# Local Index Of Spatial Association Optimization For Secondary Sample Designs In Spatial Analysis And Decision Assistance

BY

BRIAN J. STAEHLIN
BA, University of Illinois at Urbana-Champaign, 2003

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Environmental and Occupational Health Sciences
in the School of Public Health of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

Michael Cailas, Chair and Advisor
John Canar, Environmental and Occupational Health Sciences
Christos Takoudis, Bioengineering and Chemical Engineering

## TABLE OF CONTENTS

**TABLE OF CONTENTS (continued)**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF FIGURES (continued)

## LIST OF FIGURES (continued)

## LIST OF ABBREVIATIONS

| | |
|---|---|
| GPS | Global Positioning System |
| LISA | Local Index of Spatial Association |
| ppm | Parts per Million |
| RfD | Reference Dose |
| RSL | Resident Screening Level |
| SADA | Spatial Analysis and Decision Assistance |
| SL | Screening Level |
| SSL | Soil Screening Level |
| USEPA | United States Environmental Protection Agency |
| USNRC | United States Nuclear Regulatory Commission |

## SUMMARY

Sampling programs must take into account balancing the costs of operation with the necessity of accurately characterizing a contaminated site. Sampling designs that call for hundreds of samples will often be pared down drastically due to cost-cutting procedures. The samples that are implemented must then be optimally placed in order to capture the necessary information at a site. This often does not occur and crucial information is lost. The objective of this study was to assess the applicability of the Local Index of Spatial Association (LISA) secondary sampling methods provided by Spatial Analysis and Decision Assistance (SADA) software using actual data from a United States Environmental Protection Agency (USEPA) Superfund site, as well as to identify and optimize the critical parameters of these LISA methods to gain a cost effective, practical, and reliable method to place secondary samples that will ensure the characterization of the spatial distribution of contamination.

The limitations of the existing LISA parameters in SADA were observed. The LISA search window greatly affects the outcome of secondary sample designs. Guidelines were developed by mimicking real world conditions and applying them to the SADA parameters and using an iterative function for the LISA search window to develop a potential site sample distribution for each LISA secondary sample design. A methodology is recommended to reduce the redundancies that occur within the site sample distribution and that subsequently occur within the final secondary site sample design. It appears that the guidelines presented in this paper could make SADA a cost effective tool for use in Phase III and Phase IV environmental site assessments, brownfield redevelopment, or other environmental risk management or site remediation situations.

# I. INTRODUCTION

The purpose of a sampling program is to produce a set of samples representative of the source under investigation. The objective of sampling for hazardous wastes is to acquire information that will assist investigators in identifying the presence of hazardous compounds and the extent to which these compounds have become integrated into the medium under investigation. This information has the potential to be used in future litigations or to assist in the development of remedial actions (USEPA, 1983).

The USEPA, in a document detailing the characterization of hazardous waste sites, defines the term "sample" as simply a representative part of an object to be analyzed. The document (USEPA, 1983) qualifies this definition further by considering several criteria:

- Representativeness—"…the sample needs to be chosen so that it possesses the same qualities or properties as the material under consideration" (p. 2).
- Sample size—too large or too small is impractical.
- Maintenance of sample integrity—the sample must retain the properties of the parent object.
- Frequency of subsamples—is the material homogenous or heterogeneous and should a composite sample be taken?

According to Gilbert and Pulsipher (2005, 27), "[r]epresentative environmental data are essential for making defensible environmental decisions." They go on to say that sampling variability is due to the inherent variability of the environmental target population over space and time, the sample design, and the number of samples.

Barth and Mason (1984, 98) outline specific objectives when sampling a hazardous site:

- To determine the levels of contaminants and their spatial and temporal distribution.
- To determine the source, transport path, or receptor for a pollutant.
- To determine the presence of known or unknown contaminants in comparison to their presence in an appropriate background area.
- To provide input into risk assessments.
- To measure the effectiveness of control actions.
- To assist in a model validation study.

