# EnviMetric White Paper

EnviMetric

November 2020

—

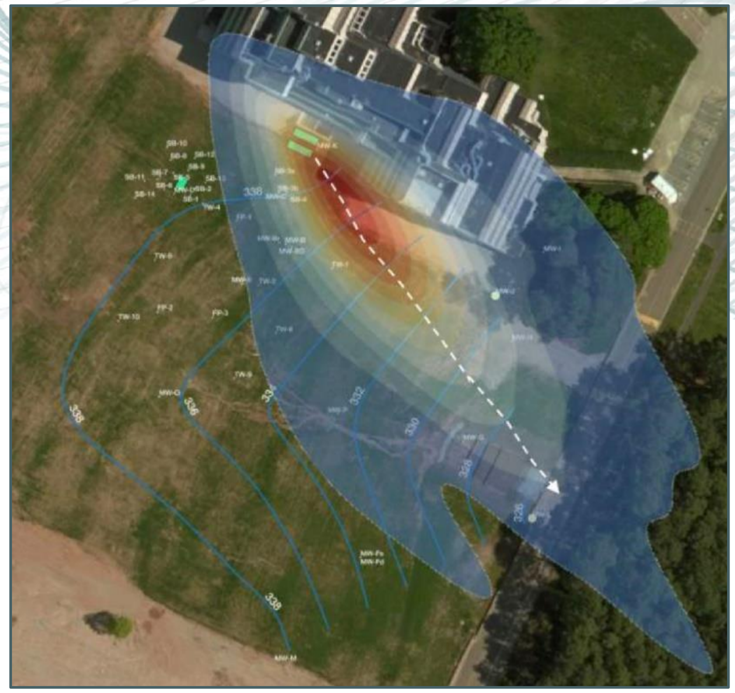
Azimuth1

—

contact@azimuth1.com

# BACKGROUND

In 2017, Azimuth1 received funding from the National Science Foundation to apply machine learning technology to the contaminated site investigation and remediation process. We did this by compiling and augmenting a database containing the findings from thousands of contaminated site investigations and then applying a machine learning model. By applying machine learning to the knowledge gained from thousands of previous site investigations, we are able to leverage thousands of observations to then predict contaminant dispersion at uncharacterized sites. The goal is the provide an additional line of evidence for an environmental investigation, emphasizing the common and most likely results consistent with sites that have similar soil, groundwater, climate, and topographic characteristics. This document will discuss the business case for applying EnviMetric to the site characterization process, along with our methods and the data set which supplies the machine learning model.



## THE PROCESS

### Business Case and Applications

Our innovation is a machine learning method called EnviMetric, which gives engineers tools to shift their business process toward A high performance approach, by characterizing and modeling sites right from the start, and improving these models with additional data rather than spending time and resources collecting initial data before a starting model is ever generated.

**The value proposition for EnviMetric is that it provides environmental consulting firms a higher success rate, and provides property owners a reduction in the total cost of remediation over current methods at a price that is less than 5% of their current project budget.**

We have found a couple different applications of this approach to be the most interesting to people we have talked to in the industry.

- Early Investigation: the first is to use the machine learning model during early investigation; when little is known about a site (perhaps only a couple of temporary monitoring wells have been installed), we take the limited data set and use it to run the machine learning model to make a prediction about the most likely extent of contamination based on what has happened in thousands of similar cases.
- Addressing Data Gaps: the second application is addressing data gaps during investigation; when dealing with issues sampling under buildings and highways or with off-site access issues, take your existing data set and use to model to find and address in any existing data gaps.
- Source Zone Identification: when dealing with multiple source zones, unknown source zones, or comingled plumes, we can run the machine learning model based on the existing data set to provide an associated prior probability for each candidate source zone.
- Portfolio Management: one final application that has come up recently from companies who manage a portfolio of sites and state regulatory agencies is applying the model to all sites within their portfolios to help prioritize their workload.

## Our Dataset

We've compiled what is, to our knowledge, the largest database of contaminated site data in the world. We did this primarily by approaching individual states' departments of environmental quality to obtain their data, which had the level of detail we needed (detailed soil, depth, and groundwater profiles). The data are then categorized according to contaminant type, soil conditions, groundwater conditions, climate, and age of the site. As you might imagine, data from different DEQs can be quite different in structure, quality, and formatting.

