Analysis of Chlorinated Hydrocarbon Concentration Data from Thousands of Groundwater Wells Using a Density-Based Cluster Analysis Approach

Walt W. McNab
Roux Associates
Oakland, California

Abstract

A meta-study was conducted of chlorinated volatile organic compounds (CVOC) detections in environmental monitoring wells, as reported in California's Groundwater Ambient and Monitoring Assessment (GAMA) database. The list of CVOCs assessed included 1,1-dichloroethane (1,1-DCA), 1,1-dichloroethene (1,1-DCE), *cis*-1,2-dichloroethene (*cis*-1,2-DCE), *trans*-1,2-dichloroethen (*trans*-1,2-DCE), tetrachloroethene (PCE), 1,1,1-trichloroethane (1,1,1-TCA), 1,1,2-trichloroethane (1,1,2-TCA), trichloroethene (TCE), vinyl chloride, and, additionally, 1,4-dioxane. A machine learning technique – DBSCAN cluster analysis – was used to delineate approximately 17,000 monitoring wells, distributed across the state of California, into 1,183 "sites" with one or more or CVOC groundwater contaminant plumes, based on well coordinates. The total number of computed sites was found to depend on the specified maximum search radius parameter of the DBSCAN method: too large of a search radius resulted in a merging of sites, while too small of a search radius generated more outlier wells that were excluded from cluster assignment, thereby reducing the number of small sites featuring only a minimal number of monitoring wells. As a subsequent step, Delaunay triangulation was used to quantify the spatial extents of the monitoring well networks at each site and to estimate CVOC mass per unit aquifer depth. The study illustrates how aggregate groundwater contaminant plume behavior can be readily quantified, given the availability of current environmental databases and accessible data analysis tools.

Keywords: Groundwater contaminant plumes, machine learning, cluster analysis

Introduction

Meta-studies of groundwater contaminant plumes, entailing statistical analyses of compiled data summaries from multiple sites, can be useful for identifying broad trends in contaminant behavior that can be difficult to extrapolate from individual sites. Meta-study results can be used to assess the efficacy of both engineered as well as natural attenuation remedies, behavioral comparisons between contaminants, and benchmarks for the extent, and therefore cost, of plume characterization in comparison to other similar conditions. Past efforts to identify and apply insights from plume meta-studies first focused on fuel hydrocarbons and methyl tert-butyl ether (MTBE), and in some case attempts to quantify impacts of biodegradation (e.g., Rice et al.,

1995; Mace t al., 1997; Mace and Choi, 1998; Connor et al., 2015). Subsequent applications of meta-analyses for quantifying CVOC plume behavior included studies by McNab et al. (2000), McNab (2001), Newall et al. (2006), and Faybishenko and Hazen (2009). More recently, Adamson et al. (2014, 2015) reviewed data from multiple California and Air Force sites to characterize frequency of co-detections of 1,4-dioxane with other CVOCs and characterized plume length, while Lee et al. (2016) employed a machine learning scheme to predict the efficacy of engineered bioremediation based on data collected across 35 sites.

The availability of (1) online groundwater quality databases in the public domain, and (2) new tools to facilitate the assimilation and processing of such data, provides a means for expeditiously conducting meta-studies on a broader scale than was previously possible. In this study, groundwater chlorinated volatile organic compound (CVOC) concentrations reported from the California Stae Water Resources Control Board's Groundwater Ambient Monitoring and Assessment (GAMA) database were compiled and quantitatively evaluated using open-source data analysis tools. The assessment entailed groundwater samples from all wells in the database across all sample dates where at least one CVOC was detected above the applicable laboratory analytical detection limit; the list of CVOCs included 1,1-dichloroethane (1,1-DCA), 1,1-dichloroethene (1,1-DCE), *cis*-1,2-dichloroethene (*cis*-1,2-DCE), *trans*-1,2-dichloroethen (*trans*-1,2-DCE), tetrachloroethene (PCE), 1,1,1-trichloroethane (1,1,1-TCA), 1,1,2-trichloroethane (1,1,2-TCA), trichloroethene (TCE), vinyl chloride, and/or 1,4-dioxane. The groundwater sample dataset was parsed into individual "sites" and CVOC plumes using density-based spatial-clustering-of-applications-with-noise (DBSCAN) cluster analysis. This exercise yielded almost 1,200 putative sites across the state, each with one or more CVOCs. Site and plume metrics quantified for this parsed dataset included monitoring well network length and associated area, maximum historic concentrations of individual CVOCs, and integrated CVOC mass per unit aquifer thickness.

The methodology and example findings of this meta-study are summarized below.

Methodology

Data from the GAMA database for all 58 California counties, representing a reporting period from approximately 2000 through the present, were used in the assessment. The data were limited to those flagged as being reported under the Electronic Data Format, or EDF, a means through which environmental monitoring well data (as opposed to public water supply well data) are submitted to GAMA. Data from individual counties were first downloaded from the GAMA website (http://geotracker.waterboards.ca.gov/gama/datadownload) as individual text files. These were compiled into a single dataframe structure using a python-language script, along with the pandas data processing package (http://pandas.pydata.org/). The data were filtered by individual sample event in each well so that at least one CVOC from the list provided above was reported as a detection. Duplicate detections were averaged; non-detects were designated as missing data (i.e., as "NA" items within the dataframe). Assessments conducted on this individual reported

detection dataset included a graphical analysis of the frequency of co-detections between individual CVOCs, and identifying possible evidence for the impact of reductive dehalogentation reactions on the TCE → *cis*-1,2-DCE → vinyl chloride sequence. The latter analysis involved using co-detections of manganese, also reported in the GAMA database for the same individual groundwater samples, as a proxy for local oxidation-reduction conditions.

Separately, individual sites were then identified from among all the wells in the GAMA database representing the individual CVOC detection dataset. This was conducted using DBSCAN cluster analysis, as implemented in the scikit-learn machine learning package for python (http://scikit-learn.org/stable/). In summary, DBSCAN clustering assigns points – in this case, spatial survey coordinates for the wells in two-dimensions – to clusters based on whether or not the points can be connected to other points directly or indirectly (via a pathway through other cluster members) within a specified search radius (Ester et al., 1996). All data from a particular well were dropped from this analysis if (1) the well was not assigned to a cluster (i.e., the well was labelled as an isolated geographic outlier), or (2) the cluster was comprised of less than five wells. Beyond this computational binning of individual wells into sites, no attempt was made to compare the posited clustering assignments against specific reported site histories as described in characterization or remediation report databases. Such a validation effort would have been infeasible, given the number of reports and the geographic distribution of monitoring well data across the entire state.

A local example of the resultant clustering of monitoring well data using the DBSCAN method is shown on Figure 1. In general, this machine learning technique groups the data points in a manner that visually appeals to intuition, in the absence of any further data about the individual "sites" that are implied. However, the maximum search radius for cluster membership is a key parameter determining cluster compositions. At distances that are too large, clusters that would otherwise be delineated as distinct are instead merged. At distances that are too short, an increasing number of wells are labelled as isolated and removed from the analysis, therefore deleting some candidate sites that fail to meet the minimum number of wells to qualify as a cluster. This effect is illustrated by a sensitivity analysis summary to the maximum search radius as shown on Figure 2. Note that the maximum search radius is expressed as degrees, (i.e., latitude-longitude), as the analysis encompasses the entire state and planar projection systems such as Universal Transverse Mercator or the state plane use different baselines for various portions of the state.

The sensitivity analysis indicates that the optimal maximum search radius which generates the largest number of clusters, and hence furnishes the best discriminating power, is approximately 0.00075 degrees. This is equivalent to approximately 80 meters along the north-south direction and to somewhat shorter distances in the east-west direction, depending on location with the state. The example clusters shown on Figure 1 are based on this inferred optimal maximum search distance.

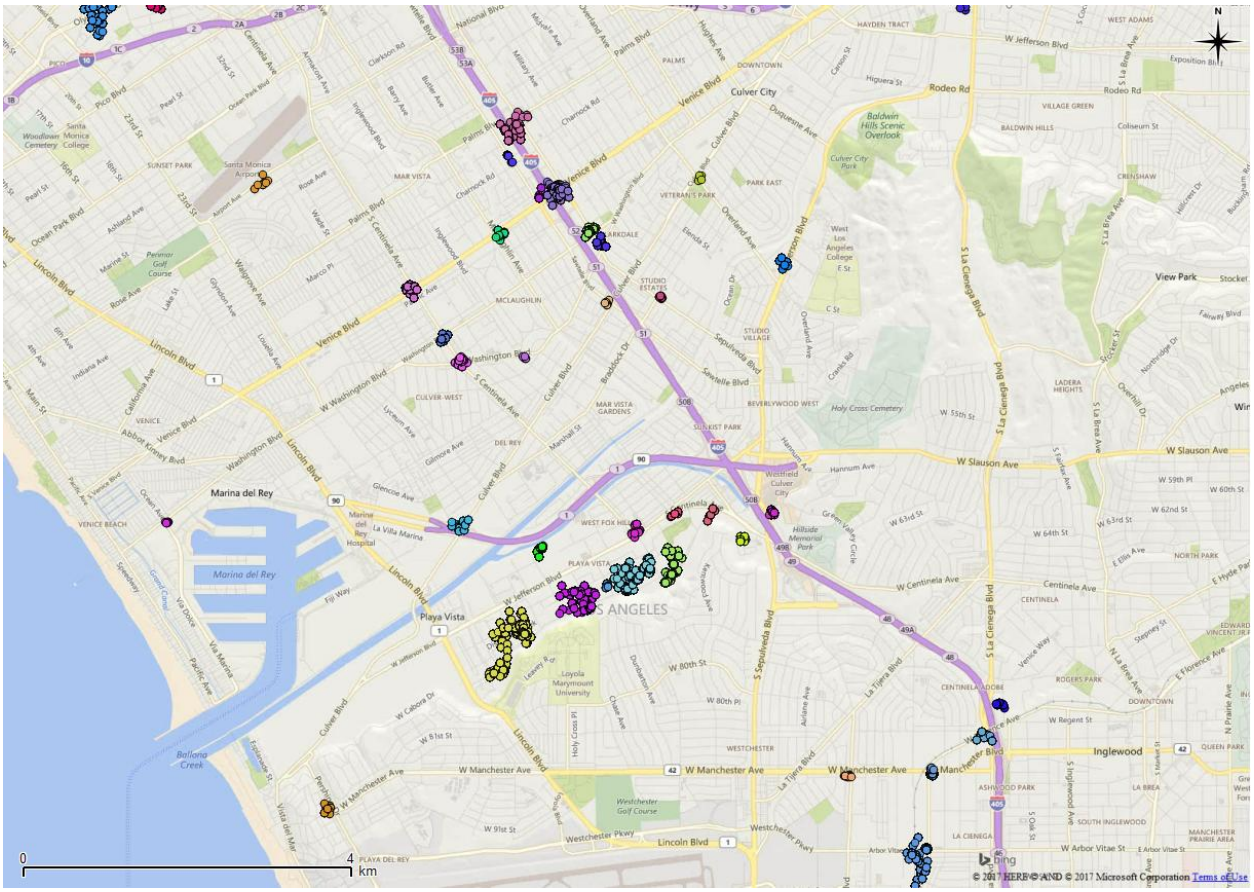


Figure 1: Example assignments of wells (circular symbols) to "sites" based upon DBSCAN clustering (indicated by color). Such assignments do not necessarily reflect any association of a particular well with any regulatory designation or remediation effort but are instead attributable purely to geographical placement.