A major component, and perhaps the most critical risk for assessment determination, of sampling a hazardous site is the determination of the levels of contaminants and their spatial distribution. Without a practical approach it may be difficult and cost prohibitive to optimally achieve the above objectives when sampling a site. Cox, Cox, and Ensor (1995) point out that an exhaustive sampling procedure may not be feasible due to the high cost of obtaining and analyzing samples. Sample design, in this case, entails balancing the costs of acquiring information with the costs of making mistakes due to insufficient information. One of the main obstacles to obtaining representative samples is the lack of understanding of the effects of spatial variability and the spatial distribution of contaminants.

Superfund and brownfield sites are unique in terms of the spatial distribution of contaminants present and their potential health risks, so risk assessments are conducted on a site-by-site basis. Given the uncertain nature of enforceable and unenforceable soil standards and background characterizations, an optimal methodology is necessary for a higher degree of certainty during the risk assessment and clean-up process. In addition, a methodology

representing an optimized rational sampling plan should demonstrate economic and scientific advantages (Osiecki, 2011).

Current regulations, both federal and state, related to certain sites, such as brownfields, contain inherent problems. Primarily, risk assessment measures, which would account for the spatial distribution of the contaminants, are not prominently factored into the process. Public domain software programs, such as SADA, have been shown to be useful in identifying sampling locations by taking into account information gained by previous sample studies. Spatial Analysis and Decision Assistance software is a cost effective and reliable tool for developing a comprehensive approach to developing sample designs. The spatially defined information would allow site investigators to visualize the extent of the contamination and minimize uncertainty while providing accurate results to reduce costs during data collection and remediation. Spatial Analysis and Decision Assistance is a useful tool in risk assessment and discovering the spatial distribution of contaminants. In particular, this has applicability for brownfield redevelopment and site characterization (Sambanis, 2012).

Spatial Analysis and Decision Assistance is developed by the University of Tennessee in Knoxville and is funded by the USEPA and the United States Nuclear Regulatory Commission (USNRC). Spatial Analysis and Decision Assistance is a free software program that incorporates tools from environmental assessments fields, such as integrated modules for visualization, geospatial analysis, statistical analysis, human health risk assessment, ecological risk assessment, cost/benefit analysis, sampling design, and decision analysis, in order to effectively characterize a contaminated site, assess risk, determine the location of future samples, and design remedial action. Spatial Analysis and Decision Assistance provides a number of useful applications, one of which is secondary sampling design. Secondary sampling designs are often applied after some

data or other information was obtained. The general objective is to further refine the model or the decision in a very specific way. Secondary designs can either be point (sample) or model (geospatial) based (SADA, 2008). Spatial Analysis and Decision Assistance offers eight secondary sampling methods:

- Judgmental Design—can be classified as either initial or secondary and relies completely on the user to place samples based on professional judgment.
- Threshold Radial (also known as Adaptive Cluster Sampling)—places samples in a radial pattern around data points that exceed a decision threshold. The user has control over the pattern of the surrounding new sample points.
- Adaptive Fill Design—places samples in the largest spatial gaps among data points.
- Ripley's K—is based on Ripley's K map. The Ripley's K statistic is a measure of neighborhood sampling density. The Ripley's K design locates samples in those areas with the lowest sampling density.
- Moran's I—places samples in areas of high local sample variance as defined by Moran's I map. The idea is to collect more data in those locations where greater heterogeneity (i.e., uncertainty or variability) exists.
- Geary's C—places samples in areas with greater (in magnitude) negative correlation among samples found in the search neighborhood. Similar to Moran's I; the idea is to collect more data in those locations where greater heterogeneity exists. The difference between this approach and Moran's I is that heterogeneity (and/or uncertainty) is measured not by local variance but local correlation.
- High Value—places samples at nodes with the highest modeled values.

- Area of Concern (AOC)—places samples along the boundary line in the AOC result. Nodes that have a value closest to the decision criteria are the targets of the design. They are selected in order to more readily distinguish between contaminated and uncontaminated zones (SADA, 2008).

Spatial Analysis and Decision Assistance offers three LISA secondary sample designs, which are Ripley's K, Moran's I, and Geary's C. The interest in the LISA designs stems from their ability to give an indication of the extent of spatial clustering, or identifying what are known as hot spots.

