METHODOLOGY NOTE V0 · DRAFT FOR REVIEW · NOVEMBER 2026

# Predictive compliance for underground storage tanks: data, model and validation

A working paper describing how SiteClear builds and tests its first-generation risk scores from public records. Author: Ananya Sudhir (draft prepared with engineering support; under review).

## 1. Problem

Underground storage tanks (USTs) at fueling, fleet, logistics and chemical-storage sites leak. EPA reports 581,676 confirmed releases from federally regulated USTs since 1984 and 53,777 still awaiting cleanup as of September 2025. Each year adds more than 4,000 new confirmed releases. Most compliance work is still calendar-driven: every site gets the same inspections and tests on the same schedule, regardless of risk. Predictive compliance asks a different question: given limited inspection, testing and maintenance capacity, *which sites should get attention first?*

## 2. Data

All inputs are public government records. A manifest of every file (URL, download date, size, SHA-256) is kept with the code.

- **EPA Semiannual Report of UST Performance Measures**, FY2021–FY2025: state counts of active and closed tanks, confirmed releases, cleanups and technical compliance rates. The tables are published as PDFs; we decode them programmatically and reconcile state sums with EPA's national totals.

- **EPA UST Finder**: a national composite of state tank, facility and release records (2018–2021 vintage), with location-based exposure fields (private wells and population within 1,500 ft, source-water and wellhead protection areas, floodplain).
- **Texas TCEQ Petroleum Storage Tank registry** (189,800 tanks, 84,393 facilities): install and removal dates, capacity, wall type, tank and piping material, corrosion protection and piping type.
- **Texas Leaking Petroleum Storage Tank (LPST) cases** (26,582 cases): reported and closure dates by facility.
- **National Response Center** incident reports, 2022–2025: self-reported incident causes.

## 3. Linking records

The same facility appears under different identifiers in different systems. In Texas, LPST cases carry the facility's PST registration number; 97.9% of case-facility links match a registry facility. EPA UST Finder's Texas IDs turned out to be TCEQ's internal facility key rather than the public facility number: joining on that key matched 99.98% of UST Finder's 61,243 Texas facilities, with name agreement of 91.3% and ZIP agreement of 83.7% as checks. 95.0% of the facilities in the 2022 test population have a linked EPA record.

## 4. Prediction design

The unit is a registered facility at a prediction date $T$ with at least one regulated tank in the ground at $T$. The label is whether a new LPST case is *reported* at that facility within 3 years of $T$. Features use only information available at $T$:

- Tanks in the ground at $T$ (installed before $T$ and not yet removed or filled in place), their number, total capacity and ages.
- Construction: single vs double wall (tank and piping), steel, fiberglass or composite tanks, corrosion protection, pressurized or suction piping, piping material, flexible connectors, shear valves, share installed before 1989.
- History: number of prior reported releases, years since the last one, and whether a case was still open at $T$.
- Facility type (retail, fleet refueling, industrial, emergency generator, etc.).

**Leakage check.** Registry files describe the present, not the past. We tested which fields change when a tank is removed. Construction fields stay populated; the "substance stored" field is cleared. Because releases are often discovered at removal, substance fields would

have leaked future information (an early model scored AUC 0.80 with them). They are excluded.

## 5. Model

The shipped v0 model is a regularized logistic regression on standardized features, chosen because each driver's direction and size can be read and challenged. A gradient-boosted tree model is trained as a comparison. Hyperparameters are chosen on the 2019 snapshot; final models are refit on 2010–2019 snapshots (labels end 2021) and scored once on the 2022 snapshot.

## 6. Validation

Test: 19,550 Texas facilities on January 1, 2022; 456 (2.3%) reported a release in 2022–2024. ROC AUC 0.705 (95% bootstrap CI 0.681–0.727); PR AUC 0.051 against a base rate of 0.023. The top 5%, 10% and 20% of facilities by score contained 14.9%, 28.5% and 46.1% of releases. Ranking by oldest tank alone captures 16.4% in the top 10%. Gradient boosting did not beat the logistic model (AUC 0.695). Adding location exposure did not improve prediction of release likelihood (AUC 0.708).

Performance varies by TCEQ region (AUC 0.62–0.76). 21% of releases occurred at facilities in the bottom half of scores, mostly newer double-walled sites. The model over-predicts slightly in the test period, so scores should be used for ranking.

## 7. National Risk & Efficiency Index (v0)

A state benchmark with three equally weighted pillars: **likelihood** (release rate per 1,000 USTs, share of pre-1989 tanks, technical non-compliance), **exposure** (share of facilities near private wells or in source-water protection areas) and **efficiency** (backlog share and releases confirmed per cleanup completed). Components are percentile-ranked across states and averaged. Rankings are stable to reweighting (Spearman 0.92–0.96) but individual states can move up to 21 places under a different normalization. Modeled facility risk will be added as more states are modeled.

## 8. Limitations

- One state. Texas is the largest UST program, but other states' data differ in structure and completeness.
- Labels are reported releases. Detection and reporting effort differ by region and over time.
- Public records lack operational signals (leak-detection alarms, inventory reconciliation, maintenance), which matter most for newer sites.
- Facility type and some attributes are current values, not values at the prediction date.
- Data vintages differ between sources (UST Finder 2018–2021; EPA measures FY2025; TCEQ 2026).

## 9. Next steps

- Extend to additional states with tank-level public data and release causes (candidates: California GeoTracker, Arizona ADEQ).
- Integrate EPA ECHO inspection and enforcement histories and PHMSA hazmat incidents, linked through EPA FRS.
- Pilot validation: compare scores with outcomes at 5–10 operator facilities and add operator data (tank tests, alarms) as features.
- Publish this note under the author's name after review, with code and data manifests.

Version 0 draft, generated from SiteClear model 0.1.0 (2026-11-22). Numbers update when the pipeline is re-run.