in raster format from MRLC [5] and pixels were extracted to each hexagon. We determined the mode (maximum frequency) of each land cover type within each hexagon. Land cover classes that are not part of the four developed classes were binned into either 'water' or 'rural/other' (Figure 3)

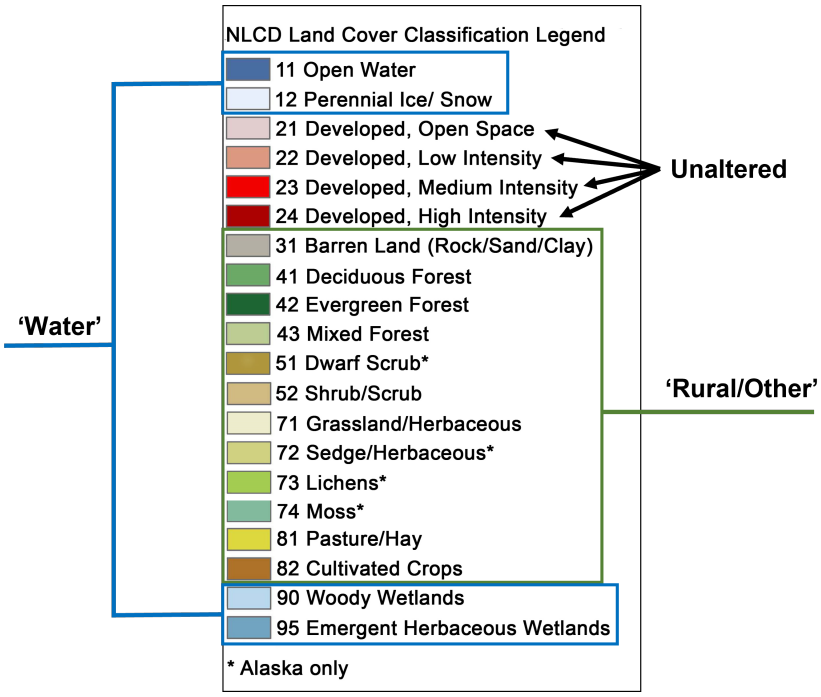

Figure 3: Bins for NLCD classifications

We also derived impervious surface using the median and mean impervious surface percentage within each hexagon.

5.4.2.5 Road Networks

Variables for roads were derived from the OSM dataset. For each hexagon, we extracted OSM 'highways' which include the following classifications:

- motorway

- trunk
- primary
- secondary
- tertiary
- residential
- unclassified

For detailed information on what each class represents, please see OSM key:highway.

We calculated the total road distance within each hexagon in meters.

5.4.2.6 Distance and Neighborhood Metrics

For each hexagon, two neighborhoods were created, one with a radius of 3 hexagons and one with a radius of 9 hexagons. Census data, land cover data and building data were aggregated by neighborhood. For all variables, the neighborhood of each variable represented the sum of the neighborhood, except land cover (which uses the statistical mode and impervious surface, which was calculated both as the mean and median of the neighborhood).

Distance was calculated in two ways. Euclidean distance was calculated as the shortest number of hexagons between the endpoint and each hexagon. Manhattan distance was calculated as the shortest number of hexagons between the endpoint and each hexagon, which have paved roads but are not large interstate-style highways (Figure 4)

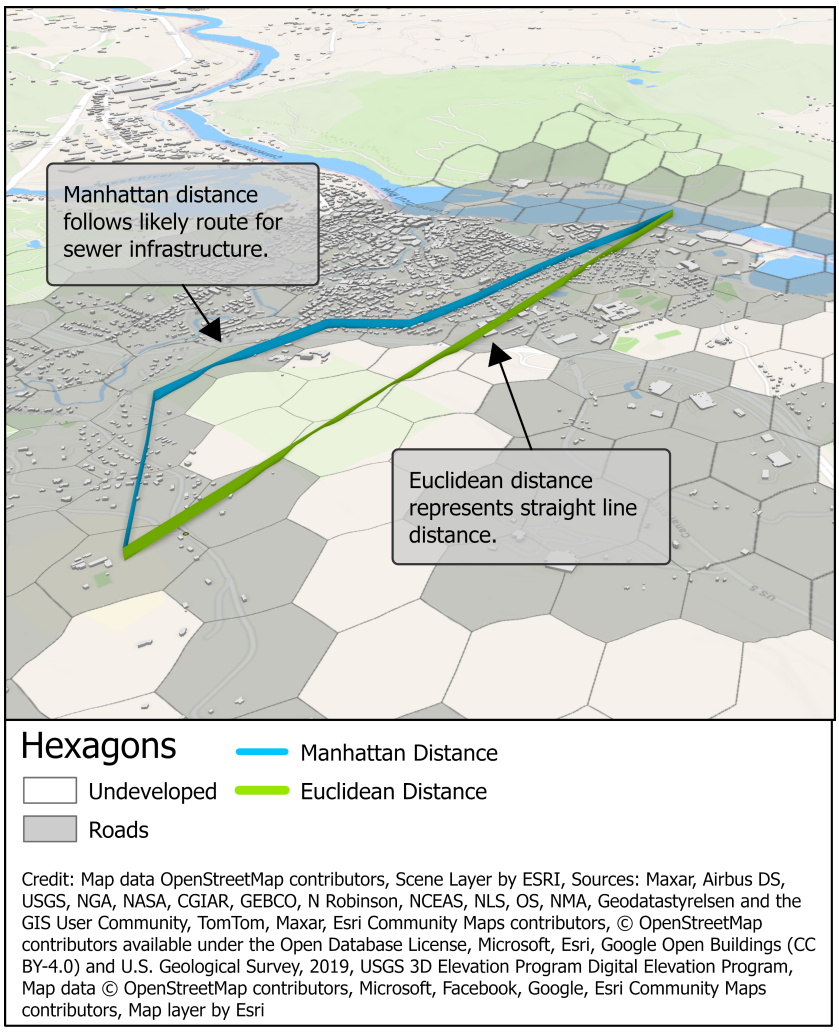

Figure 4: Euclidean vs Manhattan distance measures. Using manhattan distance more accurately accounts for the likely path of sewer pipes.

