APPENDIX A.3 — WELL REDUNDANCY/SUFFICIENCY ANALYSIS: DELAUNAY METHOD

Authors: Ling, M. and Rifai, H. S., University of Houston.

This appendix introduces the approach used in MAROS for well sufficiency analysis, the Delaunay method. The Delaunay method is designed to select the minimum number of sampling locations based on the spatial analysis of the relative importance of each sampling location in the monitoring network. The approach allows elimination of sampling locations that have little impact on the historical characterization of a contaminant plume. A well sufficiency analysis method (i.e., recommend new locations) based on the Delaunay method is also introduced.

Method Description

The Delaunay method is developed based on Delaunay triangulation, which is the triangulation of a point set with the property that no point in the point set falls in the interior of the circumcircle of any triangle in the triangulation. As seen in Figure A.3.1, all nodes (potential well locations) are joined by the blue lines, which form the edges of Delaunay triangles. The yellow lines form many polygons called Thiessen polygons or Voronoi diagrams, which are the dual parts of Delaunay triangles.

![Figure A.3.1 Illustration of the Delaunay Triangulation](image_url)
Delaunay triangles and Voronoi diagrams have been widely used for centuries for solving spatial distribution problems (Okabe et al. 1992, Watson 1994). In MAROS, Delaunay triangulation is first used to generate a grid for the studied site with potential sampling locations as its nodes. Then based on the formation of Delaunay triangles and Voronoi diagrams, spatial analyses are made to determine the relative importance of each sampling location. Finally, spatial-redundant locations are eliminated from the monitoring network.

To determine the relative importance of sample locations in the monitoring network, we define a *Slope Factor* (SF) for each potential location. The SF provides a measure of the importance of the information supplied by each sample location. Generally speaking, the SF is defined as the standardized difference between the concentration measured at a location and a concentration estimated from concentrations at its nearest neighbors.

The spatial distribution of groundwater quality data tends to follow lognormal distribution, but variance in the time-series data can be large due to artifacts of sampling and analysis and other issues. Using logarithmic scale of the concentrations smooths variance in the data, creating a more stationary data set. Using logarithmic transformations of the concentrations for estimating the average plume concentration were seen in some studies (Rice et al. 1995; Mace et al. 1997).

To be consistent, the SF calculation in MAROS is thus based on the logarithmic scale of the concentrations. In order to avoid performing calculations with negative values, concentration data is multiplied by a scaling factor (usually 10, 100 or 1000) before log transformation. The scaling factor is determined automatically in the software and is based on the magnitude of the minimum concentration result. In this way, all values are > 0 before log transformation. The following steps are used to calculate SF.

1) For a given node $N_0$, find its natural neighbors $N_i$, i.e., the set of nodes that are directly connected to this node by an edge of a Delaunay triangle (Figure A.3.2).

![Figure A.3.2 Illustration of Natural Neighbors](image-url)
2) The estimated logarithmic concentration, \( EC_0 \), of node \( N_0 \), is computed as the inverse-distance-weighted average of logarithmic concentrations of its natural neighbors:

\[
EC_0 = \frac{\sum_{i=1}^{n} \left( NC_i \cdot \frac{1}{d_{oi}} \right)}{\sum_{i=1}^{n} \frac{1}{d_{oi}}}
\]

where:

- \( n \) = number of natural neighbors
- \( NC_i \) = measured concentration in logarithmic scale at node \( N_i \) where \( i = 1, 2, \ldots, n \)
- \( d_{oi} \) = distance between node \( N_0 \) and its natural neighbor \( N_i \)

3) The SF is then calculated as:

\[
SF = \frac{\left| EC_0 - NC_0 \right|}{\text{Max}(EC_0, NC_0)}
\]

where:

- \( EC_0 \) = estimated logarithmic concentration at node \( N_0 \)
- \( NC_0 \) = measured concentration in logarithmic scale at node \( N_0 \)

The magnitude of SF ranges from 0 to 1 (not including 1). Value 0 means that the concentration at a location can be exactly estimated by its surrounding locations, thus, sampling at this location provides no extra information influencing understanding of the plume. A value larger than ‘0’ indicates the existence of estimation error. The larger is the estimation error, the larger the discrepancy would be between the estimated concentration and the measured concentration at a sampling location. SF values close to 1 indicate that the location provides unique information. Consequently, it is reasonable to keep sampling the location so that the plume can be better defined. In summary, the larger the SF value of a location is, the more important the location and vice versa.