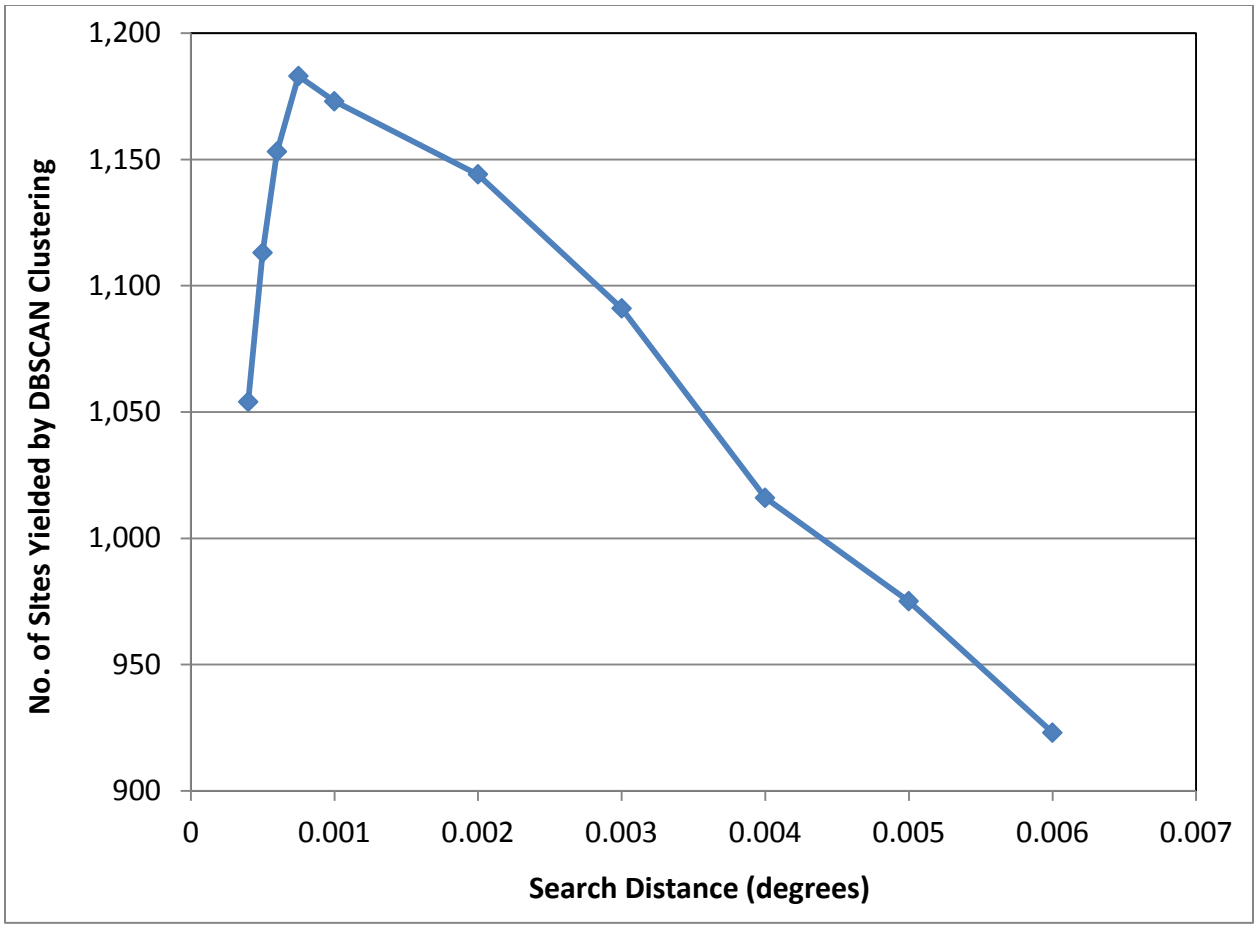


Figure 2: Response of DBSCAN cluster number count to maximum search radius for cluster membership.

Once monitoring wells were assigned to clusters, or sites, the latitude-longitude coordinates were converted to a Cartesian coordinate system using the pyproj geospatial computations library (https://pypi.python.org/pypi/pyproj). This step facilitated:

1. Calculation of maximum distances between monitoring well networks at each site; and,
2. Calculation of the areas encompassed by the site monitoring wells.

Both calculations were accomplished using the python scipy spatial package (https://www.scipy.org/). The first entailed computing the distance matrix for all monitoring well combinations with at least one historic CVOC detection at each site and then extracting the maximum value from the matrix. The second calculation involved using Delaunay triangulation to delineate each site into triangles, with the monitoring well survey points serving as the vertices. Summation of the areas associated with each triangle yields the area encompassed by the convex hull of the respective monitoring well sets. An example application of the Delaunay triangulation approach for two adjacent sites is shown on Figure 3. Note that this figure also illustrates the delineation of monitoring well survey points into separate clusters by DBSCAN

analysis: each well within a given cluster is within approximately 80 meters of another cluster member well, whereas the gap between the two sites exceeds the maximum search radisu for all possible well pairs.

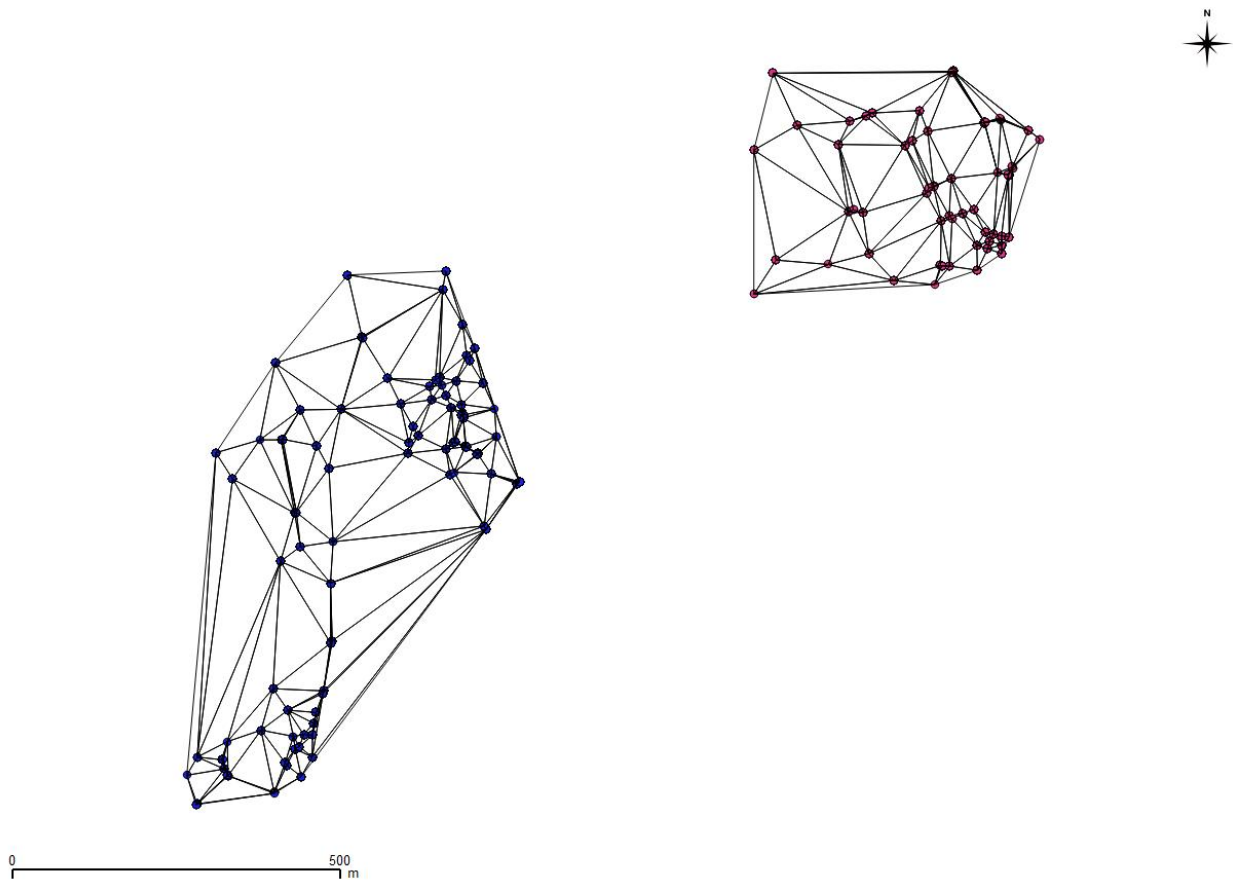


Figure 3. Results of Delaunay triangulation applied to two adjacent sites in an urban area; the figure also illustrates the efficacy of the DBSCAN algorithm in delineating sites, as constrained by the maximum search radius.

For each CVOC at each site, mass per unit depth can be estimated by multiplying the areas of each triangle by the median historical concentration of the vertices, and then subsequently summing the areas while multiplying by an assumed value for porosity. Mass per unit aquifer depth is the appropriate CVOC site-impact metric, as opposed to total CVOC integrated mass, because well screen interval data are not provided in the GAMA database. As a consequence, the three-dimensional component of the CVOC distribution is implicitly mapped into two dimensions.

Findings

Using DBSCAN, a total of 1,183 sites with at least one CVOC plume were identified from the GAMA database using the optimal maximum search radius. The distribution of these sites across California is shown on Figure 4. Unsurprisingly, these sites reside primarily in major urban areas such as Los Angeles, San Diego, the San Francisco Bay Area, and Sacramento, with scattered sites also found along the south central coast and in the San Joaquin Valley. This reflects the density of industrial sites, dry cleaners, and other potential sources of CVOCs in these areas.