5.4.2.7 Endpoint Rank

For each hexagon - endpoint relationship, the Euclidean distance rank was calculated. For example, for a relationship between a hexagon and an endpoint where there are four other endpoints that are closer, a distance rank of 5 would be assigned. Additionally, the total residential populations served for other endpoints were summed into features representing total population served of closer endpoints and total population served of farther endpoints. These variables, "closer served" and "farther served" populations, give more context about the POTWs in the area, specifically whether there are larger or smaller systems nearby and how that relates to the distance of the hexagon to the endpoint.

5.4.2.8 Area & Name Matching

Each hexagon was spatially joined to its parent sub-county, place and county geography (Census Bureau TIGER/Line). The names of each of these geographies was then compared to the facility name of each CWNS endpoint within the area of consideration by using a Jaro-Winkler string distance calculation to determine name similarity score. The Match Score was calculated as whichever of the three scores returned the best result to provide flexibility for POTWs that may be named after geographies of different sizes.

5.4.3 Selecting and Splitting the Training and Testing Datasets

Machine learning models require a training dataset, which is used to train the model to recognize patterns in the data. These patterns are used to create a model, which will be used on additional data to return a probability that a hexagon is part of the sewershed for a given endpoint location. A separate testing dataset, which has not been exposed to the model training process is then used to validate the model and measure performance. To ensure that the model was trained and validated on the best possible data, a linear regression was used to relate the total residential population served by the endpoint in the CWNS with the 2020 census population for the same area. Where the residual of the population for a sourced system was more than two standard deviations away from the mean, that sewershed was removed from the training and testing sets, but retained for inclusion in the final dataset.

CWNS endpoints were randomly sampled into either training or testing, with 70% of endpoints used for training and 30% used for testing. The area of consideration for the model to assign probabilities to hexagons is ~3,200 km^2, meaning that there is a class imbalance between hexagons that are 'FALSE' (not sewered by that specific endpoint) and 'TRUE' (sewered by that specific endpoint). This class imbalance was corrected by randomly sampling 'FALSE' hexagons to match the number of 'TRUE' hexagons in the training set. We confirmed that the distribution of variables in the corrected training data was representative of the original training set.

5.5 Model Fitting

A boosted tree model was selected for its ability to resolve complex interactions and for its improved accuracy over decision tree and random forest models, which were tested. The algorithm used was the xgBoost (extreme gradient boosting), obtained through the 'xgboost' package in R. The training data was composed of 200,000 observations (100,000 'TRUE' and 100,000 'FALSE'). The model was tuned to find the optimal parameters for the model which was done using a grid search with 5-fold cross validation. For tuning, 'gbtree' and 'gblinear' methods were tested, with maximum tree depths ranging between 3 and 10, a minimum child weight between 1 and 10, a subsample between 0.5 and 1 and a column sample by tree between

0.5 and 1. Early models incorporated all features but were later restricted to features with high importance values.

5.6 Constructing Sewersheds from Model Results

Once the model was applied nationally, probabilities for hexagon-to-endpoint pairs were aggregated into distinct sewersheds for each endpoint. To determine a 'probability threshold', meaning the probability above which is considered a positive result, the model was tested across a classification threshold between 0.01 and 0.99 and broken down by sewershed population. Bins for population served are '< 1,000', '1,000 - 4,999','5,000 - 9,999','10,000 - 99,999' and '10,000 +'. The process for hexagon selection and aggregation is as follows:

1. CWNS endpoints are ranked from smallest to largest 'Total Residential Population'. This ensures that smaller systems take priority over larger systems when a hexagon has more than one high probability and avoids overlapping sewersheds.
2. For each endpoint, hexagons are ranked by descending probability of being part of the sewershed.
3. A cutoff is determined for the number of hexagons to include in the sewershed using the following criteria:

- The row at which the cumulative sum of hexagon populations reaches the total residential population reported in the CWNS
- The row before the probability drops below the probability threshold corresponding to the endpoint's population bin.

4. Selected hexagons within the cutoff are spatially aggregated into a multi-polygon, then exploded into individual polygons. The largest of the polygons is considered the primary sewershed area. Distance is then measured (edge-to-edge) between the primary polygon and the secondary polygons. Secondary polygons are considered spatial outliers if:

- The area of a secondary polygon represents $< 5\%$ of the total sewershed area
- The distance between the secondary polygon and the primary polygon is between 5 and 10 km and the total area of the secondary polygon is $< 10\%$ of the total sewershed area
- The distance between the secondary polygon and the primary polygon is > 10 km and the area of the secondary polygon is $< 30\%$ of the total sewershed area.

Once outliers were identified, they were removed from the sewershed. Any hexagons that were within an outlier polygon or beyond the population / probability cutoff were considered to still be available for larger sewersheds. Any hexagon that was determined to be part of the sewershed (not an outlier) was removed from consideration for larger sewersheds. Once final sewersheds were delineated, interior holes were filled to aid in visualization.

5.7 Validation

Validations are conducted at two stages of the sewershed development process. The first validation is performed while tuning the boosted tree model, to determine the performance of the model in its ability to correctly predict whether a hexagon belongs to a specific sewershed or not. The second validation is performed after hexagons are aggregated into sewersheds to determine how well the model performed in estimating the area and population served by each sewershed. In each validation, the sourced sewersheds from the testing set are used to quantify accuracy.

In model tuning, the testing hexagons are compared to model outputs as a binary classification problem. In a binary classification problem, a cutoff value is used to determine a positive or negative outcome. For each model, we tested a range of cutoff values between 0.01 and 0.99. For each cutoff, we calculated total accuracy, sensitivity (true positive rate) and specificity (true negative rate). These metrics were further broken down by system size. Using the total residential population served as reported in the CWNS, systems were binned into five categories: '< 1,000', '1,000 - 4,999','5,000 - 9,999','10,000 - 99,999' and '100,000 +'. The goal of this validation is to determine the optimal cutoff value for each population bin that maximizes accuracy while balancing sensitivity and specificity.