One objective in spatial sampling is to accurately map a contaminant plume and track changes in this plume over time. It is clear that with more monitoring wells this objective can be achieved with a higher degree of accuracy. Unfortunately, there is always a trade-off between degree of accuracy and budget. The limitation of resources forces us to find a way to use as few monitoring wells as possible as far as certain degree of accuracy can be kept (no significant information loss).

To ensure that the elimination of sampling locations from a monitoring network will not cause significant information loss, two indicators are developed to measure the information loss. One is Average Concentration Ratio (CR) and the other is Area Ratio (AR), which are defined as:

\[
CR = \frac{C_{\text{avg,Current}}}{C_{\text{avg,Original}}} \quad AR = \frac{A_{\text{Current}}}{A_{\text{Original}}}
\]
where:

\[ C_{\text{avg,Current}} = \text{average plume concentration estimated after elimination of locations in the current step of optimization} \]
\[ C_{\text{avg,Original}} = \text{average plume concentration estimated from the potential locations (original network before elimination of any locations)} \]
\[ A_{\text{Current}} = \text{Triangulation area based on locations after elimination of locations in the current step of optimization} \]
\[ A_{\text{Original}} = \text{Triangulation area based on potential locations before any optimization (original network before elimination of any locations)} \]

The average plume concentration is taken as the area-weighted average of the average concentrations of all Delaunay triangles:

\[ C_{\text{avg}} = \frac{\sum_{i=1}^{N} TC_i \cdot TA_i}{\sum_{i=1}^{N} TA_i} \]

where:

\[ N = \text{number of all Delaunay triangles in the triangulation} \]
\[ TA_i = \text{area of each Delaunay triangle, } i = 1, 2, \ldots, N \]
\[ TC_i = \text{average concentration of each Delaunay triangle, } i = 1, 2, \ldots, N \]

\[ TC_i \] is computed as (refer to Figure A.3.3):

\[ TC_i = \frac{NC_1 \cdot A_1 + NC_2 \cdot A_2 + NC_3 \cdot A_3}{A_1 + A_2 + A_3} \]

where:

\[ NC_1 = \text{logarithmic concentration at vertex } N_1 \]
\[ NC_2 = \text{logarithmic concentration at vertex } N_2 \]
\[ NC_3 = \text{logarithmic concentration at vertex } N_3 \]
\[ A_1 = \text{Area of sub-part } A_1 \]
\[ A_2 = \text{Area of sub-part } A_2 \]
\[ A_3 = \text{Area of sub-part } A_3 \]

After elimination of "unimportant" locations (those with smallest SF values), the estimation of average plume concentration and triangulation area might be affected. By judging the values of CR and AR, information loss can be evaluated. CR and AR values close to 1 indicate that the information about the plume after elimination of locations is well kept. CR and AR values close to 0 represent a large estimation discrepancy and thus indicate greater information loss. By setting the acceptable level of information loss, we can judge when to stop eliminating locations.
Those eliminated locations are called "redundant" locations and the rest of potential locations are non-redundant ones and should be kept. An interpretation of the elimination decision process is given in Figure A.3.4.

The optimization process is iterative. It starts by eliminating the location(s) with smallest SF value(s), followed by a check of information loss. If information loss is not significant (within the acceptable range specified by the user), the process repeats until significant information loss happens.

![Figure A.3.3 Division of a Delaunay Triangle](image)

Two kinds of thresholds are defined to judge whether or not to 1) eliminate a location or 2) to terminate the optimization. The SF threshold is defined for the first purpose. For example, if the SF threshold for all nodes is 0.10, those nodes with SF values less than 0.10 are potential nodes to be eliminated. CR and AR thresholds are defined for the second purpose. For example, if CR threshold is 0.95, elimination of locations is valid if the CR value is greater than 0.95. In this case, the acceptable level of information loss is $1 - 0.95 = 0.05$, that is, 5%. If the CR value is less than 0.95, the optimization should be terminated and the locations eliminated at this step should be re-instated. Details about these thresholds will be discussed shortly.
The User has the option of choosing the threshold levels for the SF and the area and concentration ratios. More detailed descriptions of the threshold parameters are discussed below. Choice of the parameter values should reflect what the User considers to be an acceptable level of information loss. Parameters should be determined after consideration of the overall monitoring objectives for the network.

**Inside node Slope Factor**: The SF threshold for nodes (locations) located inside the triangulation domain. When SF of an inside node is less than this threshold, and if the node is *Removable*, it will be eliminated from the monitoring network. The current default value for this parameter is 0.1. *Removable* stands for the elimination property of a location. If the *Removable* property of a location is False, optimization cannot eliminate it no matter how small its SF value is. This is important if you want to keep a location (e.g., a POC well) in the monitoring network. The default values for all potential locations are True.