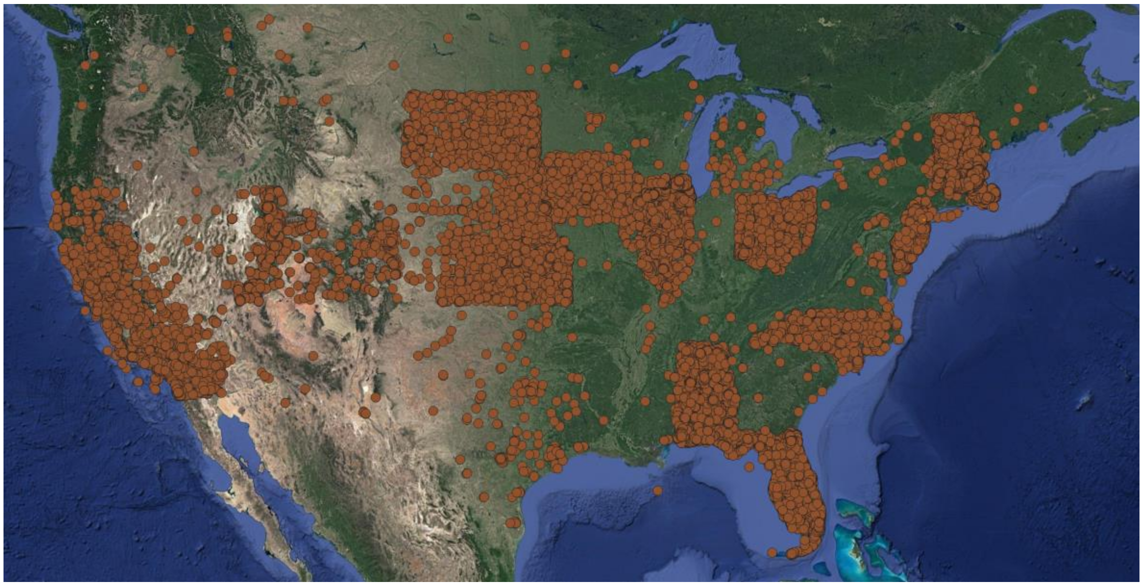


*Figure 1 Current EnviMetric Dataset*

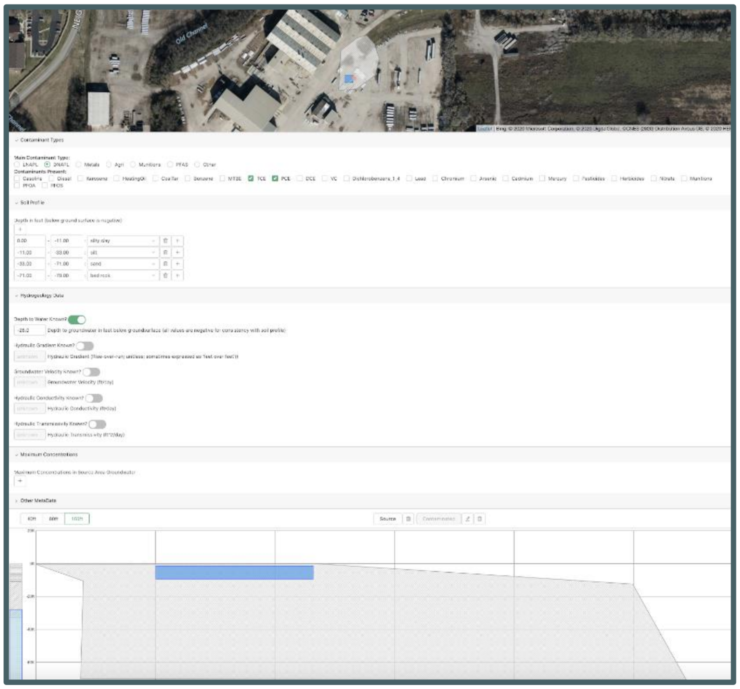
Combining everything into a cohesive national dataset has been a big effort that we are continuing still. We have over 86,000 sites collected in their raw form. This has information on the location, dates of investigation, and compounds detected. We dive into each report, digitizing the impacted area, and cataloging several dozen additional parameters from each site. We currently have 7104 augmented sites in our database.