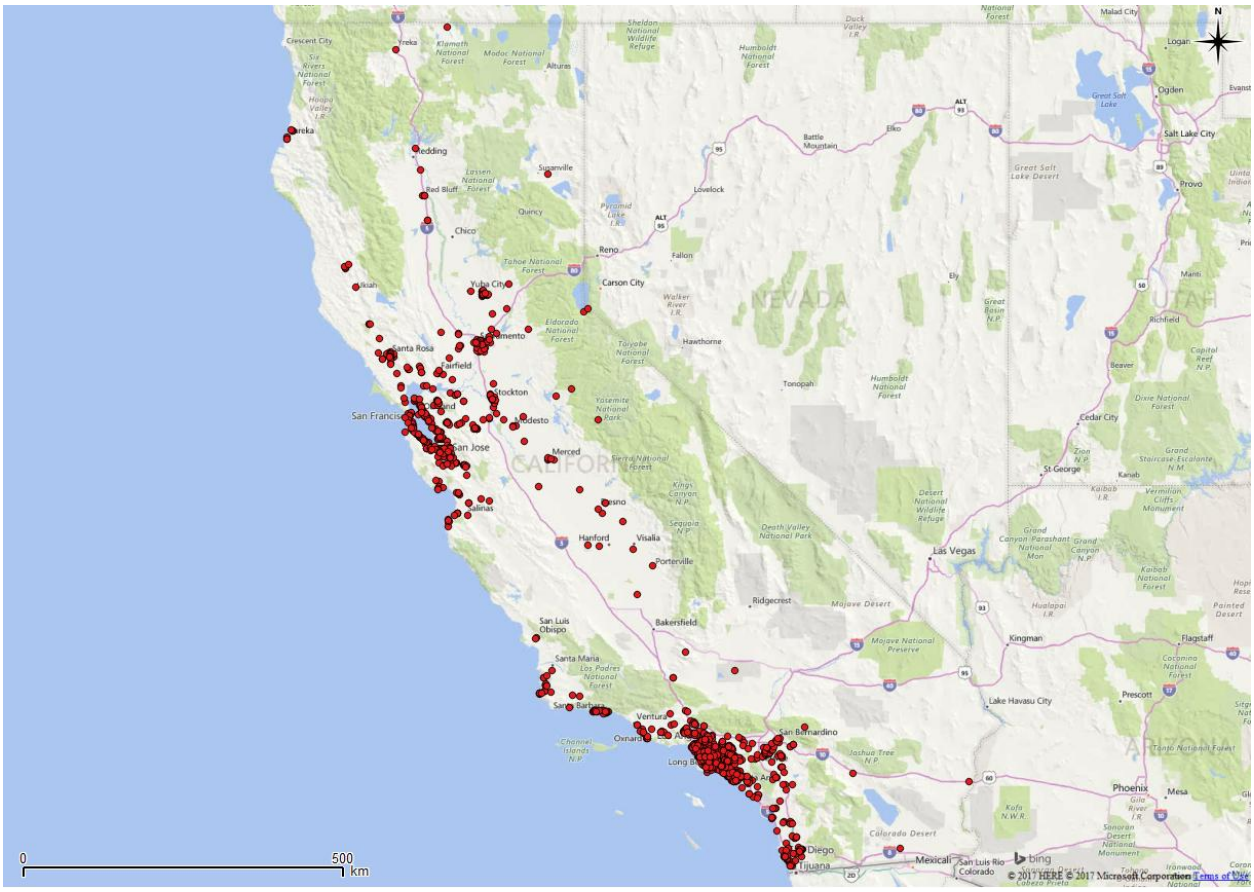


Figure 4: Locations of 1,183 sites with one or more CVOC plumes identified by DBSCAN cluster analysis.

The 1,1,83 sites represent 16,951 individual monitoring wells, or, in some instances, other surveyed groundwater sample locations, with unique survey coordinates. From among this population of wells, a total of 150,245 sample events were reported which featured at least one CVOC detection.

Patterns Among Sample Event Populations

The total number of reported detections of individual CVOCs among the sample event population is shown on Figure 5. Over 100,000 detections of TCE are represented in the dataset, with slightly fewer numbers of detections of two other chloroethenes – *cis*-1,2-DCE and PCE – with over 80,000 reported detections each. The least frequently reported detections include the two trichloroethane isomers (1,1,1-TCA and 1,1,2-TCA) and 1,4-dioxane, each with fewer than 20,000 detections.

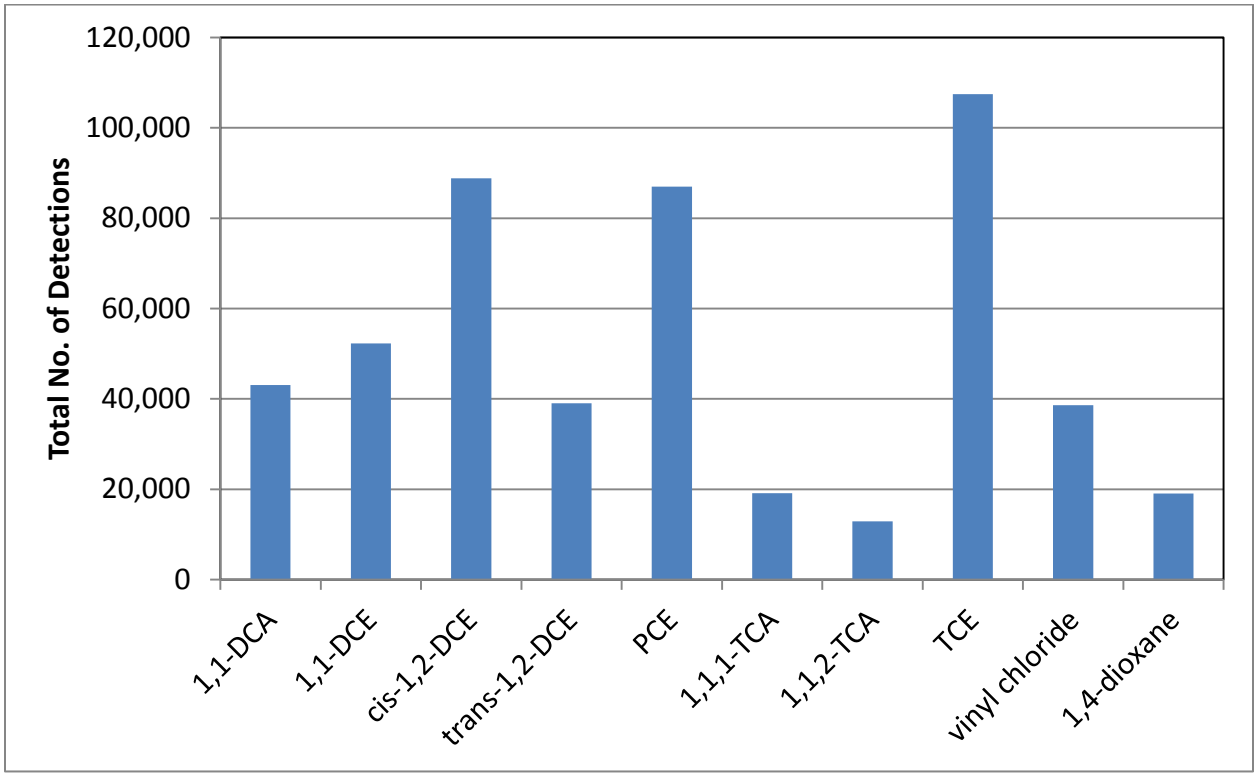


Figure 5: Total number of samples from filtered GAMA database indicating positive reported detections of individual CVOCs.

Co-detections of CVOC pairs are summarized on the graph shown on Figure 6, with the nodes representing the number of detections per analyte and the edges representing the number of co-detections. The graph indicates, for example, that the three most commonly detected CVOCs also tend to be commonly co-detected in individual samples (i.e., TCE with *cis*-1,2-DCE and/or PCE). Normalizing the number of co-detections by the number of detections of a given CVOC pair member yields an estimate of the co-detection frequency. For example, the co-detection frequencies for 1,4-dioxane with respect to the other CVOCs are summarized on Figure 7, indicating that 1,4-dioxane is more likely to be found with 1,1,1-TCA or its two potential

degradation products, 1,1-DCE and 1,1-DCA, than the other CVOCs. This finding is consistent with the use of 1,4-dioxane as a stabilizing agent for 1,1,1-TCA (Adamson et al., 2014).

Inclusion of additional analytes that are indicative of the local geochemical environment can provide additional insights into processes impacting CVOC concentrations. Specifically, reported co-detections of oxidation-reduction condition indicators, such as dissolved manganese concentrations, can be used to assess the impact of reductive dehalogenation on some CVOCs. For example, concentrations of vinyl chloride and manganese are compared on Figure 8; the highest concentrations of vinyl chloride are associated with elevated concentrations of manganese, whereas much scatter exists in the relationship at lower concentrations. A plausible explanation is that very reducing conditions are necessary for the generation of high vinyl chloride concentrations in what are presumably wells located proximal to plume source areas, whereas lower concentrations may be from more proximal plume areas and are thus impacted more by transport processes as opposed to local oxidation-reduction conditions. In contrast, a comparison of TCE and manganese concentrations from the same data set exhibits no discernible relationship. In this case, high TCE concentrations appear to be associated with low concentrations of manganese in some samples, presumably because the introduction of TCE at high concentrations from nearby sources does not require reducing conditions.

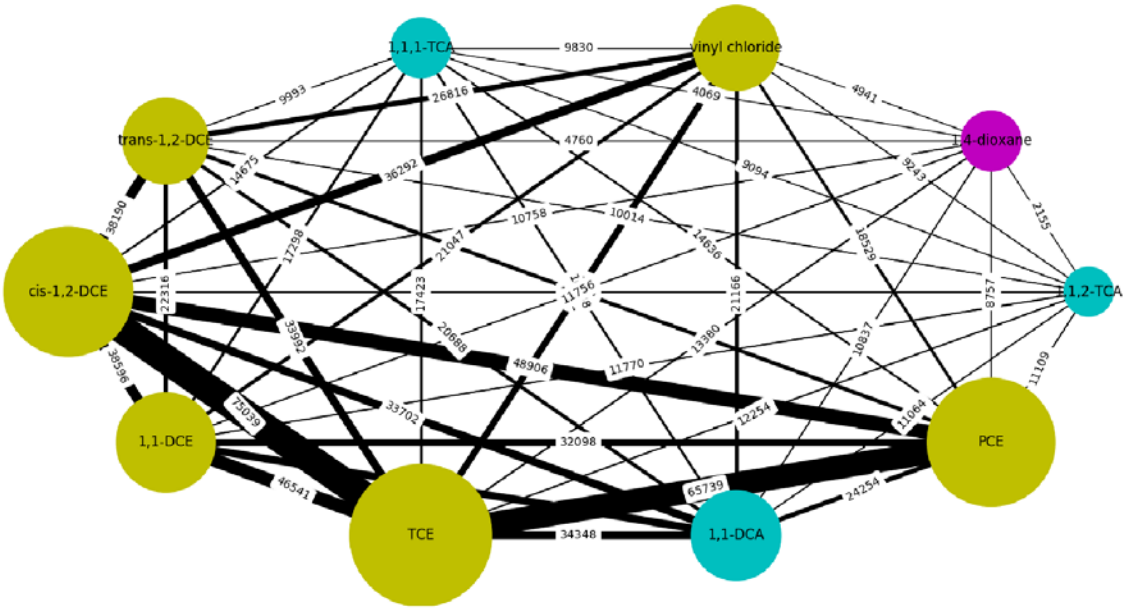


Figure 6: Graph representing the number of detections (nodes) and the number of co-detections (edges) of CVOCs in filtered individual sample dataset. Graph generated using the networkx package (https://networkx.github.io/) for python.

