CA Wastewater Needs Assessment
Overview of Wastewater Infrastructure Modeling Effort
Executive Summary for July 2025 Advisory Group Meeting

University of Massachusetts Amherst, Nelson da Luz, Emily Kumpel, Jay Taneja

Overview

Throughout the Wastewater Needs Assessment (WWNA), data on the types of wastewater infrastructure (e.g., collection systems, onsite wastewater treatment systems) serving Californians have been collected and processed. This data collection effort has been extensive, but there are still gaps in the data to be filled. To fill one of the major gaps in data, the WWNA project team will use a model developed by the University of Massachusetts Amherst (UMass) over several years (since 2021) to identify unsewered areas, i.e., likely locations of onsite wastewater treatment and disposal systems (e.g., septic tanks, cesspools, leach fields, and seepage pits). This Executive Summary describes the methods being used to develop this model and the next steps for its development and use.

Methodology

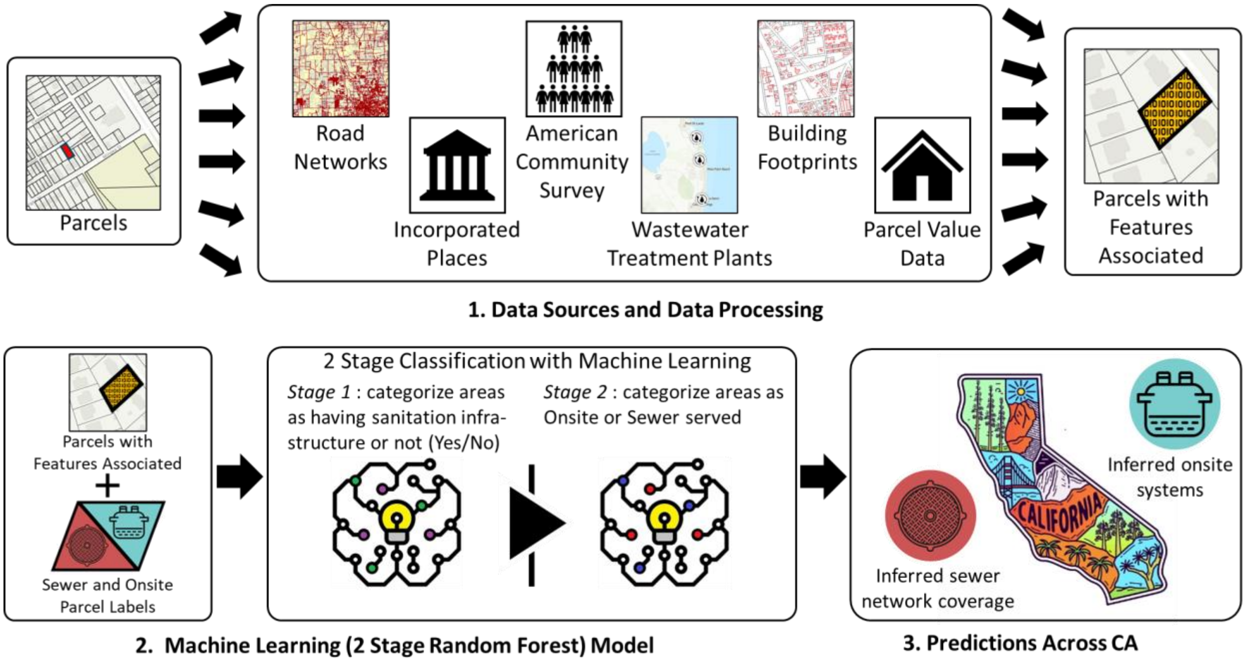
Parcels
Road Networks
Incorporated Places
American Community Survey
Wastewater Treatment Plants
Building Footprints
Parcel Value Data
Parcels with Features Associated
1. Data Sources and Data Processing
Parcels with Features Associated
+
Sewer and Onsite Parcel Labels
2 Stage Classification with Machine Learning
Stage 1 : categorize areas as having sanitation infra-structure or not (Yes/No)
Stage 2 : categorize areas as Onsite or Sewer served
2. Machine Learning (2 Stage Random Forest) Model
Inferred onsite systems
Inferred sewer network coverage
CALIFORNIA
3. Predictions Across CA

Figure 1: Identification of Unsewered Areas Method Workflow

This model leverages the patterns found among homes and businesses with known wastewater system types to predict the wastewater system type where it is unknown. The core unit of analysis for the model is the land parcel. For each parcel, a set of characteristics related to possible sewer coverage associated with that parcel is first identified. These characteristics include information about how many and what types of roads are nearby, building density and land area, distance to wastewater treatment facilities, whether the parcel is in an incorporated place, the value of the parcel, and other features from the US Census American Community Survey. In total, 59 characteristics from these sources are identified for each parcel[1]. UMass conducted an in-depth process to arrive at the selection of these characteristics, which included an in-depth evaluation of dataset availability and accuracy, elimination of highly correlated features, and evaluation of the relative importance of each characteristic in comparison to overall model performance.