**Hull node Slope Factor**: The SF threshold for nodes (locations) located on the edge (convex hull) of the triangulation domain. When SF of a hull node is less than this threshold, and if the node is *Removable*, it will be eliminated from the monitoring network. The current default value for this parameter is 0.01. The threshold for hull node is usually more stringent than that of the inside node, because the elimination of a hull node may cause reduction in the triangulation area, thereby causing greater information loss (reduction in AR). For contrast, the elimination of an inside node will only affect the average concentration ratio (CR).

**Area Ratio** (AR): The ratio of triangulation area (represents the area of a contaminant plume) at current optimization step to the original triangulation area before optimization. If the AR value in an optimization step is less than the threshold, the optimization will be stopped and locations eliminated in this step will be resumed. The default value is 0.95.

**Concentration Ratio** (CR): The ratio of average concentration of a contaminant plume at current optimization step to that of the original value before optimization. If the CR value in an optimization step is less than the threshold, the optimization will be stopped and locations eliminated in this step will be resumed. The default value is 0.95.
For the setting of these parameters, the user is referred to the corresponding parts in chapter MAROS Detailed Screen Descriptions.

**Choice of Sampling Events**

The Delaunay method performs the redundancy reduction by using an algorithm that considers all or a series of sampling events, of which optimization based on a single sampling event is a special case. Since each sampling event represents only one snapshot of the contaminant plume, we need to examine all sampling events (or parts of them) to reveal the general spatial pattern of the contaminant distribution in a specific site. This general spatial pattern is the underlying assumption for the spatial analysis. In the Delaunay method, we find the general pattern by averaging across sampling events. In addition, since the spatial patterns of COCs may be different from each other, the optimization is performed based on each COC. Therefore, results are given separately in terms of each COC. Finally, we provide the all-in-one results simply by considering the most conservative result from all COCs. The major steps of this algorithm are as follows:

1) Select a series of continuous sampling events for analysis. They could be all sampling events in the monitoring history. They could also be any segment of sampling events in the monitoring history, e.g., sampling events in the past five years.

2) Calculate SF values of potential locations for all sampling events selected by the users, and for each COC.

3) Average SF values of potential locations across the selected sampling events for each COC, weighted by the number of locations contained in each sampling event. The results are lumped SF values of potential locations for each COC.

4) Eliminate one location at a step from each COC starting from the location with smallest lumped SF value. Calculate CR and AR ratios for each sampling event and then average them across sampling events to provide sampling-events-averaged CR and AR values. Compare sampling-events-averaged CR or AR values to thresholds and if there is no significant information loss, repeat this step with the next available location.

5) Provide the COC-categorized results after eliminating all redundant locations from each COC. In this step, elimination of a location in a COC means to stop sampling for that COC at that well in the next round of sampling.

6) Provide the all-in-one results by eliminating only those locations that are eliminated from all COCs. Here elimination of a location is equivalent to abandoning it, i.e., to stop service of a well since no COC needs to be sampled at this well any more.

The user can also choose to analyze only one sampling event, e.g., the latest sampling event. In this case, the step of averaging across sampling events is skipped. Figure A.3.5 shows the detailed procedures of optimization in this simplified process.

In MAROS, two modules are developed based on the Delaunay Method. One is the **Access Module** starting with screen Well Redundancy Analysis: Delaunay Method, which is introduced in the chapter MAROS Detailed Screen Descriptions. The other one is the **Excel Module – xlsDelaunay2K**, which is a stand alone Microsoft Excel Worksheet, also discussed in chapter MAROS Detailed Screen Descriptions. The Access Module is designed to deal with multi-sampling-events analyses recognizing that a general spatial pattern may lie beneath what are revealed by each single sampling event. It can also be used to analyze a single sampling event, a special case of the multi-sampling-events analyses. The Excel Module is designed for one-sampling-event analyses, which
provides the user with graphical interface and convenient controls to the optimization process, making the process of the Delaunay method better understood.

Figure A.3.5 Steps in Sampling Location Optimization for One Sampling Event
Well Sufficiency Analysis

Augmentation of a monitoring network is needed when the existing network cannot achieve certain monitoring goals. Augmentation in this document means the addition of sampling locations and/or more frequent sampling. In this section, a method for determining new sampling locations is introduced, which is intended to enhance the spatial plume characterization. This method utilizes the SF values obtained from the previous analysis to assess the concentration estimation error or uncertainty in areas within the network. Among these potential areas, those with a high estimation error may be designated as regions for new sampling locations or increased monitoring intensity.