## II. A REVIEW OF SITE SAMPLING THEORY

Sampling methodologies espoused by entities such as USEPA are generally separated into two categories: Classical Sampling Theory and Geostatistical Theory. What follows is a review of each theory as well as the general strengths and weaknesses of each. In addition, both theories are directly compared to each other and their differences outlined. Finally, brief consideration is given to designs that incorporate elements from both theories.

### A. Classical Site Sampling Theory

Sampling designs such as random sampling, systematic sampling, and stratified sampling—often employed by USEPA at Superfund sites—are based on probability sampling theory and have the following mathematical properties in common:

- A set of distinct samples can be defined, $S_1$, $S_2$, …, $S_v$, which the procedures are capable of selecting if applied to a specific population. We can say precisely what sampling units belong to $S_1$, to $S_2$, and so forth (Cochran, 1977).
- Each possible sample $S_i$ is assigned a known probability of selection $\pi_i$ (Cochran, 1977).
- One of the $S_i$ is selected by a random process in which each $S_i$ receives its appropriate probability $\pi_i$ of being selected. As an example, if we had three samples we might assign equal probabilities to them. The draw can then be made by choosing a random number between 1 and 3. If this number is j, $S_j$ is the sample that is taken (Cochran, 1977).

- The method for computing the estimate from a sample must be stated and must lead to a unique estimate for the specified sample. As an example, the estimate could be the average of the measurements on the individual units in the sample (Cochran, 1977).

Probability sampling refers to methods that satisfy these properties. The frequency distribution of the estimates these methods generate can be calculated if the sampling procedures are repeatedly applied to the same population. We then know how frequently any particular sample $S_i$ will be selected, and we know how to calculate the estimate from the data in $S_i$ (Cochran, 1977).

This theory assumes that sample estimates are approximately normally distributed. With a normally distributed estimate, the whole shape of the frequency distribution is known if the mean and standard deviation, or variance, is known (Cochran, 1977). A problem occurs with this assumption, however, when applied to contaminated sites since environmental contaminants tend to be lognormally distributed, or, highly positively skewed to the right.

1. **Simple random sampling**

According to Cochran (1997, 18), "[s]imple random sampling is a method of selecting n units out of the N such that every one of the ${}_NC_n$ distinct samples has an equal chance of being drawn," where N refers to the population, C refers to the value of the individual, and n refers to the sample number. The units of the population are numbered from 1 to N, and a series of random numbers between 1 and N is then drawn, either by a random number table or a computer program that produces one. Typically, sampling without replacement is the method used, which means that at any draw the process used must give an equal chance of selection to

any number in the population not already drawn, and numbers already drawn are removed from the population for all subsequent draws.

The USEPA states that simple random sampling is generally employed when little information exists concerning the material or location. It is effectively employed when the population of available sampling locations is large enough to lend statistical validity to the random selection process (USEPA, 1983). Elsewhere, USEPA states that randomization is necessary to make probability or confidence statements about the results of the sampling. Judgment sampling[1] has no randomization component, but may be justified for preliminary assessment and site investigation stages if the sampler has substantial knowledge of the sources and history of contamination (USEPA, 1989a).

Simple random sampling is considered most useful when the population of interest is relatively homogenous and no major patterns of contamination or hot spots[2] are expected. Many hazardous waste sites, however, are likely to contain one or more hot spots. To combat this, USEPA suggests using adaptive cluster sampling. This entails using simple random sampling to take n samples, and additional samples are taken at nearby locations where measurements exceed a particular threshold value. Additional sampling is driven by the results of the initial random sample. Adaptive cluster sampling is useful for delineating the boundaries of hot spots. Simple random sampling, due to its non-symmetric pattern, may miss pockets of higher concentration (USEPA, 2002).

---

[1] United States Environmental Protection Agency defines judgment sampling as a sample of data selected according to non-probabilistic methods (USEPA, 1989a).

[2] Hot spots are localized circular or elliptical areas with concentrations exceeding the cleanup standard. These areas are either a volume defined by the projection of the surface area through the soil zone that will be sampled or a discrete horizon within the soil zone that will be sampled (USEPA, 1989a).