Once these characteristics are assigned to each parcel, ground truth data – information about whether a parcel is served by either sewer or an onsite system – is associated with parcels to allow for model training and testing. This ground truth data includes known locations of onsite wastewater treatment systems from permits and digital maps of sewer systems collected by the WWNA project team and the State Water Resources Control Board (State Water Board). These parcels with features associated with them are used to train a machine learning model. Machine learning is a part of the artificial intelligence (AI) field, and in this context, the method used allows a computer model to identify complex patterns in large datasets and use that information to make predictions to fill gaps. The model is built to make 1 of 3 possible classifications:

1) Sewer - served by a sewer connection (collection system),

2) Onsite - served by an onsite solution (e.g., septic tank, cesspool, or straight pipe), or

3) Not Applicable - not served by any wastewater infrastructure.

UMass employs a two-stage machine learning approach to model wastewater infrastructure coverage[2]. Stage 1 identifies whether a parcel needs wastewater infrastructure, and Stage 2 identifies whether it is served by an onsite wastewater treatment or disposal system or by a sewer connection to a wastewater collection system. In Stage 1, UMass implements a simple classification of the need for wastewater infrastructure based on the presence or absence of a building on the parcel. In Stage 2, UMass trains a machine learning model using a machine learning technique called Random Forest, which combines multiple decision trees trained on subsets of data to produce robust predictions through majority voting or averaging. For parcels that are assigned a classification of Sewer or Onsite in Stage 2, the model also outputs a confidence value. The confidence values describe how confident the model is that the assigned classification is correct on a scale of 0 to 100 (100 being the highest).

[1] See Appendix.

[2] Many more details on the overall approach used to develop the model can be found here: https://dx.doi.org/10.21203/rs.3.rs-6656886/v1

Current State of the Model

To train the model, over the last 2 years, UMass has worked with the WWNA project team, State Water Board, and the California Regional Water Quality Control Boards (collectively Water Boards) to collect available wastewater infrastructure data (lists of parcels that include sewer systems or onsite wastewater treatment or disposal systems from parts of 10+ states including California). From California specifically, UMass currently includes ground truth data samples from 12 counties (more than 616 thousand samples out of 2.7 million total samples used for training). Examples from California, as well as other states, are used to help train the model and improve the model's performance in California.

The current model achieves 93.8% accuracy. Here, accuracy is defined as the number of correct predictions divided by the total number of predictions made. Accuracy is evaluated against a testing dataset that includes held-out samples from census tracts in California. UMass held out certain census tracts from training to make sure that the model was not overly influenced by having nearby samples that were very similar. Making predictions with the current model results in a statewide prediction confidence of 87.8%. Table 1 shows the breakdown of what the model currently predicts across the state of California by classification category. It is important to note that these estimates correspond to the number of parcels served by each type of wastewater service; they do not represent population estimates.

Table 1: Summary of Model Predictions across California

Category	**Count of Parcels**	**Percentage**[3]
Onsite	1,545,917	12.13%
Sewer	9,419,937	73.94%
Not Applicable	1,774,714	13.93%

Future Work

UMass will work in the coming months to further improve the model, including adding in more 'true' data (e.g., onsite wastewater treatment system (OWTS) location data from Local Agency Management Programs (LAMPs) and sewer system maps) as part of WWNA Phase 1C Data Collection and Gap Analysis. The WWNA project team will also perform checks on the model to evaluate its suitability in a variety of geographical conditions (e.g., elevation, rural vs urban) that are present across California. The model will ultimately be used in the Groundwater Impacts

[3] This is the percentage of land parcels served by each category of wastewater service relative to all land parcels in California.

Assessment and Onsite Sewage Treatment Systems (OSTS) Connection Opportunities portions of the broader WWNA project, as well as by the Water Boards for future broader purposes.