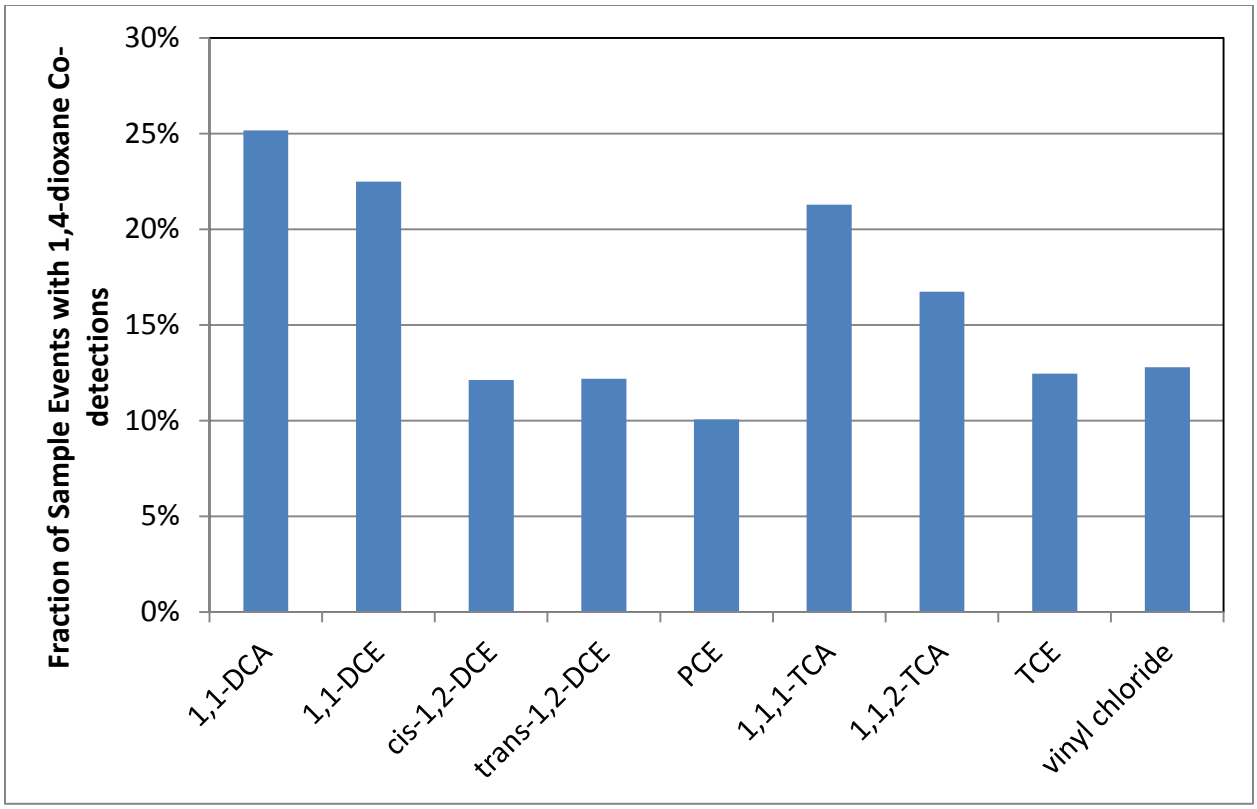
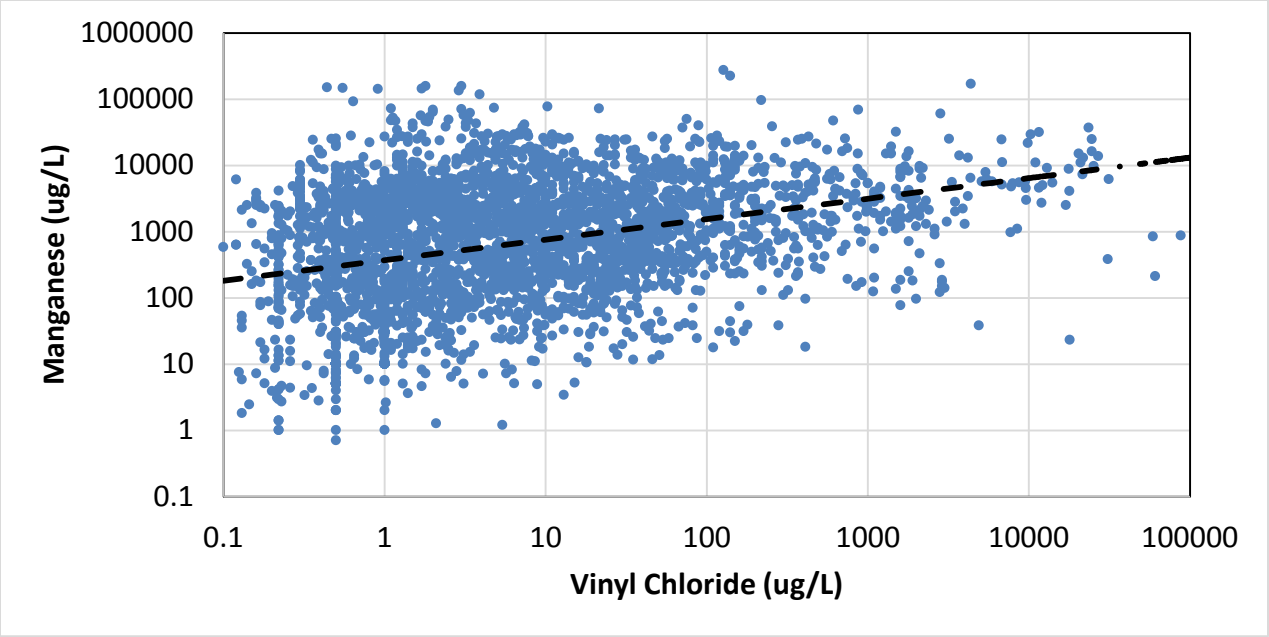


Figure 7: Fraction of individual samples characterized by co-detection of 1,4-dioxane.



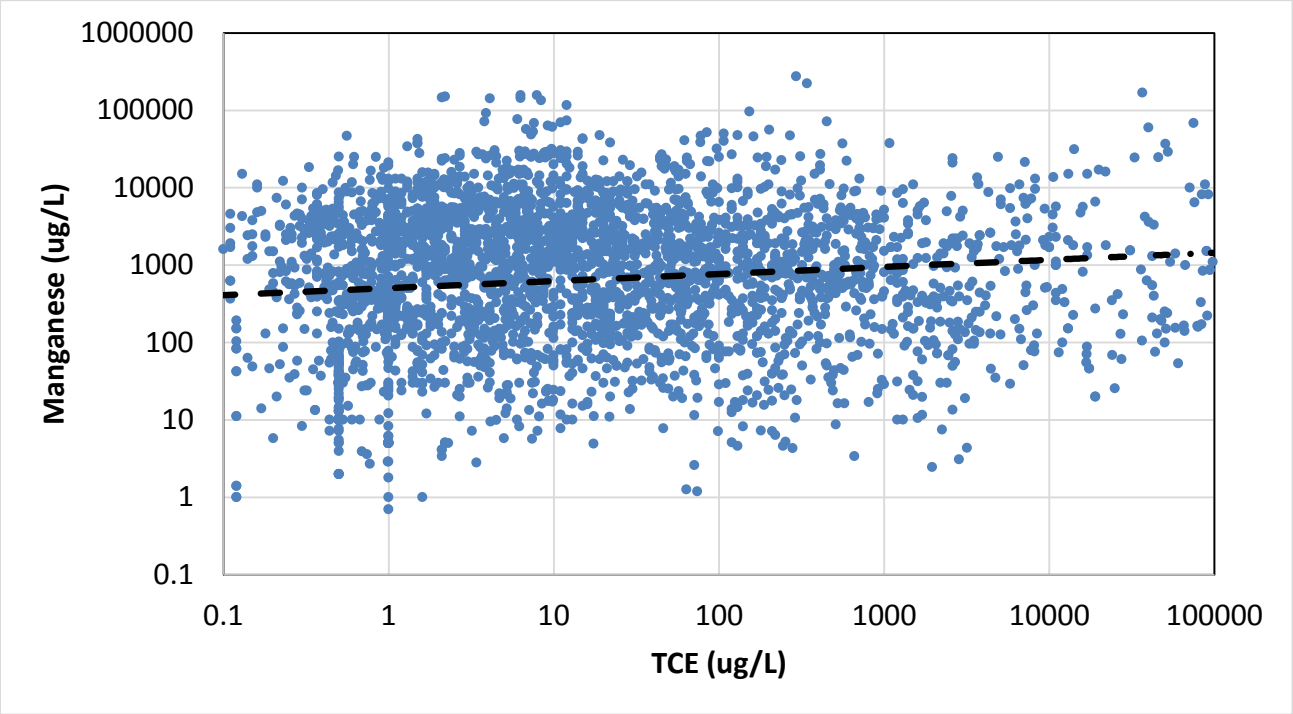


Figure 8: Relationship between vinyl chloride and manganese (as an indicator of oxidation-reduction conditions), top, and TCE and manganese, bottom, in samples with co-detections of all three species across multiple sites throughout California.

Patterns Among Plume Populations

Among the 1,183 sites, a variety of monitoring well network summary CVOC statistics can be extracted. The distribution of maximum well-to-well distances, a proxy for plume length, across the site population is summarized on Figure 9. Similarly, the distribution of areas associated with the convex hulls of the site monitoring well networks, or plume extent, is shown on Figure 10. Both distributions exhibit approximate lognormal shapes. The number of historic monitoring wells per unit area of each site is summarized on Figure 11. These results indicate a median value close to 10 wells per hectare, although a small subset of sites are characterized by higher densities (e.g., 20-40 wells/hectare). These sites may include monitoring wells that sample across multiple aquifers, or may feature high-concentration source areas that have been subject to additional characterization or remediation.

Assuming a classical idealized ellipsoidal shape for CVOC plumes, where the plume area is given by the product of the semi-major and semi-minor axes, and π, a distribution of lumped plume "widths" (per site) can be inferred from the corresponding lengths and areas. The relationship between inferred plume widths and lengths is shown on Figure 12. Regression of these data indicates that, on average, plume width is on the order of one-fifth of plume length, albeit with much scatter in the relationship. Commonly, the major axes of plumes tend to align with the prevailing groundwater flow direction, as would be expected (Figure 13).