Two validations were performed to evaluate the aggregated sewersheds relative to total area and population served. Percent area captured is defined as the number of hexagons that the model correctly placed into a sewershed divided by the total number of hexagons within that sewershed as calculated using the sourced sewershed data. Percent sewered population captured is therefore the percent of the sewered population as calculated from the sourced sewershed data that was correctly placed into the sewershed by the model.

6 Results

In total, we started with 54,738 rows of data, which were queried down to a universe of 17,272 individual endpoints to be included in the modeling efforts. Of these endpoints, 276 were found to either have duplicated geo-locations with multiple POTWs at the same location or were otherwise found to have insufficient data reported in the CWNS. Therefore, the final universe of endpoints used for sewershed delineation is 16,996.

A total of 3,187 sewersheds were successfully matched with their corresponding CWNS endpoints, leaving the model to estimate sewersheds for 13,809 endpoints.An application was built to enable multiple team members to perform validation of the matches (Figure 5)

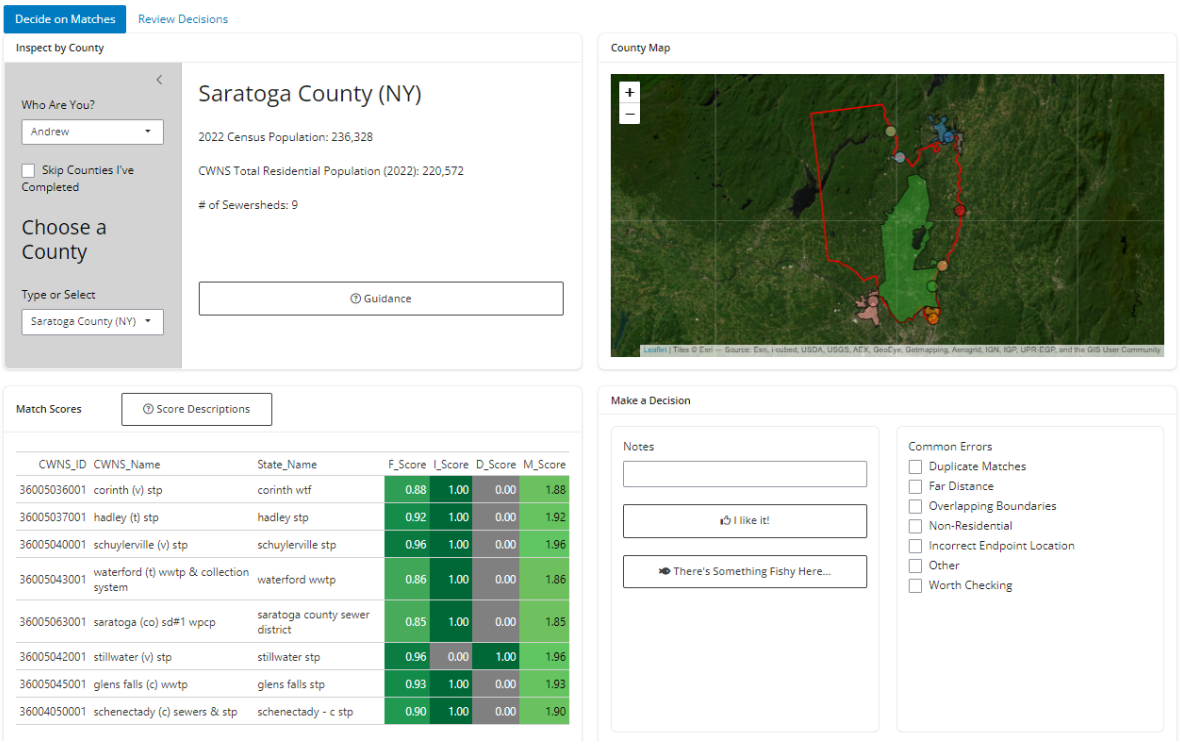

Figure 5: A screenshot of a shiny application used by the research team to validate sewershed to CWNS endpoint matches.

6.1 Selecting Training and Testing Sewersheds

A linear regression was fit using the aggregate population of hexagons within sourced sewersheds to predict the total residential population served in 2022 as reported in the CWNS. The regression returned an adjusted R-squared value of 0.83, with a p-value of < 0.001, indicating a strong relationship between the two variables. The residuals were then normalized by total residential population and z-scores were calculated (Figure 6).