Gilbert and Pulsipher (2005) suggest that in order to find hot spots efficiently, simple random sampling should be avoided and a systematic sampling design adopted. They also identify adaptive cluster sampling as a viable tool to delineate hot spots.

Cox et al. (1995) elaborate on hot spots:

> Hot spots are often selected using expert or prior knowledge, such as knowledge of sources of the contamination or topography or by visual inspection. This may be augmented by random grid sampling. A pattern for the contamination (e.g., elliptical) may be assumed, as well as the relative size of hot spots to grids. To the extent that true hot spots are located, hot spot sampling addresses the problem of area heterogeneity. Hot spot sampling is simple to execute, often yields a large number of samples, and is supported by well-documented procedures (Gilbert, 1987). However, it can be costly, and, for Superfund applications, results must be reinterpreted in terms of average contamination. (21)

A drawback, however, to adaptive cluster sampling is that as additional rounds of sampling and analysis accumulate to detect the shape of the hot spot, so too will the time and costs associated with the multiple phases. The process of sampling, testing, quality control, resampling, and testing could take considerable time. Quick and inexpensive field measurement capabilities must be available to deter total sampling costs from growing too large. In addition, the rule is that the process stops when no more units are found with the characteristic of interest. Thus, the final overall sample size is of unknown quantity, and the total cost is also an unknown quantity (USEPA, 2002). Cox et al. (1995) suggest that current theory may need to be extended (such as adaptive sampling combined with line-transect sampling) to further the use of adaptive sampling in the spatial context.

Despite this, USEPA acknowledges that multiple phase sampling can be effective. The first, or preliminary, phase can be designed to develop estimates of the variability found in the soil/waste combination, and to work out the necessary sampling protocols for later phases. Later

sampling can be more efficient in the use of both time and financial resources to meet the goals of the sampling program (USEPA, 1992).

## 2. **Stratified random sampling**

In stratified sampling, the population of N units is divided into subpopulations of $N_1$, $N_2$,…,$N_L$, units. The subpopulations do not overlap, and comprise the whole population, so that

$$N_1 + N_2 + \ldots + N_L = N$$

These subpopulations are called strata. After the strata have been determined, a sample is drawn from each, with drawings being made independently in different strata. Sample sizes are designated by $n_1$, $n_2$,…, $n_L$. Random samples taken within each stratum marks the procedure as stratified random sampling (Cochran, 1977).

Cochran (1977) wrote:

> Stratification is a common technique. There are many reasons for this; the principal ones are the following.
>
> 1. If data of known precision are wanted for certain subdivisions of the population, it is advisable to treat each subdivision as a "population" in its own right.
> 2. Administrative convenience may dictate the use of stratification; for example, the agency conducting the survey may have field offices, each of which can supervise the survey for a part of the population.
> 3. Sampling problems may differ markedly in different parts of the population. With human populations, people living in institutions (e.g., hotels, hospitals, prisons) are often placed in a different stratum from people living in ordinary homes because a different approach to the sampling is appropriate for the two situations. In sampling businesses we may possess a list of the large firms, which are placed in a separate stratum. Some type of area sampling may have to be used for the smaller firms.
> 4. Stratification may produce a gain in precision in the estimates of characteristics of the whole population. It may [be] possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous. This is suggested by the name strata, with its implication of a division into layers. If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any

> stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population. (89–90)

Knowledge of sample characteristics in each stratum is essential into dividing the sample population into homogeneous subpopulations. The main purpose of stratified random sampling is to increase the precision of the estimates made by sampling, which is accomplished when units within each subpopulation are more homogenous than the total population (USEPA, 1983).

The major limitation of stratified random sampling is that reliable prior knowledge of the population is necessary to effectively define the strata and allocate the sample sizes. Any gains in precision or reductions in cost depend on the quality of the information used to set up the stratified sampling design. Often times, samplers go into a site without adequate information to implement this design. In addition, an investigator may encounter difficulties gaining access to sampled locations placed randomly in the field (USEPA, 2002).