To manage this data, we built an internal tool called Groundtruth that our team uses to extract data from report text and graphics, cataloging it in a neatly formatted database structure. We collect the following parameters in Groundtruth:

- 3D Contaminant Extent
- 3D Source Zone Location
- Contaminant Type
- Maximum Contaminant Concentrations
- Hydrogeology Data
    - Depth to water
    - Hydraulic gradient
    - Groundwater velocity
    - Hydraulic conductivity
    - Hydraulic transmissivity
- Geochemistry Data
    - pH
    - Oxidation reduction potential

- Dissolved oxygen
- Total organic carbon of soil

- Geologic Profile of site
- Other site-specific parameters



Contaminant Types

Main Contaminant Type:
○ LNAPL ● DNAPL ○ Metals ○ Agri ○ Munitions ○ PFAS ○ Other

Contaminants Present:
□ Gasoline □ Diesel □ Kerosene □ HeatingOil □ CoalTar □ Benzene □ MTBE ☑ TCE ☑ PCE □ DCE □ VC □ Dichlorobenzene_1_4 □ Lead □ Chromium □ Arsenic □ Cadmium □ Mercury □ Pesticides □ Herbicides □ Nitrate □ Munitions □ PFOA □ PFOS

Soil Profile

Depth in feet (below ground surface is negative)

| | | |
|---|---|---|
| 0.00 | -11.00 | silty clay |
| -11.00 | -33.00 | silt |
| -33.00 | -71.00 | sand |
| -71.00 | -79.00 | bed rock |

Hydrogeology Data

Depth to Water Known? ●

-25.2 Depth to groundwater in feet below groundsurface (all values are negative for consistency with soil profile)

Hydraulic Gradient Known? ○

unknown Hydraulic Gradient (Rise-over-run) unitless: sometimes expressed as 'feet over feet')

Groundwater Velocity Known? ○

unknown Groundwater Velocity (ft/day)

Hydraulic Conductivity Known? ○

unknown Hydraulic Conductivity (ft/day)

Hydraulic Transmissivity Known? ○

unknown Hydraulic Transmissivity (ft^2/day)

Maximum Concentrations

Maximum Concentrations in Source Area Groundwater

Other MetaData

Old Channel
Leaflet | Bing © 2020 Microsoft Corporation, © 2020 Digital Globe, ©CNES (2020) Distribution Airbus DS, © 2020 HERE
40ft
80ft
160ft
Source
Contaminated
20ft
0ft
-20ft
-40ft
-60ft

*Figure 2 Groundtruth Portal*

## Using Machine Learning to make Site Specific Predictions

The machine learning algorithm is trained using subsets of the database containing similar sites; the algorithm then produces a statistical prior probability model automatically. There are millions of combinations of input parameters that must be weighted and selected in the right proportions to predict the output variables. The machine learning model uses the provided input parameters to make decisions about which sites are most likely to have similar conditions to the site in question.

The algorithm's purpose is not to predict exactly the resulting contaminated zone, but to provide the weight of evidence from many previously observed cases, and serve as a starting point for further refinement, reducing uncertainty and providing confidence bounds on the extent or source of a known contaminant. Using many observations serves to filter out variation and determine if there are consistent migration patterns that can be used to guide a site investigation. To do this we first shift each site polygon into a common orientation. Next, each plume geometry is oriented so that the groundwater gradient is oriented in the X axis direction. We then import the plume geometries into the machine learning model to create a forecast for future plumes. Specifically, we are using a geospatial kernel density estimation that produces a probability estimate for the extent and dispersion of future plumes.

The EnviMetric model then outputs two results- an unknown source model and an unknown destination model. The unknown destination model shows the highest probability estimated for the location of the farthest detectable contaminated area down hydraulic gradient from the contamination source zone. The second output is the unknown source model. This is for situations where contamination is detected, but the source of the contaminant is unknown. This often occurs when one property is contaminated, while a neighboring property may be the source of the contamination.