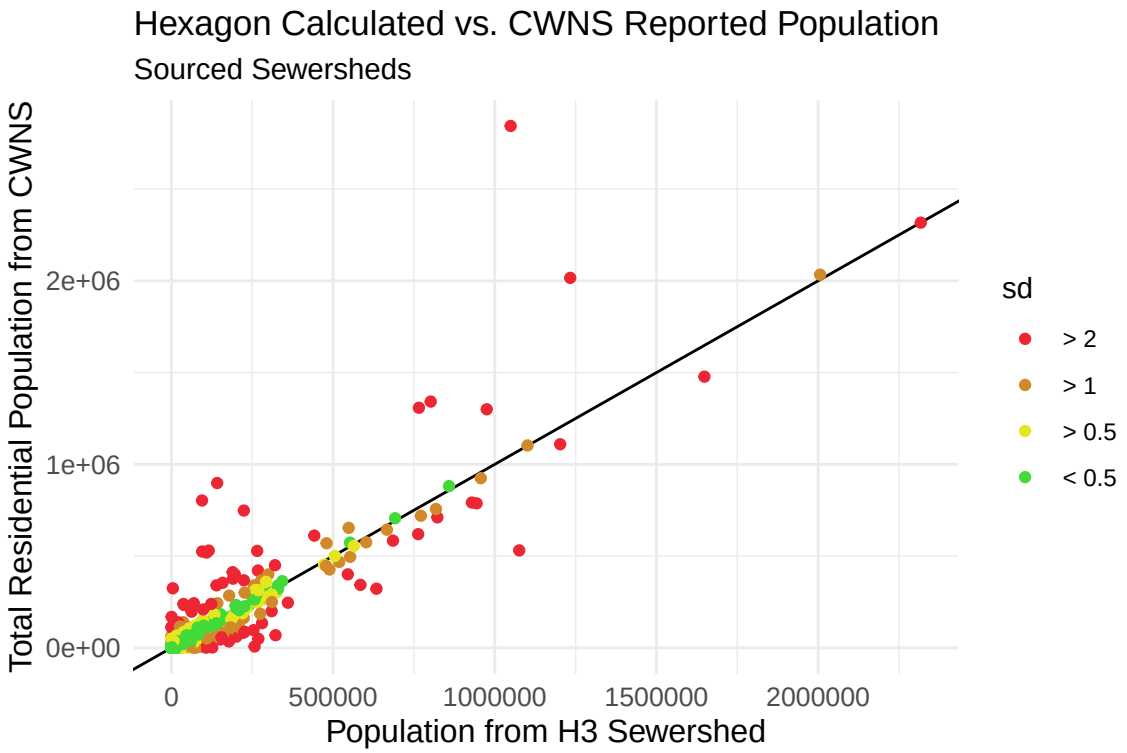

Figure 6: A scatter plot showing the relationship between measured population and reported population served, colored by z-scores.

For the 2,942 sewersheds, whose residuals were less than 0.5 standard deviations from the predicted value, the residuals were normalized by the total residential population served as reported in the CWNS. The sourced systems with the 2,000 smallest normalized residuals were used for the training and testing datasets (Figure 7). A 70:30 split was used to randomly select POTWs for the training and testing sets. The model was trained on 1,400 sewersheds and tested on 600 sewersheds.

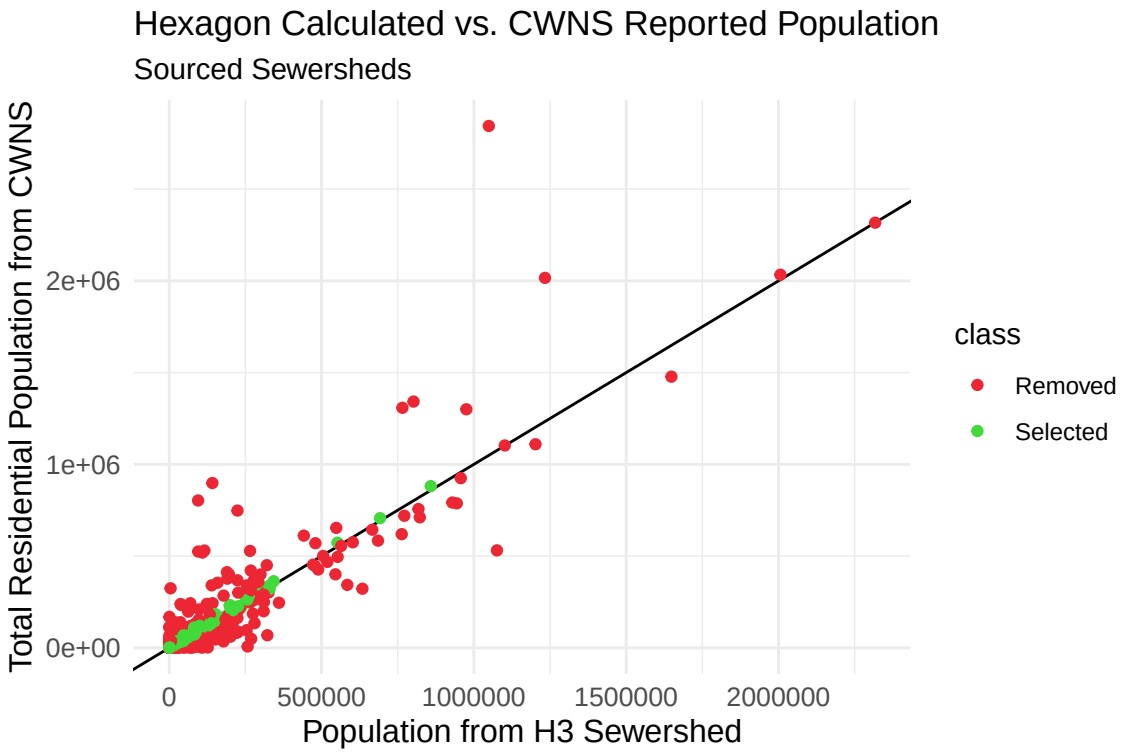

Figure 7: A scatter plot showing the relationship between measured population and reported population served, colored by wether the sewershed will be considered for model training and validation or left out.

6.2 Model Tuning

25 variables were identified as the most important features for the model (Figure 8):

- Euclidean Distance ('E_Distance')
- Total Residential Population of the endpoint (CWNS) ('TOTAL_RES_POPULATION_2022')
- Endpoint Near Rank ('Near_Rank')
- % of the 9-hexagon neighborhood that was sewered in 1990 ('Pub_S_90_9')
- Population Served by Closer endpoints ('Closer_Served')
- Count of Urban population within the 9-hexagon neighborhood ('Urban_Pop_9')
- Hexagon and endpoint are in the same county ('County_Match')
- Best name match score between the facility name and geographic area of the hexagon ('Match_Score')
- The difference between euclidean and Manhattan distance ('S_Distance')
- Hexagon and enpoint are within the same zip code ('Zip_Match')
- The total population between the hexagon and the endpoint ('Pop_B')

- The total urban population between the hexagon and the endpoint ('Urban_B')
- Elevation of the endpoint ('EP_Elevation')
- Hexagon and the endpoint are within the same sub-county ('CouSub_Match')
- The count of urban population within the 3-hexagon neighborhood ('Urban_Pop_3')
- Mean Imperviousness of the hexagon ('Imprv_Mean')
- % of the hexagon that was sewered in 1990 ('Pct_Sewer_90')
- Median area of buildings within the hexagon ('Med_Bldg_Area')
- No name match was found between the hexagon and endpoint ('No_Match')
- Mean Area of buildings within the hexagon ('Mean_Bldg_Area')
- Hexagon and endpoint are within the same city/place ('Place_Match')
- The count of buildings within the hexagon ('nBldgs')
- A name match was found between the endpoint facility and the name of the city the hexagon was in ('City_Place')
- The elevation difference between the endpoint and the hexagon ('EP_Elev_Dif')
- Mean elevation of the 3-hexagon neighborhood ('mean_Elev_3')