Some trends in the overall site data pertaining to relationships between various CVOCs are also apparent. For example, the distributions of maximum historic site concentrations of CVOCs belonging to the chloroethene reductive dehalogenation sequence TCE → *cis*- and *trans*-1,2-DCE → vinyl chloride is shown on Figure 14. Given the prevalence of TCE in the individual sample event dataset, frequent detections of the respective reductive dehalogenation daughter products suggests that dechlorination reactions are widespread, either stemming from natural conditions or as a product of engineered bioremediation. *cis*-1,2-DCE is the most common member of the sequence, and occurs at the highest concentrations, followed by vinyl chloride and then *trans*-1,2-DCE. Associations between various CVOCs are also apparent in overall site plume metrics. Maximum concentrations of CVOC that are associated through degradation reactions, such as the abiotic conversion of 1,1,1-TCA to 1,1-DCE (Figure 15), appear as correlations in the data across sites where both CVOCs are found, whereas those that are presumably unrelated by either degradation reaction or use exhibit little or no obvious correlation (Figure 16). Similarly, the mass per unit aquifer depth of 1,4-dioxane is correlated with both 1,1,1-TCA (Figure 17) and 1,1-DCE (Figure 18), but less so with respect to PCE (Figure 19), with which it has no known use-relationship.

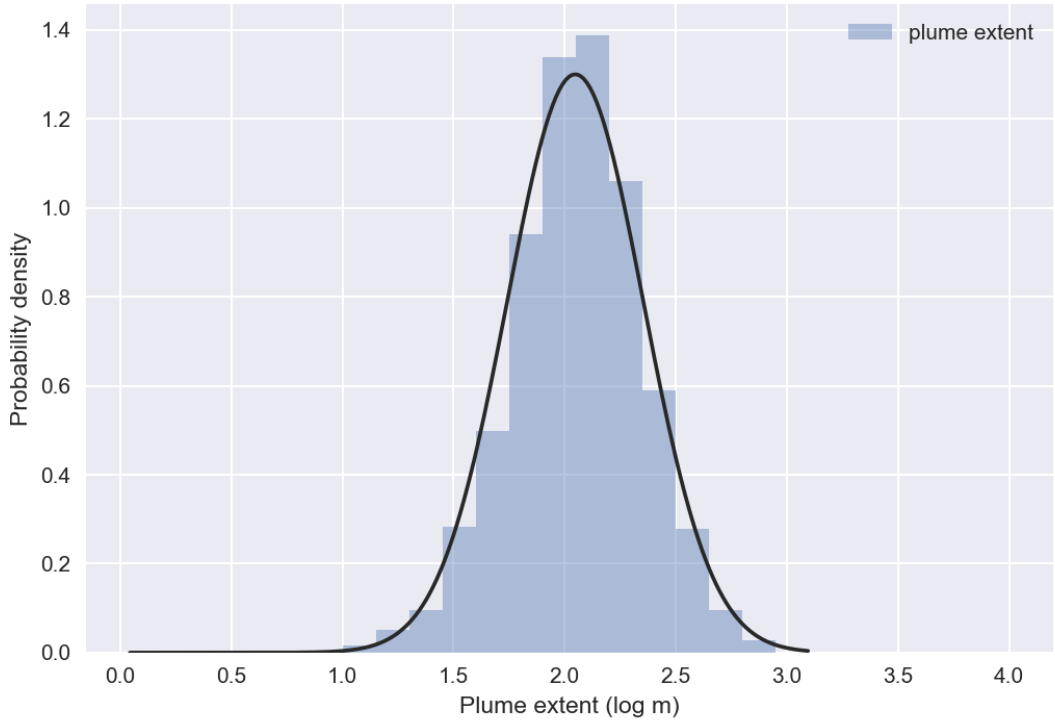


Figure 9: Distribution of maximum extent of respective monitoring well networks among the 1,183 sites identified by cluster analysis.

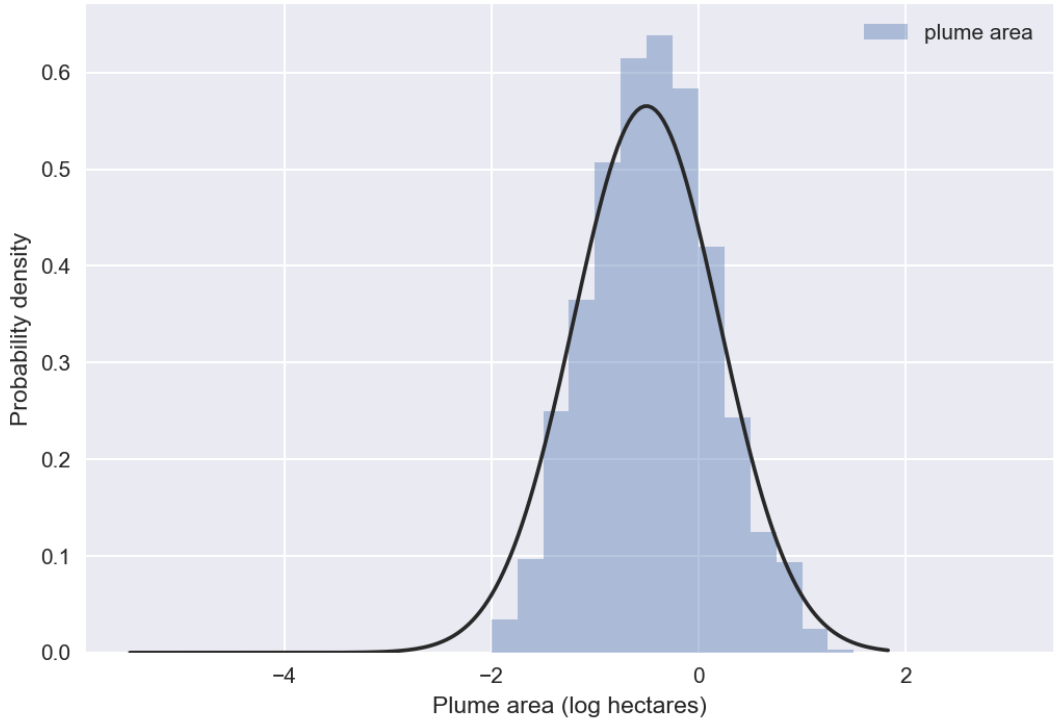


Figure 10: Distribution of integrated areal footprints of respective monitoring well networks, computed via Delaunay triangulation, across the site population identified by cluster analysis.

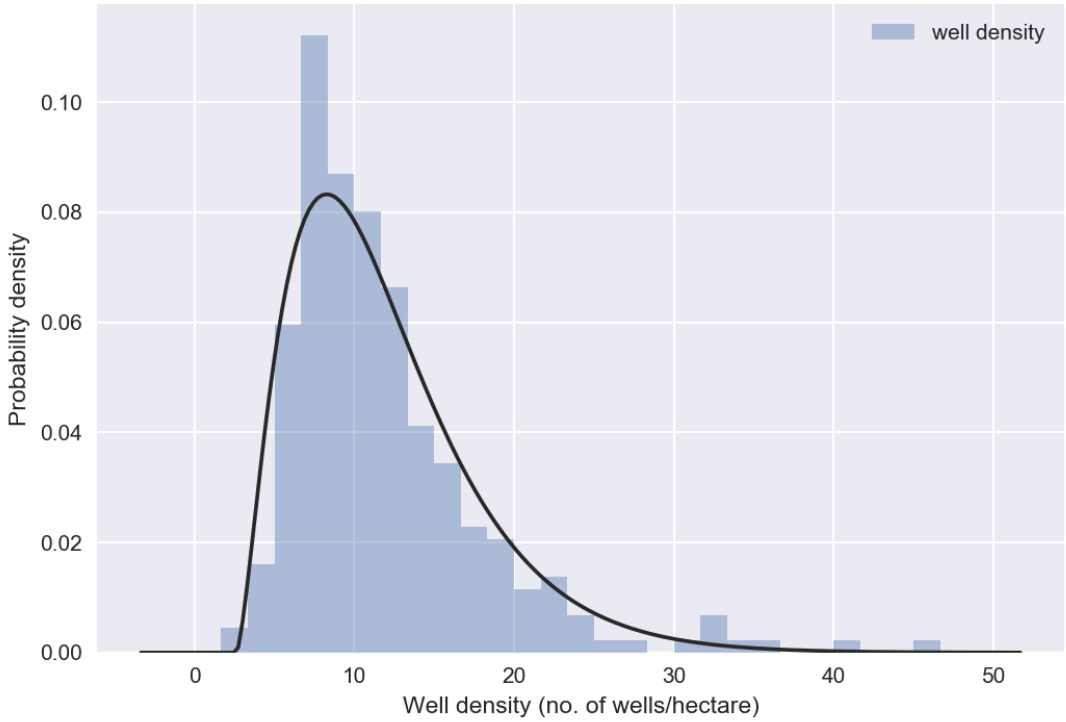


Figure 11: Computed monitor well density across the population of sites (for sites with areas greater than 1 hectare).