Conceptually, the method is to overlay a grid onto the study area and interpolate the SF values at existing sampling locations to grid cells that do not contain sampling locations. These grid cells serve as potential areas for new sampling locations. Those areas with a high estimated SF value (i.e., high estimation error) are therefore candidate regions for new sampling locations. This approach is further simplified in MAROS in order to adapt to the visualization limitations of Microsoft Access and Excel. In the simplified approach, each Delaunay triangle in the triangulated monitoring network is used as a potential area for new sampling locations (Figure A.3.7).

The SF value at a Delaunay triangle is estimated as follows. Consider a Delaunay triangle with vertices $N_1$, $N_2$, and $N_3$ (Figure A.3.6). Assume $A_1$, $A_2$, and $A_3$ are sub-parts of the triangle divided based on the centroid of the triangle. The average SF value for this triangle is estimated as:

$$SF_{avg} = \frac{SF_1 \cdot A_1 + SF_2 \cdot A_2 + SF_3 \cdot A_3}{A_1 + A_2 + A_3}$$

where:

- $SF_1$ = the sampling-events-averaged SF value at vertex $N_1$
- $SF_2$ = the sampling-events-averaged SF value at vertex $N_2$
- $SF_3$ = the sampling-events-averaged SF value at vertex $N_3$

![Figure A.3.6 Division of a Delaunay triangle for estimating its average SF value.](image)
The estimated SF values at these potential areas reflect the concentration estimation error at these regions for the time period specified by the sampling events. For example, a value of 0.9 indicates the ratio of the estimated to measured concentration is 1:10 or 10:1, a large discrepancy. A value of 0.5 indicates the ratio of the estimated to measured concentration is only 1:2 or 2:1, a relatively small estimation error.

In MAROS, a Microsoft Excel module, *xlsNewLocation*, is developed to implement the method. To help visualize the analysis results in *xlsNewLocation*, potential areas (the triangles) for new sampling locations are marked by blue dash lines. A colored label is placed around the center of each triangle to indicate the estimated SF level at a potential area. The estimated SF values are classified into four levels: *S—Small* (<0.3), *M—Moderate* (0.3~0.6), *L—Large* (0.6~0.9), and *E—Extremely large* (>0.9). Those potential areas with the estimated SF value at the Extremely large or Large level are candidate regions for new sampling locations. New sampling locations can then be placed inside these regions, e.g., at the centroid of a triangle region. Refer to the MAROS Detailed Screens Description chapter for details on the usage of *xlsNewLocation*.

It is emphasized that recommendations from the well sufficiency analysis are derived solely from the spatial configuration of the monitoring network and the spatial pattern of the contaminant plume. No hydrogeologic conditions are considered in the analysis. Therefore, professional judgement and regulatory considerations must be used to decide whether an area for new sampling locations recommended using the above method is appropriate.

![Figure A.3.7 Illustration of the potential areas for new sampling locations.](image-url)

**Other Considerations**

One thing to keep in mind is that if the coordinates of a sampling location are not available, this location will be excluded and will not be shown in the analysis. The potential locations for
analysis are only those with coordinates from the raw set of locations in the original database (ERPIMS or others). The minimum number of wells valid for analysis is 6. If there are less than 6 sample locations with detected concentrations, the Delaunay method cannot be applied. (Note: for datasets with less than 6 locations, the software will not function at the Moment Analysis step).

Also, before applying the Delaunay method for spatial redundancy analysis, it is important to select the appropriate set of wells for analysis, i.e., only the wells that contribute to the spatial delineation of the plume. For example, if wells are far from the plume and contribute little or nothing to the delineation of the plume (e.g., some sentry wells or background wells far from the plume), they should be excluded from the analysis. One reason not to use these wells is that these wells usually are on the boundary of the triangulation and are hard to be eliminated since the Delaunay method protects boundary wells from being easily removed. The elimination status of these wells, in fact, should be determined from the regulatory standpoint. Another well type that could be excluded from analysis is one of a clustered well set because the Delaunay method is a two-dimensional method. Generally, only one well is picked from the clustered well set to represent the concentration at this point. This well can be the one that has the highest concentration or is screened in the representative aquifer interval with the geologic unit. Data from clustered wells can also be averaged to form a single sample and then used in the Delaunay method.

References

Mace, R. E. et al., 1997, Extent, Mass, and Duration of Hydrocarbon Plumes from Leaking Petroleum Storage Tank Sites in Texas, University of Texas at Austin and TNRCC.