## EnviMetric Output

EnviMetric employs an ensemble model which trains many diverse models and combines the outputs of these models to provide a better prediction than could be obtained by a single model alone. The models evaluate all available information including (but not limited to): site location, lithology, contaminant types, contaminant concentrations, groundwater flow conditions, and release type. In addition to providing a more accurate combined prediction, the ensemble approach also provides a distribution of predictions, rather than simply a single point estimate.

In addition to the output of the EnviMetric model, we also provide summary statistics of sites in our database - both broad categories, such as all sites in the US, and more specific applicable subsets, such as all LNAPL sites in the state. These summary statistics are designed to provide a broader context in which to interpret the results of the EnviMetric model and where this particular site fits into the broader set of previously investigated contaminant plumes.

In addition to calculating probability ranges for contaminant distribution, EnviMetric also generates a single value, i.e. the model's 'best guess'. This is the value the model predicts with the highest confidence (the 50th percentile EnviMetric model output).

The contours in Figure 3 represent total probability that the contaminant is fully contained in that contour volume. Each contour represents a 3D volume that can be associated with remediation cost of treatment or excavation.

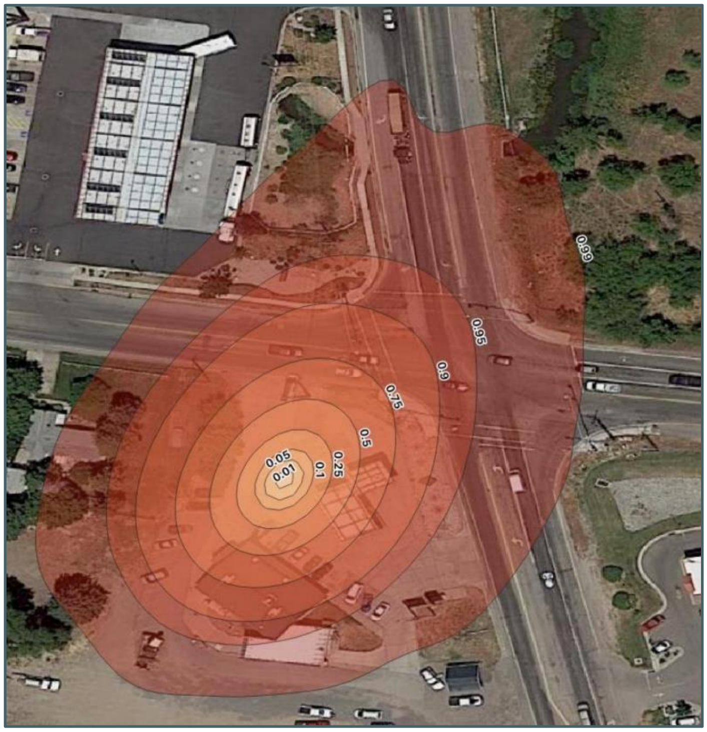
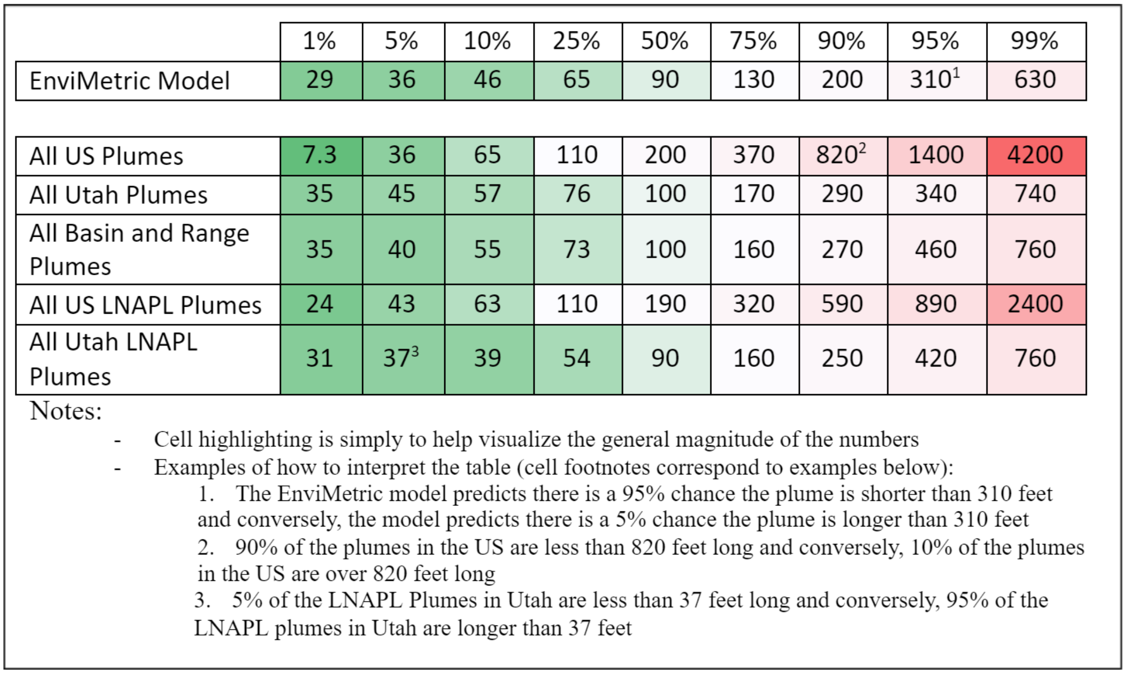