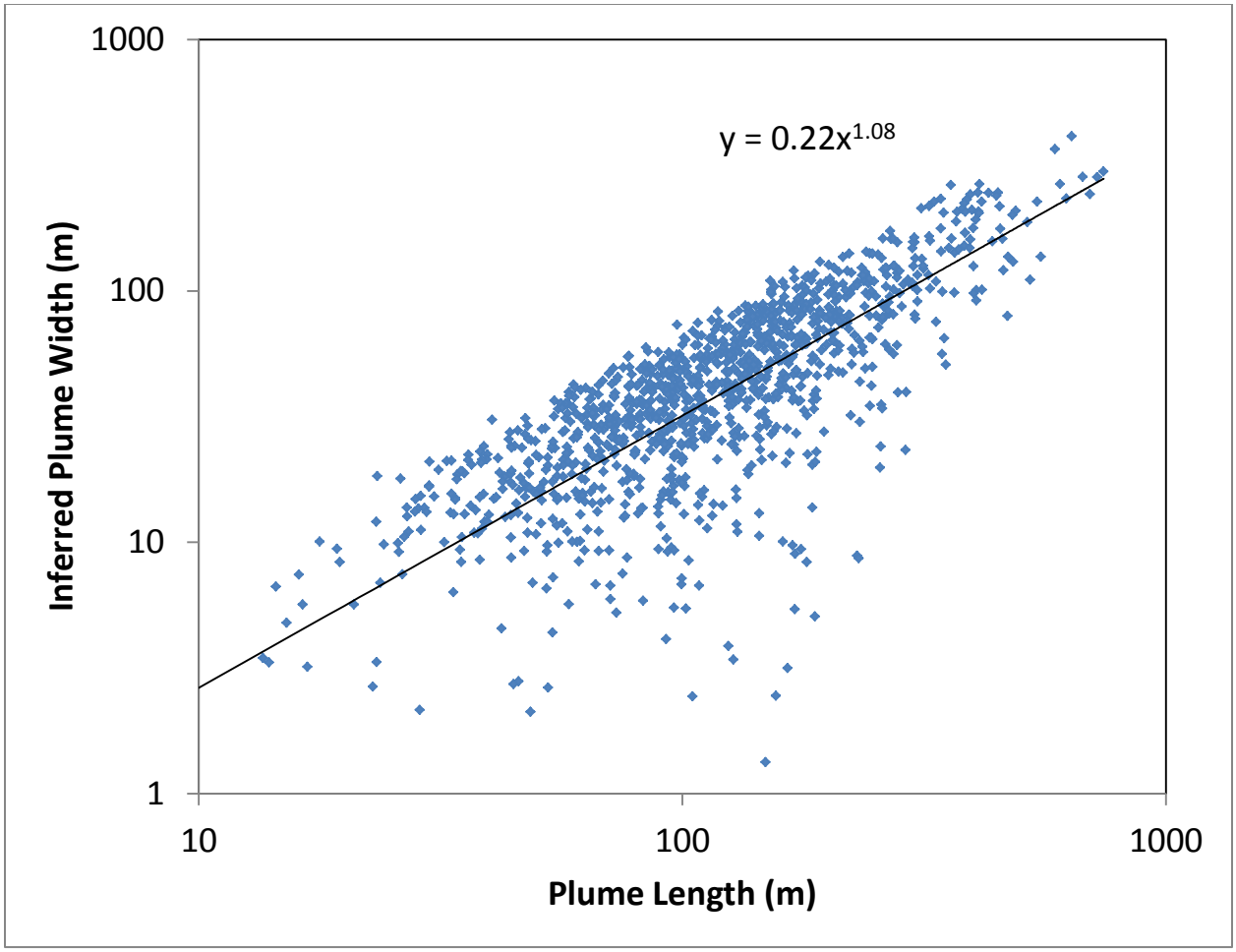


Figure 12. Inferred plume width, based on idealized ellipse shape, versus corresponding plume length.

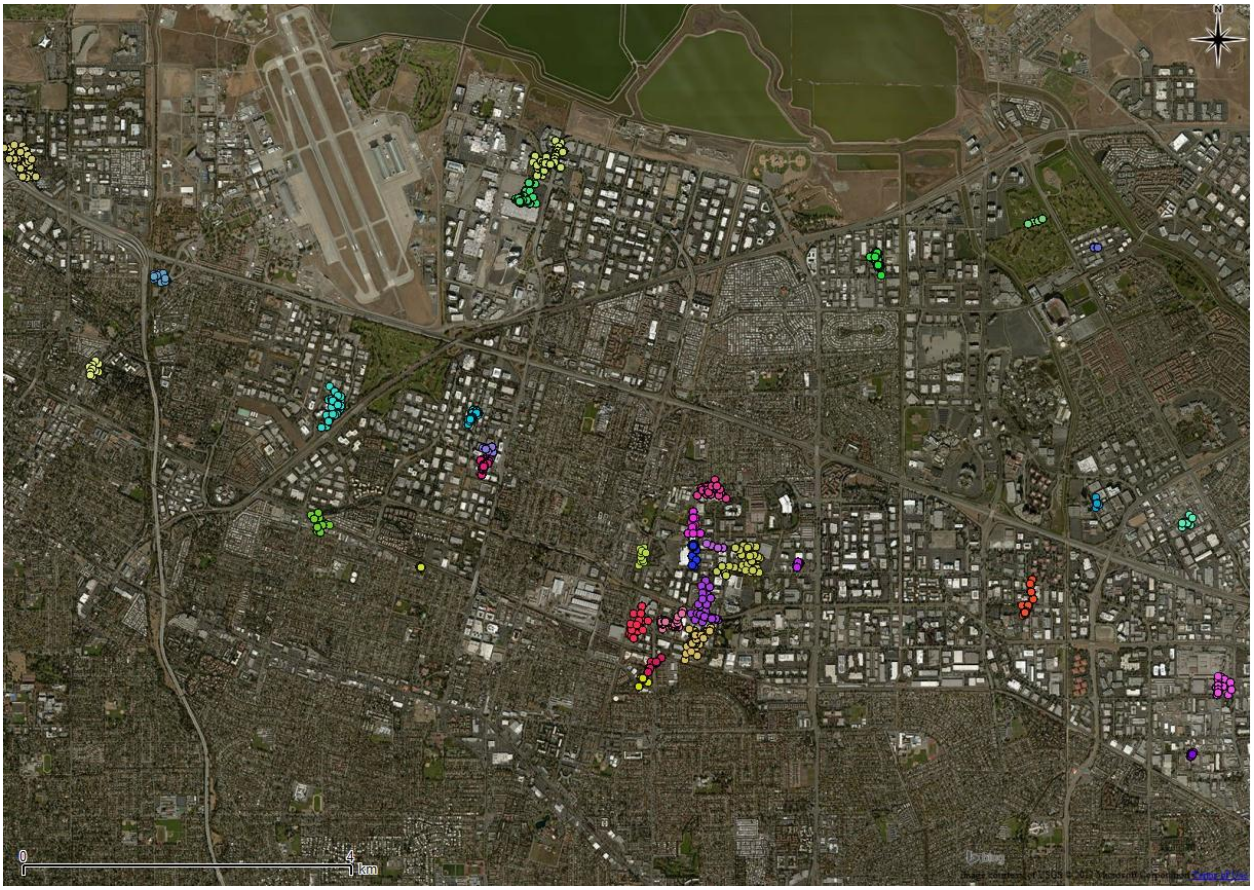


Figure 13. Configurations of multiple individual plumes identified by the cluster analysis – indicated by different color symbols marking well locations – near the southern end of San Francisco Bay. Groundwater flow direction is generally northerly. Posited assignments of wells to plumes do not necessarily reflect any regulatory designation or remediation plan but are instead attributable purely to geographical placement.

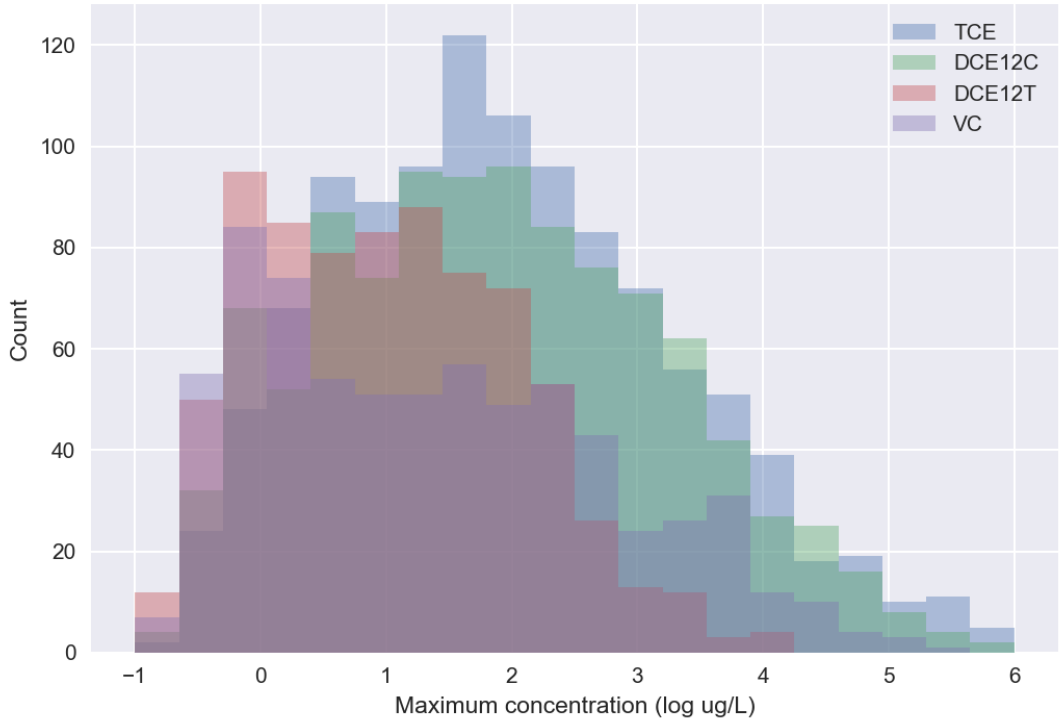


Figure 14: Distributions of maximum historic concentrations, among the population of sites, of the chloroethenes comprising the commonly recognized reductive dehalogenation sequence.

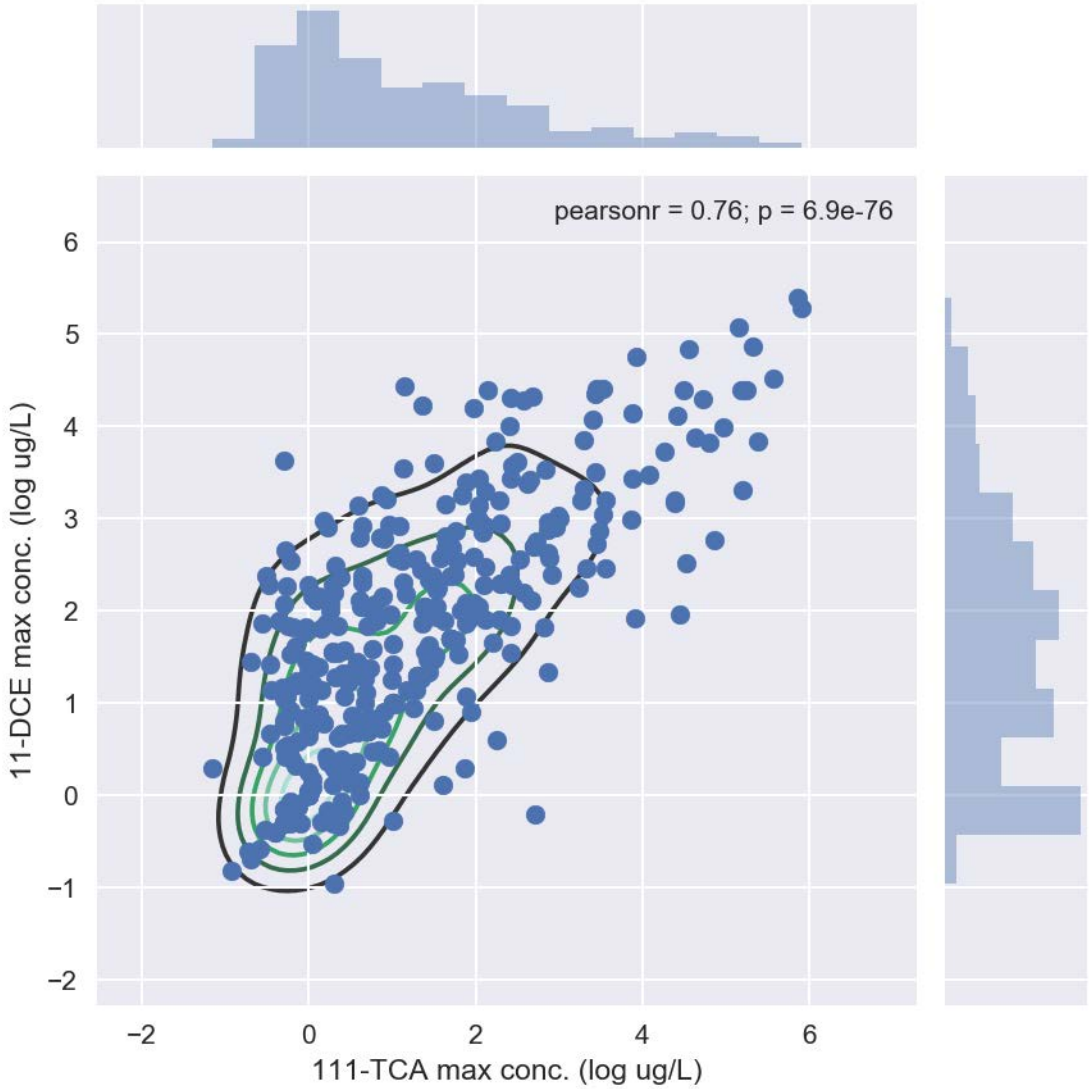


Figure 15: Comparison between maximum concentrations of 1,1-DCE and 1,1,1-TCA among the population of sites, with the former often presumably existing as an abiotic degradation product of the latter.

