

Evaluación de las necesidades de aguas residuales de California
Resumen del Esfuerzo de Modelación de la Infraestructura de Aguas Residuales
Resumen Ejecutivo para la Reunión del Grupo Asesor de Julio de 2025

Universidad de Massachusetts Amherst, Nelson da Luz, Emily Kumpel, y Jay Taneja

Resumen

A lo largo de la Evaluación de Necesidades de Aguas Residuales (WWNA, por sus siglas en inglés), se han recopilado y procesado datos sobre los tipos de infraestructura de aguas residuales (por ejemplo, sistemas de recolección, sistemas de tratamiento de aguas residuales en el sitio) que prestan servicio a los habitantes de California. Este esfuerzo de recopilación de datos ha sido extenso, pero aún existen vacíos que deben ser llenados. Para llenar una de las principales brechas de datos, el equipo del proyecto WWNA utilizará un modelo desarrollado por la Universidad de Massachusetts en Amherst (UMass) durante varios años (desde 2021) para identificar áreas sin conexión a alcantarillado, es decir, ubicaciones probables de sistemas de tratamiento y disposición de aguas residuales en el sitio (por ejemplo, fosas sépticas, pozos negros, campos de infiltración y pozos de absorción). Este Resumen Ejecutivo describe los métodos que se están utilizando para desarrollar este modelo y los próximos pasos para su desarrollo y aplicación.

Metodología

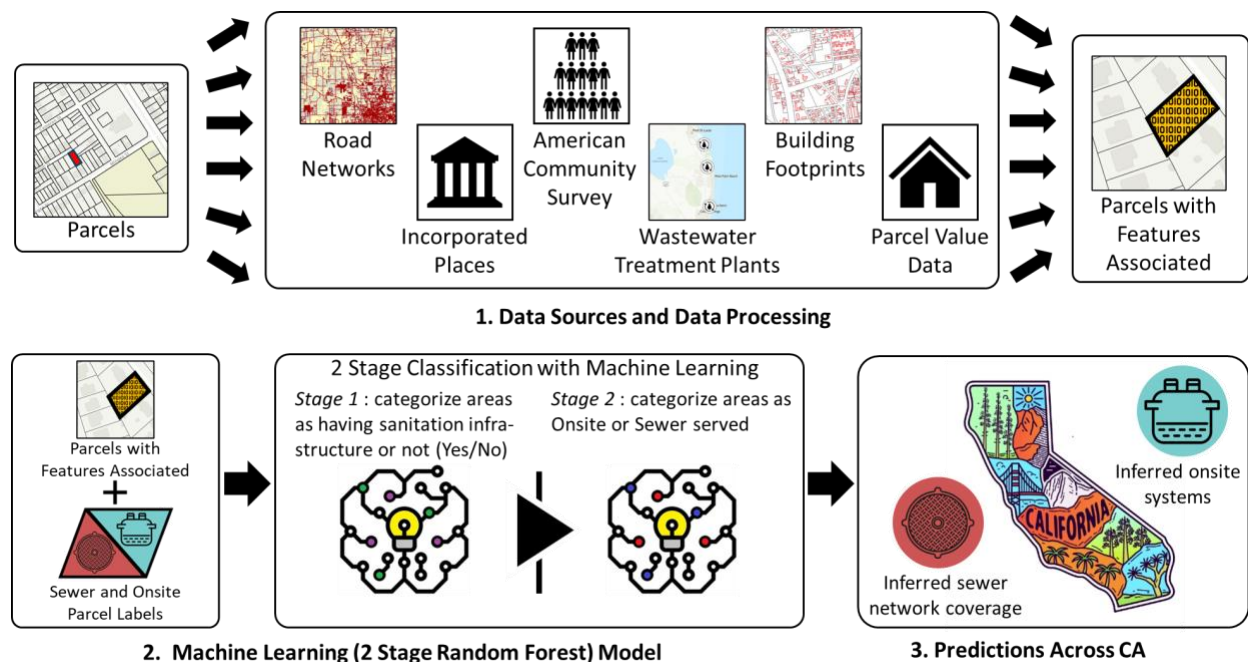


Figura 1: Flujo de Trabajo del Método de Identificación de Áreas sin Alcantarillado

Este modelo aprovecha los patrones encontrados entre viviendas y negocios con tipos de sistemas de aguas residuales conocidos para predecir el tipo de sistema donde no se conoce. La unidad principal de análisis del modelo es la parcela de terreno. Para cada parcela, primero se identifica un conjunto de características relacionadas con la posible cobertura de alcantarillado asociada a dicha parcela. Estas características incluyen información sobre cuántas y qué tipos de carreteras hay cerca, la densidad de edificaciones y el área del terreno, la distancia a las plantas de tratamiento de aguas residuales, si la parcela se encuentra en una zona incorporada, el valor de la parcela y otras características provenientes de la Encuesta sobre la Comunidad Estadounidense del Censo de EE. UU. En total, se identifican 59 características de estas fuentes para cada parcela¹. UMass llevó a cabo un proceso profundo para llegar a la selección de estas características, que incluyó una evaluación en profundidad de la disponibilidad y precisión del conjunto de datos, la eliminación de características altamente correlacionadas y la evaluación de la importancia relativa de cada característica en comparación con el rendimiento general del modelo.