*Figure 3 Example EnviMetric Model Output*

Table 1 shows an example of the model's confidence that the plume length is shorter than the distance shown (in feet) along with the percentages of different data subsets with plumes shorter than the distance shown (see Table 1 footnotes for further explanation). For the EnviMetric Model (row 1 of Table 1), the provided confidence intervals are based on fitting a kernel density estimation to the individual model predictions. For confidence intervals provided using subsets of the EnviMetric database (rows 2 through 6 of Table 1), the provided confidence intervals are based on known plumes within the dataset. We generate this table for the length, width, and depth of each modeled plume.

*Table 1 Example EnviMetric Distribution Plume Length Table*



| | 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% |
|---|---|---|---|---|---|---|---|---|---|
| EnviMetric Model | 29 | 36 | 46 | 65 | 90 | 130 | 200 | 310[1] | 630 |
| | | | | | | | | | |
| All US Plumes | 7.3 | 36 | 65 | 110 | 200 | 370 | 820[2] | 1400 | 4200 |
| All Utah Plumes | 35 | 45 | 57 | 76 | 100 | 170 | 290 | 340 | 740 |
| All Basin and Range Plumes | 35 | 40 | 55 | 73 | 100 | 160 | 270 | 460 | 760 |
| All US LNAPL Plumes | 24 | 43 | 63 | 110 | 190 | 320 | 590 | 890 | 2400 |
| All Utah LNAPL Plumes | 31 | 37[3] | 39 | 54 | 90 | 160 | 250 | 420 | 760 |

Notes:

- Cell highlighting is simply to help visualize the general magnitude of the numbers
- Examples of how to interpret the table (cell footnotes correspond to examples below):
  1. The EnviMetric model predicts there is a 95% chance the plume is shorter than 310 feet and conversely, the model predicts there is a 5% chance the plume is longer than 310 feet
  2. 90% of the plumes in the US are less than 820 feet long and conversely, 10% of the plumes in the US are over 820 feet long
  3. 5% of the LNAPL Plumes in Utah are less than 37 feet long and conversely, 95% of the LNAPL plumes in Utah are longer than 37 feet

Figure 4 shows the distribution of the kernel density estimates and represents the information provided in Table 1 graphically. The Y-axis of Figure 4 is the percentage of predictions (i.e. percentage of plumes within the EnviMetric database, or within the referenced subset of the EnviMetric database) at that particular plume length. Figure 4 shows that the EnviMetric model follows a similar distribution shape as the data subsets. For this site, the majority of models in the ensemble predict a short plume, while a smaller subset predicts a significantly longer plume.