3. **Systematic sampling**

In systematic sampling, N units of the population are numbered 1 to N. To select a sample of n units, a unit is taken at random from the first k units and then every *k*th unit after. The selection of the first unit determines the whole sample, and is known as an every *k*th systematic sample. In effect this creates n strata, which consists of the first k units, the second k units, and so forth. The difference between this and typical stratified random sampling is that with the systematic sample the units occur within the same position in the stratum, while in stratified random sampling the position of the unit is determined separately by randomization within each stratum. The systematic sample is spread more evenly over the population and this can make systematic sampling more precise than stratified random sampling (Cochran, 1977).

Spatially, systematic sampling is applied in a grid pattern, with the grid randomly placed at a starting location. If the sampling objective is to estimate spatial patterns or trends in the target population using geostatistical methods, then this design is optimal. It is also useful in estimating statistical parameters of the target population, such as the mean and the variance, when the systematic pattern of locations does not coincide with a spatial pattern of contamination that could cause a bias in estimating those parameters (Gilbert, 2005).

This particular technique has an advantage over random sampling in that it is often easier to execute without mistakes, especially if the units are chosen at or near the center of the strata (Cochran, 1977). Holmes (1970) notes that there is a viewpoint among statisticians that randomization is an inferior approach in plane sampling. He goes on to say that the systematic sample will be more precise provided care is taken in selecting the appropriate sampling interval. For plane sampling, the loss of information inherent in randomization is wasteful and unnecessary. The USEPA (2002) states that the benefits of systematic sampling include multiple options for implementing a grid design, which can be useful for multiphase sampling. Regularly spaced samples allow for spatial correlations to be calculated if the pattern of interest is larger than the spacing of the sampled nodes. In addition, grid designs can be implemented when little or no prior information exists about a site, and are often used for pilot studies and exploratory studies if the assumption is that there are no patterns or regularities in the distribution of the contaminant of interest.

However, systematic sampling may not be as efficient as historical records if prior information is available about the population. Background knowledge from exploratory or pilot studies can be used as a basis for stratification or identifying areas of higher likelihood of finding properties of interest, such as hotspots. In addition, if the properties of interest are aligned with

the grid, there is a possibility that systematic sampling may overestimate or underestimate a characteristic of the population (USEPA, 2002). Milne (1959) explains:

> Both Finney [1947, 1948, 1950] and Yates [1953] have pointed out the danger to systematic sampling arising from unsuspected: (a) periodic variation, (b) consistent increase, in unit value along the direction of the sampling lines; and (c) "marked strip effects running in straight lines across the material in such a manner that the whole of one line of sample points falls on the same strip." [Yates, 1953, 286]

Marked strip effects are relatively rare in nature, but can be imposed by drastic human activities, such as draining and cultivating. Likewise, periodic variation rarely occurs naturally, but could be caused by anthropogenic sources.

## B. **Geostatistical Methodology**

Most classical statistical methods do not make use of the spatial information in earth science data sets. Geostatistics offers a set of tools to analyze the spatial continuity that is an essential feature of many natural and anthropogenic phenomena, and provides adaptations of classical regression techniques to take advantage of this continuity (Isaaks & Srivastava, 1989).

Some believe that a geostatistical methodology can cut down on sampling costs and time. A Department of Energy estimate indicates that the department will spend between $15 and $45 billion dollars for analytical services alone over the next 30 years to support environmental restoration activities at its facilities (Johnson, 1996). Johnson (1996) proposes an adaptive sampling framework that relies on a coupled Bayesian/geostatistical methodology as the potential for substantial savings in the time and cost associated with characterizing the extent of contamination. Bayesian analysis allows the quantitative integration of "soft" information—such as historical information, non-intrusive geophysical survey data, preliminary transport modeling results, and past experience with similar sites—with hard data. Geostatistical analysis provides a

means for interpolating results from locations where hard data exists to areas where it does not using methods such as indicator kriging.[1] Johnson notes that the challenge for adaptive sampling programs is providing real-time sampling program support that incorporates the significant amount of soft information available and accounts for the spatial autocorrelation that is typically present.

McBratney, Webster, and Burgess (1981) describe a study in which the design of an optimal sampling scheme is based on the theory of regionalized variables,[2] and assumes that spatial dependence is expressed quantitatively in the form of the semi-variogram.[3] It also assumes that the maximum standard error of a kriged estimate is a reasonable measure of sample spacing. If variation is isotropic[4] a regular triangular or rectangular grid is used, and the maximum standard error is kept to a minimum for any given sampling. If the variation is geometrically anisotropic[5] grid spacing is greatest in the direction of major correlation, and smallest in the direction of minor correlation.