Appendix: Variables used to train the model

Dataset	Variable Code	Description
Building Footprints	BCBGC1000	Number of Buildings completely contained within an intersecting 1000m x 1000m grid cell
Building Footprints	BCBGC250	Number of Buildings completely contained within an intersecting 250m x 250m grid cell
Building Footprints	BCBGC500	Number of Buildings completely contained within an intersecting 500m x 500m grid cell
Building Footprints	MaxBldgArea	Maximum Building Area on Parcel
Building Footprints	MBAwGC1000	Median Building Area within an intersecting 1000m x 1000m grid cell
Building Footprints	MBAwGC250	Median Building Area within an intersecting 250m x 250m grid cell
Building Footprints	MBAwGC500	Median Building Area within an intersecting 500m x 500m grid cell
Building Footprints	NBApPA	Ratio of Non-Building Area to Parcel Area
Building Footprints	NumBldgCont	Number of Buildings contained on the parcel
Building Footprints	TBAwGC1000	Total Building Area within an intersecting 1000m x 1000m grid cell
Building Footprints	TBAwGC250	Total Building Area within an intersecting 250m x 250m grid cell
Building Footprints	TBAwGC500	Total Building Area within an intersecting 500m x 500m grid cell
Building Footprints	TotBldgArea	Total Building Area on Parcel
Parcel Value	parval_Fill	Parcel (Monetary) Value
Parcels	MPAIGC1000	Median Parcel Area within an intersecting 1000m x 1000m grid cell
Parcels	MPAIGC250	Median Parcel Area within an intersecting 250m x 250m grid cell
Parcels	Parc_Area	Parcel Area
Parcels	PCBGC500	Number of Parcels completely contained within an intersecting 500m x 500m grid cell
Parcels	PIGC1000	Number of Parcels within an intersecting 1000m x 1000m grid cell
Parcels	PIGC250	Number of Parcels within an intersecting 250m x 250m grid cell
Parcels	PIGC500	Number of Parcels within an intersecting 500m x 500m grid cell
Roads	GC1000Len_S1400_max	Total length of S1400 Roads (Local Neighborhood Roads, Rural Roads, City Streets) within an intersecting 1000m x 1000m grid cell
Roads	GC1000Num_S1100_max	Number of S1100 Roads (Primary Roads) within an intersecting 1000m x 1000m grid cell
Roads	GC1000Num_S1200_max	Number of S1200 Roads (Secondary Roads) within an intersecting 1000m x 1000m grid cell

Dataset	Variable Code	Description
Roads	GC1000Num_S1400_max	Number of S1400 Roads (Local Neighborhood Roads, Rural Roads, City Streets) within an intersecting 1000m x 1000m grid cell
Roads	GC1000Num_S1740_max	Number of S1740 Roads (Private Roads for service vehicles) within an intersecting 1000m x 1000m grid cell
Roads	GC2000Len_S1400_max	Total length of S1400 Roads (Local Neighborhood Roads, Rural Roads, City Streets) within an intersecting 2000m x 2000m grid cell
Roads	GC2000Num_S1100_max	Number of S1100 Roads (Primary Roads) within an intersecting 2000m x 2000m grid cell
Roads	GC2000Num_S1200_max	Number of S1200 Roads (Secondary Roads) within an intersecting 2000m x 2000m grid cell
Roads	GC2000Num_S1400_max	Number of S1400 Roads (Local Neighborhood Roads, Rural Roads, City Streets) within an intersecting 2000m x 2000m grid cell
Roads	GC2000Num_S1740_max	Number of S1740 Roads(Private Roads for service vehicles) within an intersecting 2000m x 2000m grid cell
Roads	RdLenGC1000_max	Total length of roads within an intersecting 1000m x 1000m grid cell
Roads	RdLenGC2000_max	Total length of roads within an intersecting 2000m x 2000m grid cell
US Census American Community Survey	B25010e1	Census Block Average household size of occupied housing units
US Census American Community Survey	B25010e2	Census Block Average household size of owner-occupied housing units
US Census American Community Survey	B25010e3	Census Block Average household size of renter-occupied housing units
US Census American Community Survey	B25017e1	Census Block Total: housing units
US Census American Community Survey	B25017e2	Census Block Total: 1 room housing units
US Census American Community Survey	B25017e3	Census Block Total: 2 room housing units
US Census American Community Survey	B25017e4	Census Block Total: 3 room housing units
US Census American Community Survey	B25017e5	Census Block Total: 4 room housing units
US Census American Community Survey	B25017e6	Census Block Total: 5 room housing units
US Census American Community Survey	B25017e7	Census Block Total: 6 room housing units
US Census American Community Survey	B25017e8	Census Block Total: 7 room housing units

Dataset	Variable Code	Description
US Census American Community Survey	B25017e9	Census Block Total: 8 room housing units
US Census American Community Survey	B25017e10	Census Block Total: 9 or more rooms housing units
US Census American Community Survey	B25037e1	Census Block Median year structure built of occupied housing units
US Census American Community Survey	B25037e2	Census Block Median year structure built of owner-occupied housing units
US Census American Community Survey	B25037e3	Census Block Median year structure built of renter-occupied housing units
US Census American Community Survey	B25041e2	Census Block Total: No bedroom housing units
US Census American Community Survey	B25041e3	Census Block Total: 1 bedroom housing units
US Census American Community Survey	B25041e4	Census Block Total: 2 bedrooms housing units
US Census American Community Survey	B25041e5	Census Block Total: 3 bedrooms housing units
US Census American Community Survey	B25041e6	Census Block Total: 4 bedrooms housing units
US Census American Community Survey	B25041e7	Census Block Total: 5 or more bedrooms housing units
US Census Places	CDP	US Census Designated Place (Yes or No)
US Census Places	Incorp	Incorporated Place in US Census (Yes or No)
US Census Places	Undesignated	Place undesignated by US Census (Yes or No)
Wastewater Treatment Plants	WWTP_Dist	Distance to the nearest wastewater treatment facility