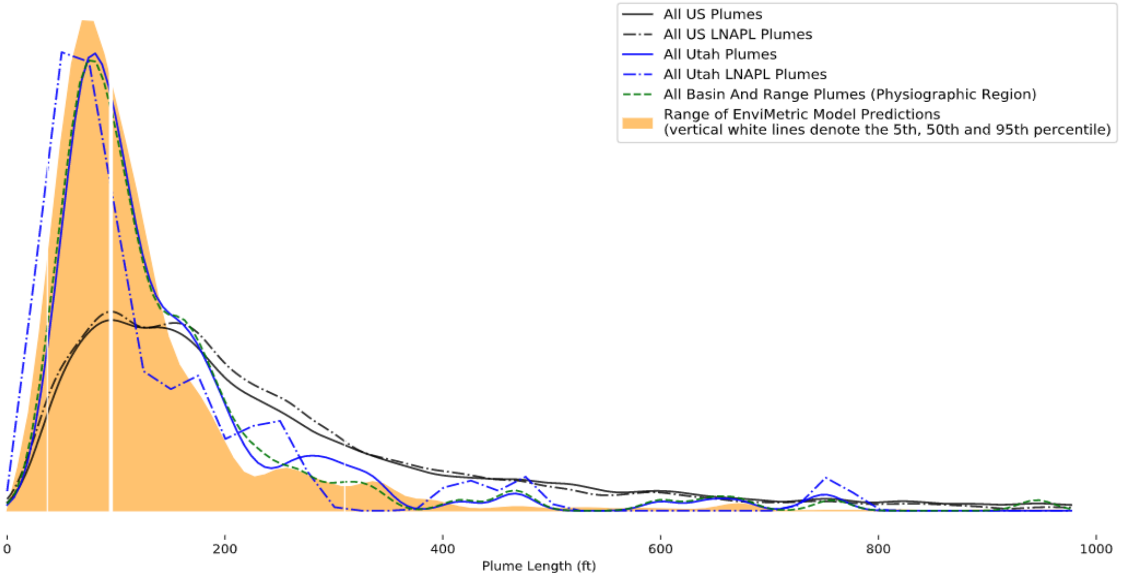
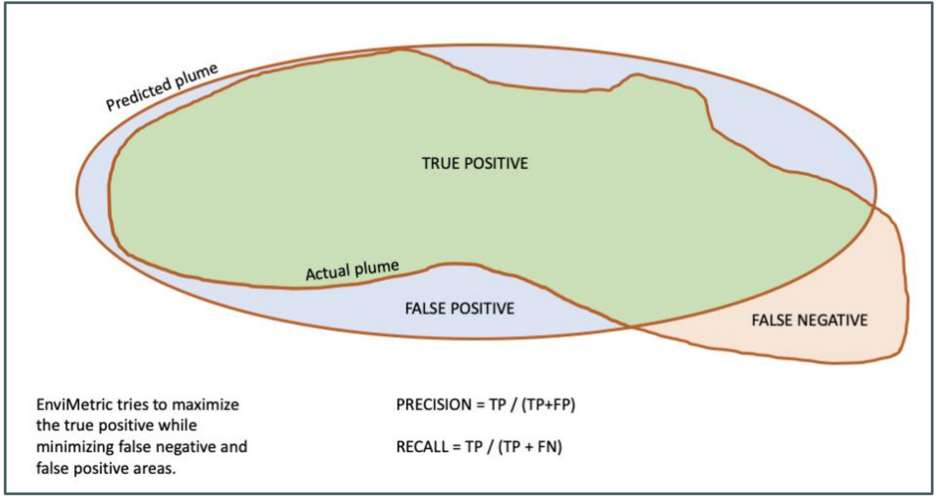


*Figure 4 Example EnviMetric Distribution of Plume Length*

## Metrics of Performance

We measure the performance of our model using area statistics. Our true positive area is where we predicted that a plume is in an area, and the actual plume covers the same area. A false positive is when we predicted the plume in an area, but the contamination is absent. The false negative area is where we predicted there was no plume, but there is actual contamination. We measure these parameters across our entire database using a process called cross validation, where we train the model on a partial set of data and test on the rest. Then we rotate the hold out sets many times to get a statistical view of how accurate the model is overall.