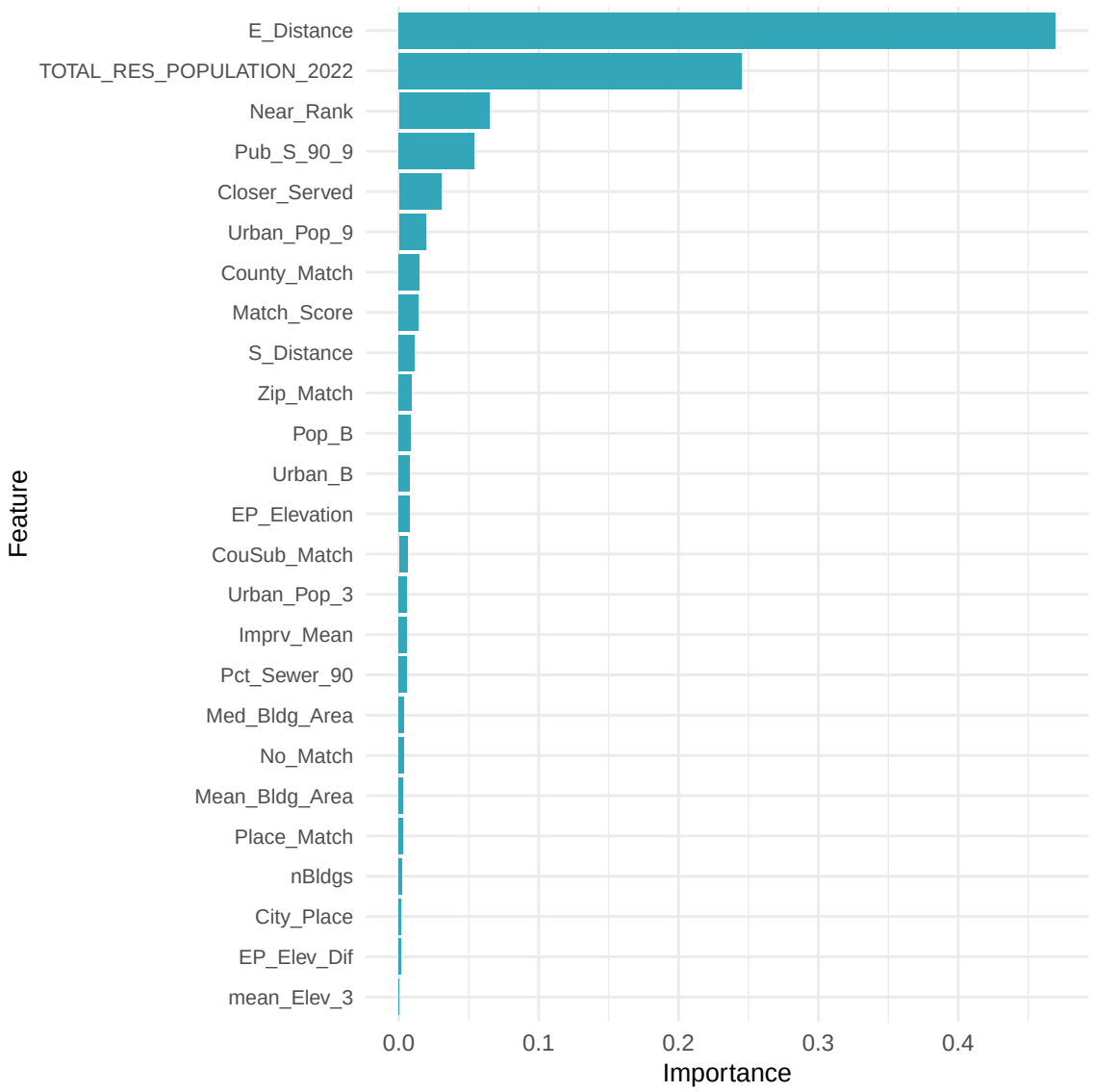

Figure 8: Variable importance for the sewershed model.

The results of the model tuning returned an optimal model using the gbtree method, a maximum tree depth of 7, a minimum child weight of 9.33, a sub sample of 0.994 and column sample by tree of 0.966. The overall accuracy of the final model was 90.65% with a sensitivity of 92.6% and a specificity of 88.9%.

The probability cutoff was calculated from the final model predictions on the testing dataset. The optimal cutoff ranged from 0.11 for large sewersheds (> 100,000 people) to 0.77 for small sewersheds (<1,000 people) (Figure 9)