Una vez que se asignan estas características a cada parcela, los datos de verdad fundamental (información sobre si una parcela cuenta con servicio de alcantarillado o de un sistema en el sitio) se asocian con las parcelas para permitir el entrenamiento y la prueba del modelo. Estos datos de verdad fundamental incluyen ubicaciones conocidas de sistemas de tratamiento de aguas residuales en el sitio a partir de permisos y mapas digitales de sistemas de alcantarillado recopilados por el equipo del proyecto WWNA y la Junta Estatal de Control de Recursos del Agua (Junta Estatal del Agua). Estas parcelas, con sus características asociadas, se utilizan para entrenar un modelo de aprendizaje automático. El aprendizaje automático forma parte del campo de la inteligencia artificial (IA) y, en este contexto, el método empleado permite que un modelo informático identifique patrones complejos en grandes conjuntos de datos y utilice esa información para realizar predicciones que permitan subsanar deficiencias. El modelo está diseñado para realizar una de tres clasificaciones posibles:

- 1) Alcantarillado - servido por una conexión de alcantarillado (sistema de recolección),
- 2) En el sitio - atendido por una solución en el sitio (por ejemplo, tanque séptico, pozo negro o tubería recta),
- 3) No aplicable - no cuenta con ninguna infraestructura de aguas residuales..

UMass emplea un enfoque de aprendizaje automático de dos etapas para modelar la cobertura de la infraestructura de aguas residuales ². La etapa 1 identifica si una parcela necesita

¹ Ver Apéndice.

² Se pueden encontrar muchos más detalles sobre el enfoque general utilizado para desarrollar el modelo aquí: <https://dx.doi.org/10.21203/rs.3.rs-6656886/v1>

infraestructura de aguas residuales, y la Etapa 2 identifica si cuenta con un sistema de tratamiento o eliminación de aguas residuales en el sitio o con una conexión de alcantarillado a un sistema de recolección de aguas residuales. En la Etapa 1, UMass implementa una clasificación simple de la necesidad de infraestructura de aguas residuales según la presencia o ausencia de un edificio en la parcela. En la Etapa 2, UMass entrena un modelo de aprendizaje automático utilizando una técnica de aprendizaje automático llamada Random Forest, que combina múltiples árboles de decisión entrenados en subconjuntos de datos para producir predicciones sólidas a través de la votación mayoritaria o el promedio. Para las parcelas clasificadas como Alcantarillado o In situ en la Etapa 2, el modelo también genera un valor de confianza. Este valor describe la certeza del modelo de que la clasificación asignada es correcta en una escala de 0 a 100 (100 es el valor más alto).

Estado Actual de el Modelo

Para entrenar el modelo, durante los últimos 2 años, UMass ha trabajado con el equipo del proyecto WWNA, la Junta Estatal del Agua y las Juntas Regionales de Control de Calidad del Agua de California (colectivamente Juntas del Agua) para recopilar datos disponibles de infraestructura de aguas residuales (listas de parcelas que incluyen sistemas de alcantarillado o sistemas de tratamiento o eliminación de aguas residuales en el sitio de partes de más de 10 estados, incluida California). Específicamente de California, UMass actualmente incluye muestras de datos de campo de 12 condados (más de 616 mil muestras de un total de 2.7 millones utilizadas para el entrenamiento). Se utilizan ejemplos de California, así como de otros estados, para ayudar a entrenar el modelo y mejorar su rendimiento en California.

El modelo actual alcanza una precisión del 93,8 %. En este caso, la precisión se define como el número de predicciones correctas dividido entre el número total de predicciones realizadas. La precisión se evalúa con un conjunto de datos de prueba que incluye muestras excluidas de los distritos censales de California. UMass excluyó ciertos distritos censales del entrenamiento para garantizar que el modelo no se viera excesivamente influenciado por la presencia de muestras cercanas muy similares. Al realizar predicciones con el modelo actual, se obtiene una confianza de predicción estatal del 87,8 %. La Tabla 1 muestra el desglose de las predicciones actuales del modelo en el estado de California por categoría de clasificación. Es importante señalar que estas estimaciones corresponden al número de parcelas atendidas por cada tipo de servicio de aguas residuales; no representan estimaciones de población.

Tabla 1: Resumen de las predicciones del modelo en California

Categoría	Recuento de Parcelas	Porcentaje ³
En el Sitio	1,545,917	12.13%
Alcantarillado	9,419,937	73.94%
No Aplicable	1,774,714	13.93%

Trabajo Futuro

UMass trabajará en los próximos meses para mejorar aún más el modelo, incluyendo la incorporación de más datos "reales" (por ejemplo, datos de ubicación del sistema de tratamiento de aguas residuales en el sitio (OWTS) de los Programas de Gestión de Agencias Locales (LAMP) y mapas del sistema de alcantarillado) como parte de la Recopilación de datos y análisis de brechas de la Fase 1C de WWNA. El equipo del proyecto WWNA también realizará controles en el modelo para evaluar su habilidad en una variedad de condiciones geográficas (por ejemplo, elevación, rural vs. urbano) presentes en toda California. El modelo se utilizará en última instancia en las secciones de Evaluación de impacto de aguas subterráneas y Oportunidades de conexión de sistemas de tratamiento de aguas residuales en el sitio (OSTS) del proyecto más amplio WWNA, así como también por las Juntas de Agua para futuros propósitos más amplios.

Apéndice: Variables utilizadas para entrenar el modelo

Conjunto de Datos	Código de Variable	Descripción
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

³ Este es el porcentaje de parcelas de tierra atendidas por cada categoría de servicio de aguas residuales en relación con todas las parcelas de tierra en California.

Conjunto de Datos	Código de Variable	Descripción
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
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

Conjunto de Datos	Código de Variable	Descripción
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
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)

Conjunto de Datos	Código de Variable	Descripción
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