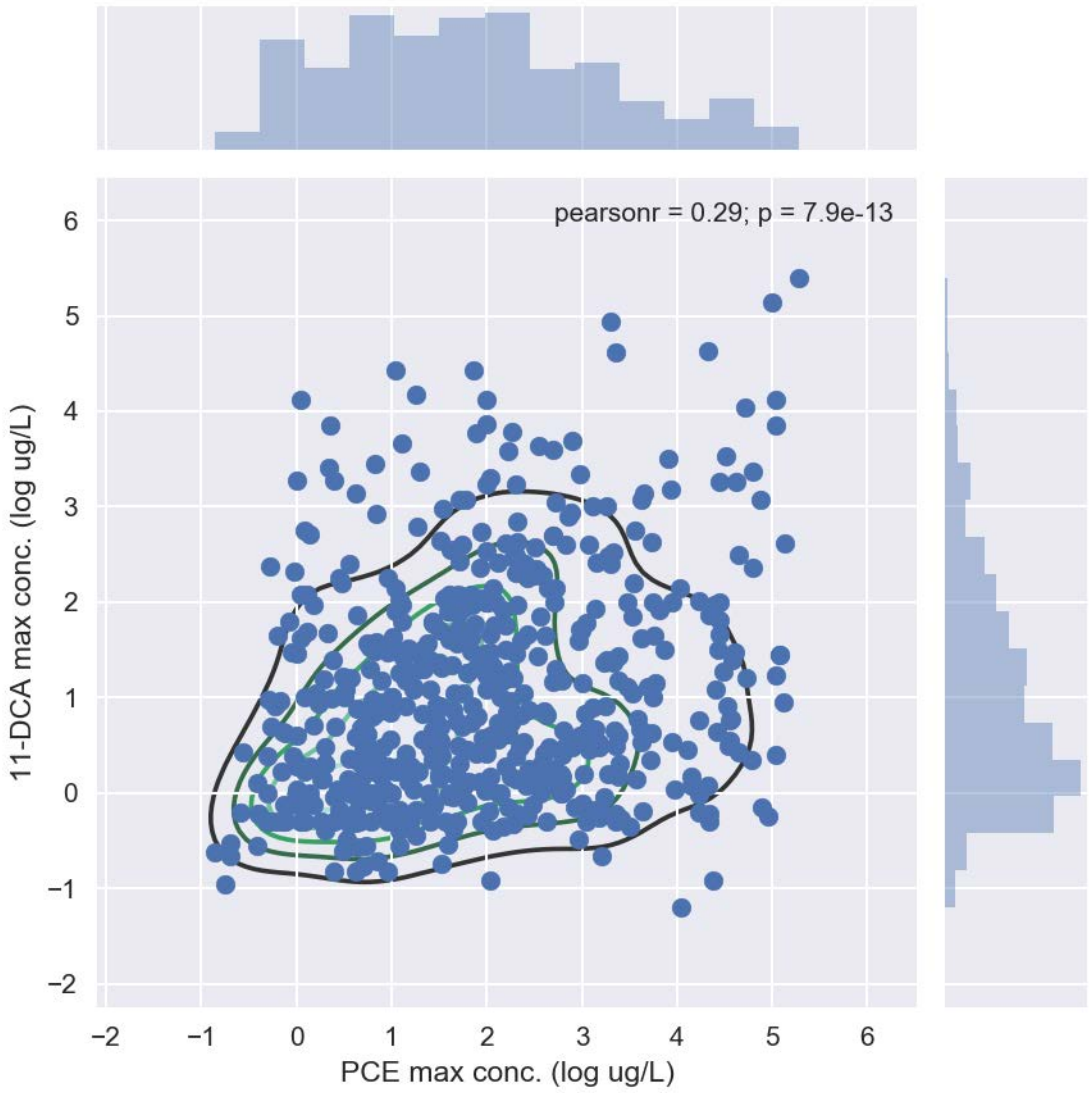


Figure 16: Comparison between maximum concentrations of 1,1-DCA and PCE among the population of sites; reference example where no particular use- or degradation reaction relationship is known to exist.

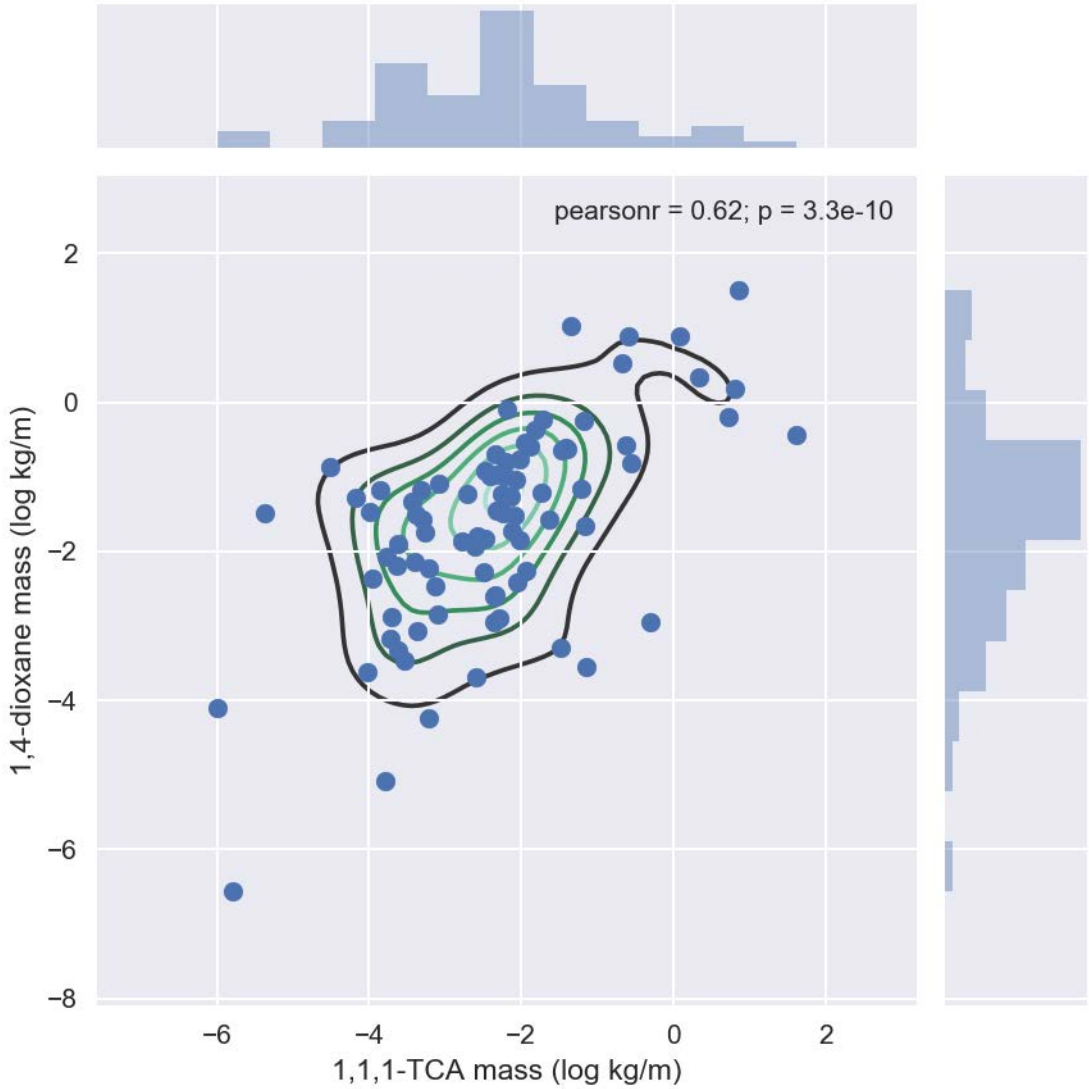


Figure 17: Comparison between mass/aquifer depth of 1,4-dioxane and 1,1,1-TCA across the population of sites.