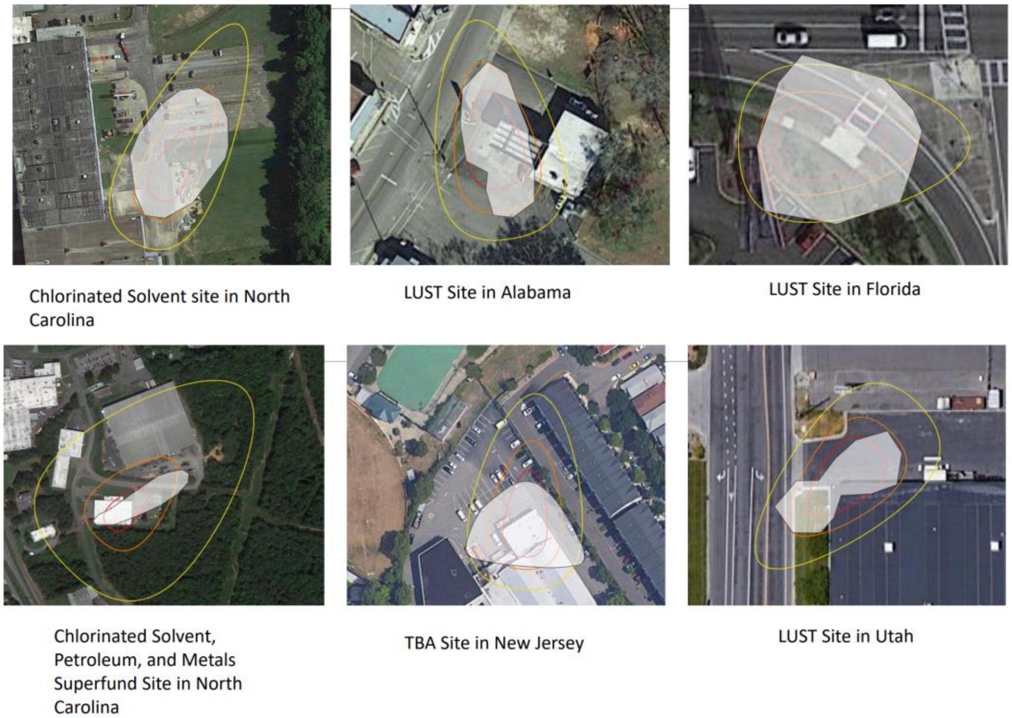
EnviMetric tries to maximize the true positive while minimizing false negative and false positive areas.

PRECISION = TP / (TP+FP)

RECALL = TP / (TP + FN)

*Figure 5 Metrics of Performance*

Here are a couple of examples of where an application of our machine learning model was applied to estimate the contaminant extent. The partially opaque white polygon is the real observed site conditions. The red, orange and yellow outlines show the machine learning model's output; the red is the 25th percentile estimate, the orange is the 50th percentile estimate (this is our 'best guess' as to what is happening at the site), and the yellow is the 75th percentile estimate.



Chlorinated Solvent site in North Carolina

LUST Site in Alabama

LUST Site in Florida

Chlorinated Solvent, Petroleum, and Metals Superfund Site in North Carolina

TBA Site in New Jersey

LUST Site in Utah

*Figure 6 Example EnviMetric Applications*

For the first row of examples, the model's 50th percentile estimate performed quite well. For the second row of examples, you can see that in these instances the actual plumes are such different shapes than the model output. There are a couple different reasons for this: for example discontinuities in the subsurface, infrastructure and others. For the following sites we've identified site conditions which contributed to the poorer performance of the model.

- Chlorinated solvent site in NC: In addition to having several different contaminant types present at this site, the site is located at the crest of a star shaped hill, which we believe impacts groundwater flow
- TBA site in NJ: This site is right next to a river, so we think that groundwater-river interactions contributed to the disparity between real world conditions and the modeled output
- LUST site in Utah: for this leaking underground storage tank site in Utah the model does a good job of predicting the real plume's length, but our has a model regular formation; a non-standard plume geometry is one limitation of the model.

**INTERESTED IN LEARNING MORE?**

**Please Contact us at:**

- **Azimuth1.com**
- **contact@azimuth1.com**
- **(703)-618-8866**
- **1751 Pinnacle Drive, Suite 600, McLean, VA, 22102**