Table 3

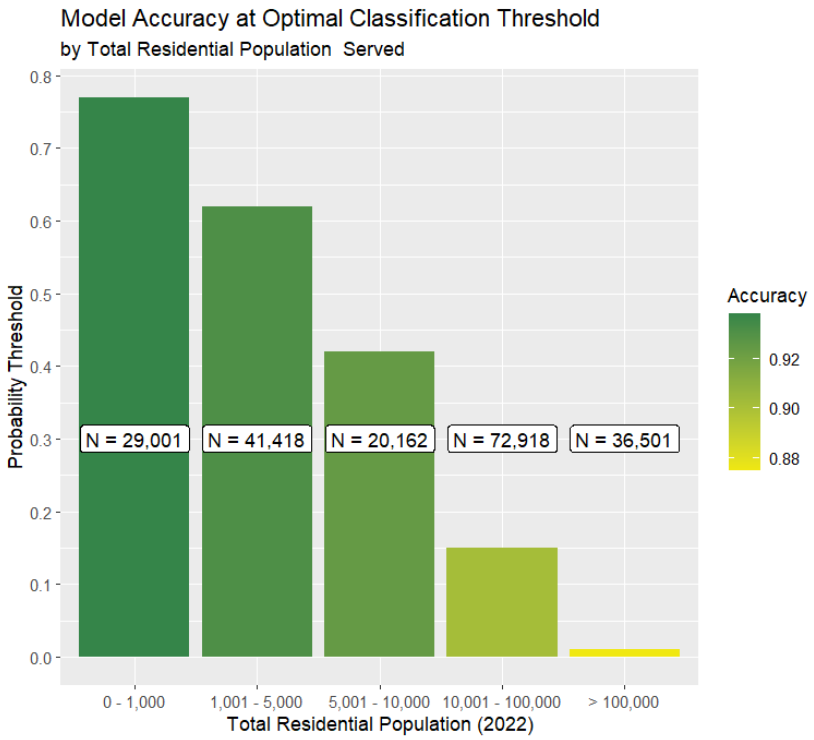

Figure 9: A bar plot showing that as sewershed size increases, the optimal cutoff value to predict a positive outcome (sewered) decreases.

6.3 Sewershed Validation

Compared with the utility sources sewersheds, modeled sewersheds from the testing dataset had a median population capture of 89.4% and a median area capture of 78.9% (Table 4). Accuracy measures for each of the 600 POTWs in the training set can be seen in Figure 10 and Figure 11.

Table 4: A table showing the percent of sewered population captured by modeled sewersheds compared with sourced sewersheds.

Sewershed Accuracy Metrics				
Metric	25th Percentile	Mean	Median	75th Percentile
Population Capture	74.80	81.53	89.40	96.80
Area Capture	61.80	74.16	78.90	92.90

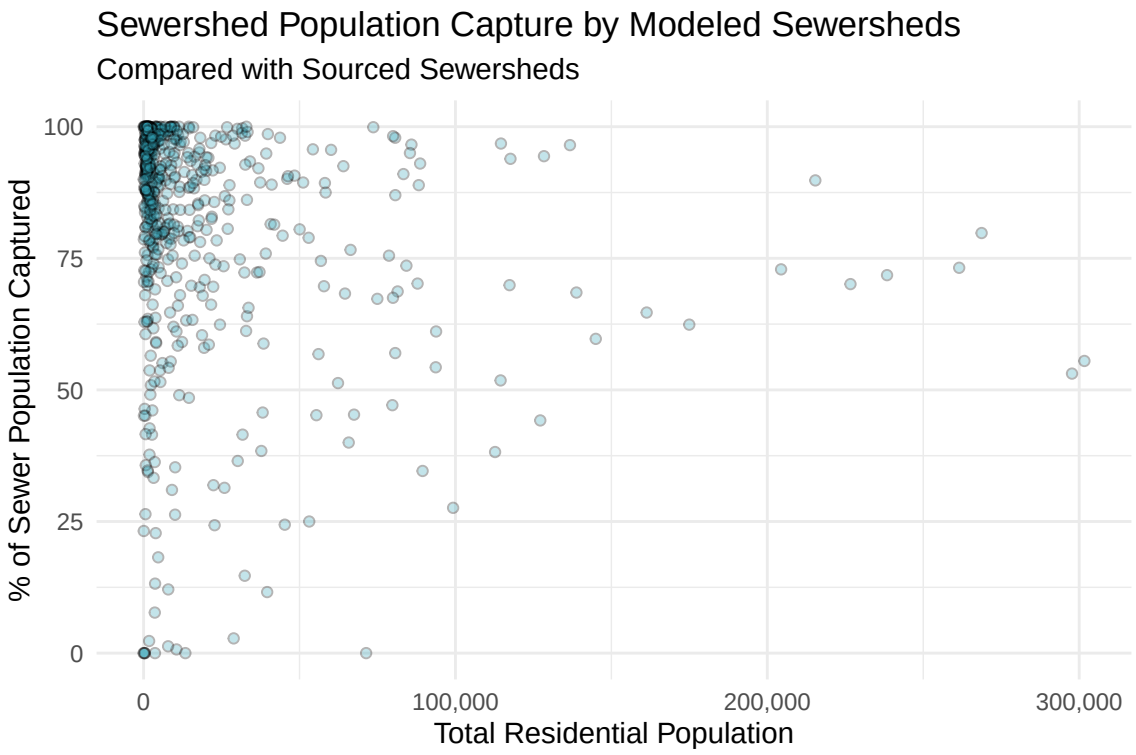

Figure 10: Percent of sewered population captured by modeled sewersheds compared with sourced sewersheds.