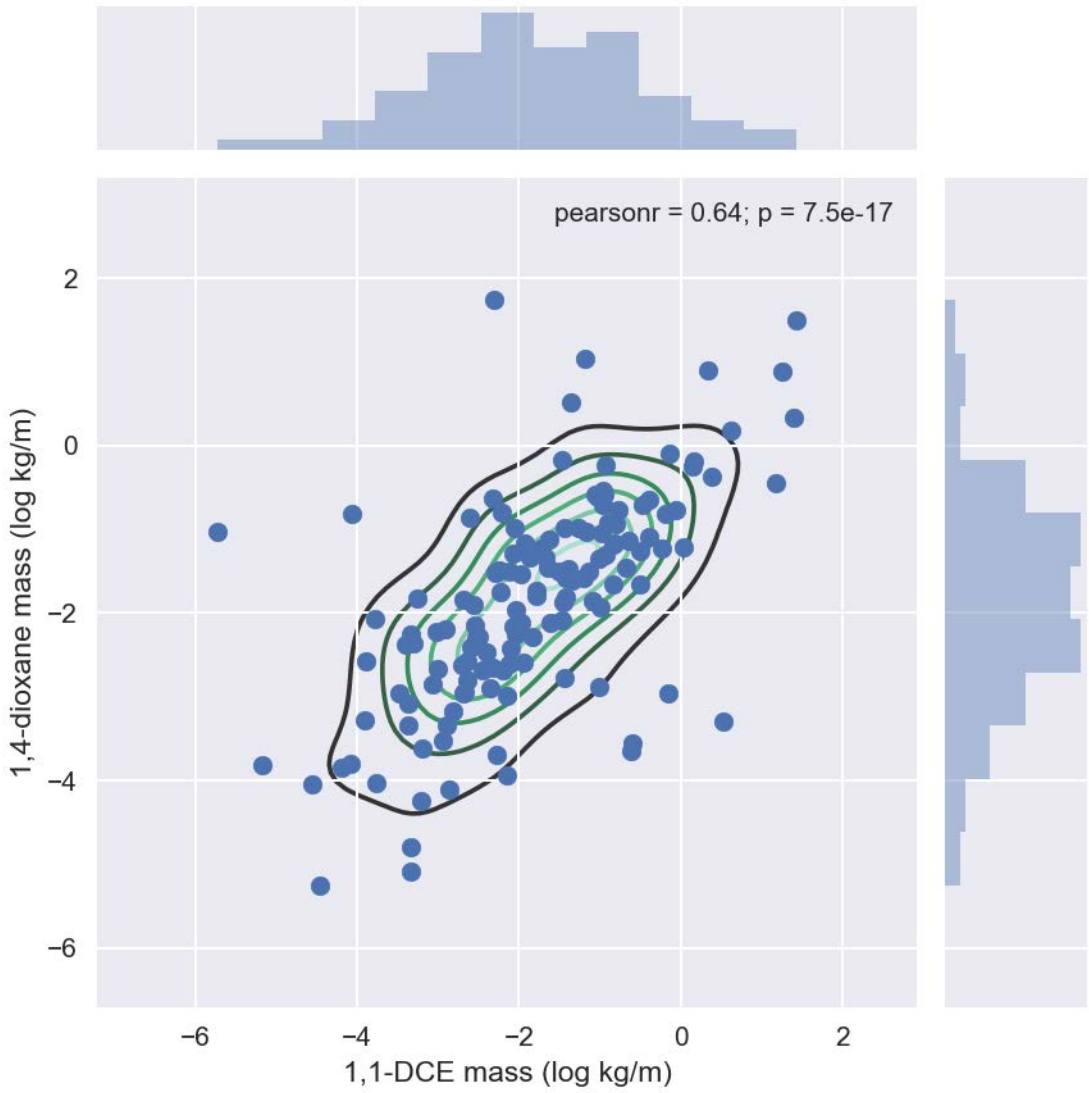


Figure 18: Comparison between mass/aquifer depth of 1,4-dioxane and 1,1-DCE across the population of sites.

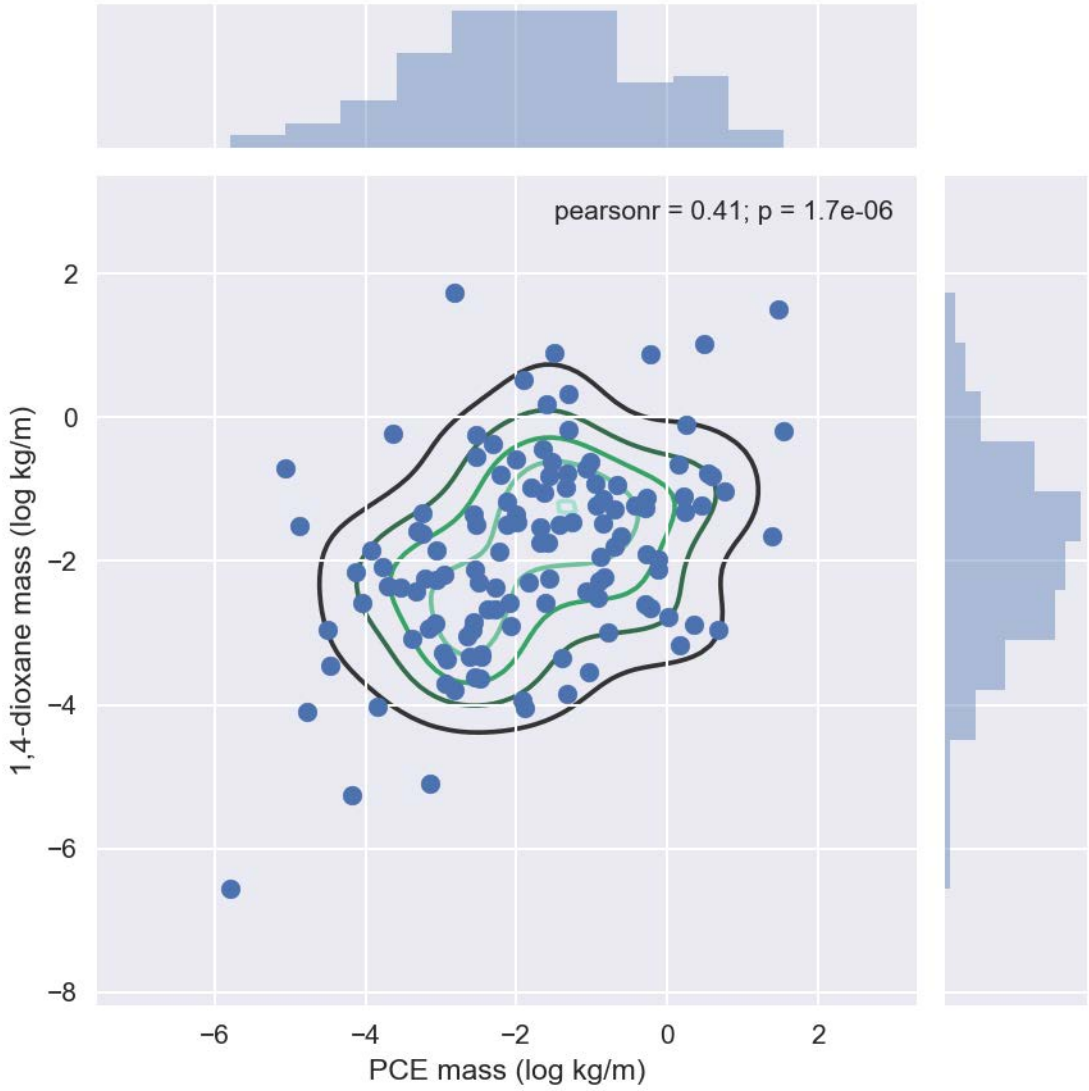


Figure 19: Comparison between mass/aquifer depth of 1,4-dioxane and tetrachloroethene across the population of sites.

Discussion

Application of accessible machine learning tools to public groundwater quality databases permits complex plume meta-analysis with relative ease compared to more labor-intensive past data gathering efforts. For databases consisting of hundreds of thousands of individual groundwater samples, tens of thousands of wells, and hundreds to thousands of sites/plumes, a key requirement is a capability to parse wells into sites based on survey coordinates. DBSCAN cluster analysis is ideal, provided that maximum search distance is optimized. This approach has been demonstrated for California's GAMA database, which lists CVOC concentration histories and survey locations for environmental monitoring wells distributed throughout the state.

Subsequent site-specific analyses can take advantage of methods such as Delaunay triangulation to automate the summary of putative site-specific data. Example features in both the raw, individual groundwater sample data from the GAMA database and the parsed site dataset have been extracted from the data that are consistent with expected trends in the data, including associations between 1,1-DCE and 1,1,1-TCA, 1,4-dioxane and 1,1,1-TCA and its daughter products, and probability distributions describing monitoring well network spacing.

The python code used for this assessment, and an extraction of CVOC data from the GAMA database across all 58 counties, can be found at, https://github.com/NumericalEnvironmental/VOC_Plume_Meta-analysis_with_Python.

Limitations to this machine learning-based approach to lumping environmental data from such large datasets are numerous and should be recognized. Clearly, the assignment of individual monitoring wells to clusters is not equivalent to properly matching every well to the correct corresponding site, where site-specific historical operations information and hydrogeologic data would more precisely characterize plume behavior at the local scale. This type of mis-assignment error would be more common in urban or industrial areas where sites may be closely juxtaposed, the results shown on Figures 1 and 2 notwithstanding. However, other sources of error also exist, including, at a minimum:

- Biases and/or extensive information omission inherent in ignoring non-detections;
- Temporal changes in both plume characterization – in particular the number of wells sampled – as well as plume movement over time;
- Inability to identify impacts and timing of various remediation approaches at individual sites;
- Historic sampling biases, as CVOC data, provided as EDF, are generally from 2000 or later in the GAMA database; and,
- As noted, three-dimensional data distribution information, which may be critically important for some sites, is not listed in GAMA and so cannot be represented.

References

Adamson, D.T., R. H. Anderson, S. Mahendra, and C.J. Newell, 2015. Evidence of 1,4-dioxane attenuation at groundwater sites contaminated with chlorinated solvents and 1,4-dioxane, Environmental Science and Technology, 49 (11), pp 6510–6518.

Adamson, D.T., S. Mahendra, K.L. Walker, S.R. Rauch, S. Sengupta, and C.J. Newell, 2014. A multisite survey to identify the scale of the 1,4-dioxane problem at contaminated groundwater sites, Environmental Science and Technology Letters, 2014, pp 254–258.

Connor, J.A., R. Kamath, K.L. Walker, and T.E. McHugh, 2015. Review of quantitative surveys of the length and stability of MTBE, TBA, and benzene plumes in groundwater at UST sites, Groundwater, 53(2), pp. 195-206.

Ester, M., H.P. Kriegel, J. Sander, and X. Xu, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, pp.226–231.

Faybishenko, B. and T.C. Hazen, 2009. Multiple factor analysis and k-means clustering-based classification of the DOE Groundwater Contaminant Database (abstract), American Geophysical Union, Fall Meeting.

Lee, J., J. Im, U. Kim, and F. Löffler, 2016. A data mining approach to predict in situ detoxification potential of chlorinated ethenes, Environmental Science and Technology, 50 (10), pp 5181–5188.

Mace, R.E., R.S. Fisher, D.M. Welch, and S.P. Parra, 1997. Extent, mass, and duration of hydrocarbon plumes from leaking petroleum storage tank Sites in Texas, Bureau of Economic Geology, University of Texas at Austin, Geologic Circular 97-1.

Mace, R.E., and W.J. Choi. 1998. The size and behavior of MTBE plumes in Texas, in Proceedings of the Petroleum Hydrocarbons and Organic Chemicals in Ground Water, 1–11, November 11–13, Houston, Texas.

McNab, W.W., D.W. Rice, and C. Tuckfield, 2000. Evaluating chlorinated hydrocarbon plume behavior using historical case population analyses, Bioremediation Journal, 4(4), pp. 311–335.

McNab, W.W., 2001. Forensic analysis of chlorinated hydrocarbon plumes in groundwater: A multi-site perspective, Environmental Forensics, 2(4), pp. 313–320.

Newell, C.J., I. Cowie, T.M. McGuire, and W. McNab, 2006. Multi-year temporal changes in chlorinated solvent concentrations at 23 MNA sites, Journal of Environmental Engineering, 132(6), pp. 653–663.

Rice, D.W., R. D. Grose, J.C. Michaelsen, B.P. Dooher, D.H. MacQueen, S.J. Cullen, W.E. Kastenberg, L.G. Everett, and M.A. Marino, 1995. California Leaking Underground Fuel Tank (LUFT) Historical Case Analyses, Lawrence Livermore National Laboratory, UCRL-AR-122207, 65 p.