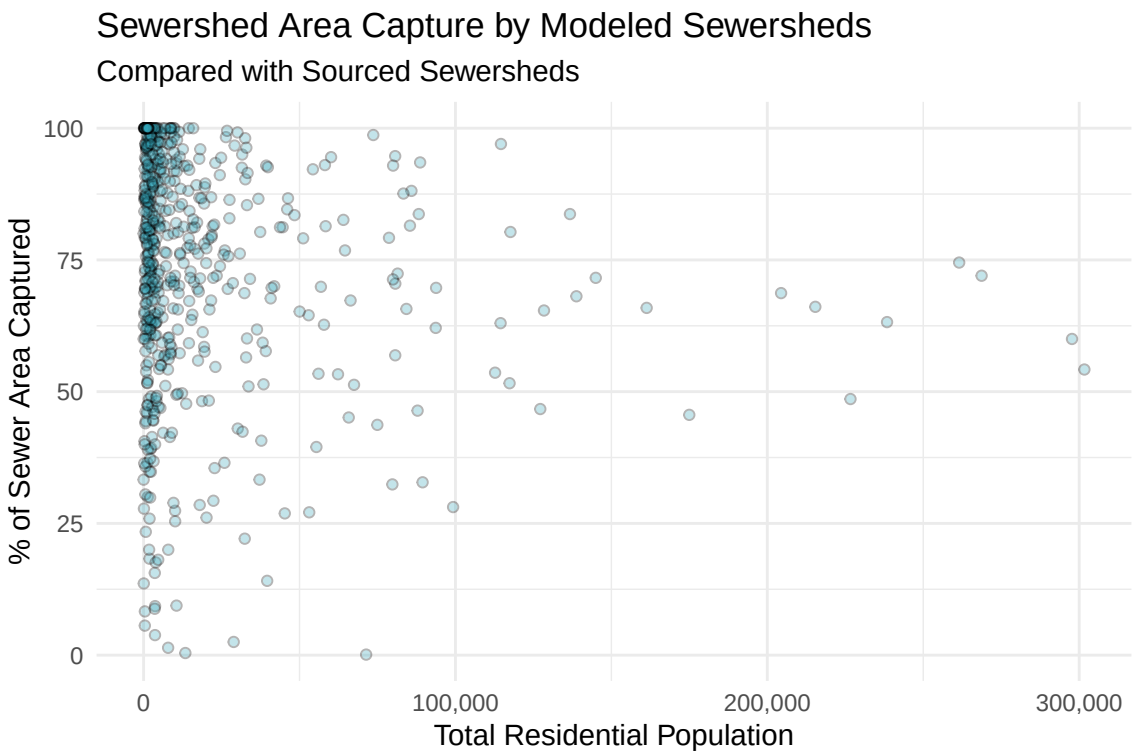

Figure 11: Percent of sewered area captured by modeled sewersheds compared with sourced sewersheds.

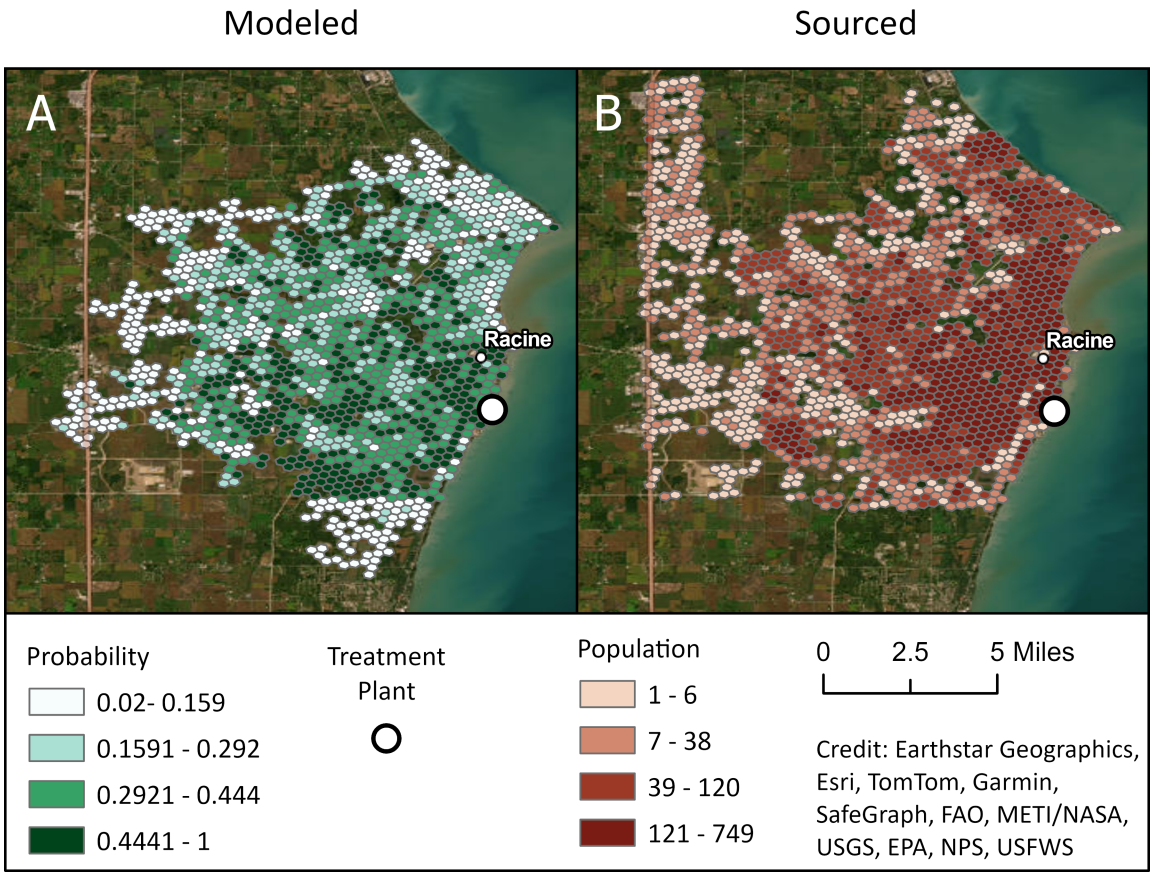

Figure 12: Modeled (Panel A) and sourced (Panel B) sewersheds for the Racine Sewage Treatment Plant in Racine, WI. Sewersheds are illustrated as pre-aggregated hexagons colored by model probability (Panel A) and 2020 census population (Panel B).

In total, the final dataset includes 16,961 sewersheds, meaning only 30 endpoints were not able to be modeled. The dataset can be accessed and viewed on the EPA internal web application

7 Discussion

The model and process to estimate sewersheds from the input data available represents the most advanced attempt to date for mapping wastewater infrastructure in the United States.

7.1 Known limits

As with all models, there are limitations. As an example, refer to Figure 12, which shows the modeled and sourced sewersheds for the Racine Sewage Treatment Plant in Racine, WI. The modeled sewershed captures 97% of the population served by the sourced sewershed, but only 84% of the area. You can see the area the model struggled with was the northwest part of the sourced sewershed, which is both rural and far from the treatment plant. The model struggles with sewered areas that are farther away from treatment plants and are more rural, however that is how we expect it to perform given our understanding of wastewater collection systems. Another known issue with modeling sewersheds is the need for accurate input data. For example, if a treatment plant incorrectly reports its population served, the model will struggle to accurately delineate a sewershed. We found evidence of this in some relatively rare scenarios such as neighboring treatment plants double-reporting populations, resulting in oversized sewersheds.

7.2 Planned improvements

EPA plans to continue to improve this dataset in multiple ways. First, we will continue to explore ways to improve the model, including the addition of new input data as they become available. Second, we plan to publish a web-based application to allow the public to provide feedback and suggest edits to the dataset. The web application will allow for three types of submissions:

1. The ability to upload a spatial file for a sewershed and provide a justification for why EPA should consider replacing the currently published version.
2. An interactive map where users and click and drag vertices to manually edit a sewershed and provide a justification for why EPA should consider replacing the currently published version.
3. An address search tool where users can enter an address and see which sewershed EPA has assigned to that location. If the user believes the assignment is incorrect, they can suggest either the correct sewershed or suggest that the address is not sewered.

The primary goal of this dataset is to provide the most accurate national dataset of sewersheds possible. We believe that by engaging the public and allowing for feedback, we can continue to improve the dataset over time, which will benefit the EPA mission, as well as other state and federal agencies, NGOs, academics and the public.

8 Appendices

8.1 Appendix I - Existing Sewershed Data

Table 5: Sources for sourced sewersheds

State	Detail	Source
AR	Little Rock	Link
AZ	Lakeside	Link
AZ	Pima County	Link
CA	Bakersfield	Link
CA	Coachella	Link
CA	San Diego	Link
CA	San Francisco	Link
CA	Sonoma County	Link
CO	Statewide	Link
CO	Denver Suburbs	Link
CO	Palisade	Link
CO	Southeast Metro	Link
CT	Statewide	Link
DE	Statewide	Link
FL	Central	Link
FL	Bradenton	Link
FL	North Port	Link
FL	Sarasoto	Link
FL	Tampa	Link
GA	Eastern	Link
GA	Canton	Link
HI	Honalulu	Link
HI	Maui County	Link
IA	Des Moines	Link
IA	Durant	Link
IL	Chicagoland Area	Link
IL	Lake County	Link
IL	Naperville	Link
KS	Olathe	Link
KY	Statewide	Link
MD	Statewide	Maryland Department of the Environment
MI	Detroit MI	Link
MN	Twin Cities	Link

State	Detail	Source
MO	Statewide	Missouri Department of Health and Human Services, Missouri Department of Natural Resources
MO	Kansa City Area	Link
MO	Missoula	Link
MS	Statewide	Link
MT	Southwest	Link
NC	Statewide	Link
NC	Statewide	Link
NH	Statewide	New Hampshire Department of Environmental Services
NJ	Statewide	Link
NJ	Monmouth County	Link
NM	Santa Fe	Link
NV	Douglas County	Link
NY	Statewide	Link
NY	Statewide	Link
OH	Statewide	Ohio Department of Health
OK	Norman	Link
OR	Statewide	Oregon Department of Health
OR	Portland Suburbs	Link
OR	Woodburn	Link
PA	Berks County	Link
PA	Central	Link
PA	Chester County	Link
PA	Penn Hills	Link
PA	Philadelphia	Link
RI	Statewide	Link
RI & MA	Partial	Link
SC	Statewide	Link
TX	Statewide	Link
TX	Statewide	Link
TX	La Porte	Link
TX	San Antonio	Link
TX	Woodlands	Link
UT	Statewide	Utah Department of Health and Human Services, Utah Department of Environmental Quality

State	Detail	Source
VA	Newport News & Hampton Roads	Email correspondence
VA	Richmond	Link
VT	Statewide	Link
WA	Partial	Link
WA	King County	Link
WI	East Central	Link
WI	Marathon County	Link
WI	South Eastern	Link
WV	Statewide	Link

8.2 Appendix II (Data Query)

The CWNS data we are working with was received as a Microsoft Access Database, which is located in 'Data/CWNS 2022 Database_April2024.accdb'. We use 7 tables from this access database. Note that 'POPULATION_WASTEWATER_CONFIRMED' was manually updated by ERG in June, 2024 and does not reflect original data in the access database. Each table was exported from Access to a comma delimited file. The seven files used are:

- 'Data/FACILITIES.txt'
- 'Data/POPULATION_WASTEWATER.txt'
- 'Data/FACILITY_TYPES.txt'
- 'Data/PHYSICAL_LOCATION.txt'
- 'Data/FACILITIES_CONFIRMED.txt'
- 'Data/POPULATION_WASTEWATER_CONFIRMED_updated06242024.csv'
- 'Data/DISCHARGES.csv'

The following steps describe the process used to query endpoints to be included in modeling efforts. The code can also be viewed in the expandable code tab.

We create a flat file where each row represents a POTW and perform the following steps: - In 'POPULATION_WASTEWATER", if no value is given for the column 'END_FACILITY', we assign it "N", meaning it is not considered an end facility. - 'END_FACILITY' and 'FACILITY_TYPE' columns for the 'FACILITIES_CONFIRMED' file are set to 'Y' and 'Treatment Plant' respectively. - Discharge is calculated using the 'DISCHARGES' file. Because systems can discharge to multiple other systems, we calculate the sum of discharges to determine the total % discharged elsewhere. If discharge is not reported for a system, we default the value to zero. - Data tables are combined and the filter is applied:

- `TOTAL_RES_POPULATION_2022 > 0 & FACILITY_TYPE=="Treatment Plant" & PRESENT_DISCHARGE_PERCENTAGE < 50`

Total receiving population = total number of people the facility is receiving from, regardless of whether its sent on elsewhere

References

[1] Jonathan Manson Steven; Schroeder. *IPUMS National Historical Geographic Information System*. http://doi.org/10.18128/D050.V19.0. 2024.

[2] Microsoft. *Global ML Building Footprints*. https://github.com/microsoft/GlobalMLBuildingFootprints. 2025.

[3] Alex Murray Andrew; Hall. *Community Water System Service Areas*. https://github.com/USEPA/ORD_SAB_Model. 2024.

[4] Alex Murray Andrew; Hall. *National Sewershed Service Areas*. https://github.com/USEPA/sewersheds. 2025.

[5] *National Land Cover Database (NLCD)*. U.S. Geological Survey, 2020.

[6] Inc. Uber Technologies. *H3 Hexagons*. https://github.com/uber/h3. 2025.