---

[1] Kriging is a method that produces a distribution of possible estimates for each unsampled point. The estimations are a function of the surrounding neighbors (SADA Documentation, 2008, Ch. 28)

[2] Regionalized variable theory is a set of statistical principles that mathematically considers spatial function properties but that neglects the physical nature of the phenomenon under study. Regionalized variable theory uses random variables to model spatial functions (Olea, 1984).

[3] The semi-variogram method returns a measure of variance for any given distance of separation and essentially calculates the degree to which data are more or less "alike" for any given distance. This measure is defined as half of the average squared difference between values separated by distance h. The term h is the lag distance. The equation is:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (x_i - y_i)^2$$

where N(h) is the number of pairs separated by distance h, $x_i$ is the starting sample point (tail), and $y_i$ is the ending sample point (head) (SADA Documentation, 2008, Ch. 30).

[4] Isotropy refers to the spatial phenomenon where data do not tend to be more alike in any one direction than any other (SADA Documentation, 2008, Ch. 28).

[5] Anisotropy refers to the spatial phenomenon where data tend to be more alike in a particular direction than another (SADA Documentation, 2008, Ch. 29).

Their methodology can be summed up as follows, (D) being decisions, (C) being computations, and (F) being field-work:

D1 Choose the maximum error allowed $\sigma K_{max}$ and block size.
D2 Decide the level of presurvey information required.
(a) If the semi-variogram is known or can be inferred, then go to C3.
(b) If the scale of variation is known or can be inferred, then go to D3.
(c) Else, nothing is known or can be inferred about the variable in the region of interest and the scale of variation first should be obtained using F1 and C1.
F1 Obtain the scale of variation used a nested design (Youden and Mehlich, 1937).
C1 Calculate nondirectional semi-variograms for nested design (Miesch, 1975).
D3 Choose transect sample interval from nondirectional semi-variogram and preset $\sigma^2 K_{max}$ remembering that this sample interval should be considerably less than the final grid spacing to obtain a useful experiment semi-variogram.
F2 Sample transects in 3 or more directions with randomly-located starting points.
C2 Calculate experimental semi-variograms and fit a model.
C3 Obtain grid spacing $a$ for direction of maximum variation $\phi$ using the method described previously for both triangular and square grids. The grid spacing in direction $\phi + \pi/2$ is $ra$, where $r$ is the anisotropy ration. If the semi-variogram is isotropic, the grid can be oriented in any direction and the grid spacing are equal in both directions.
D4 Choose to sample on a triangular or square grid. Only in exceptional circumstances will the efficiency advantage of the triangular grid out-weight the inconveniences of extra travelling, site location and computer handling.
F3 Sample on grid in direction $\phi$ rad with spacing $a$ and $\phi + \pi/2$ rad with spacing $ra$ (McBratney et al., 1981, 334).

Olea (1984) presents a procedure to minimize the sampling requirements necessary to estimate a mappable spatial function at a specified level of accuracy. The technique is based on universal kriging, an estimation method within the theory of regionalized variables. The *average standard error* and *maximum standard error of estimation* over the sampling domain are used as global indices of sampling efficiency. These measures depend on several unmanageable factors, such as the semivariance and the drift, and several manageable factors, such as the size of the sample subset of nearest neighbors considered by the estimate, the sample pattern, and the sample density. The procedure optimally selects those parameters controlling the magnitude of

the indices, including the density and spatial pattern of the sample elements and the number of nearest sample elements used in the estimation.

Olea states:

> Most spatial functions of a geologic nature can be known only partially through scattered sets of expensively gathered measurements. Observations of a spatial function constitute a statistical sample. However, because spatial functions possess continuity and each location is unique, classical statistical theory and sampling procedures are not applicable. Rather, we must turn to a special statistical theory which explicitly considers spatial properties, the theory of regionalized variables. (369–370)

Regionalized variable theory and the kriging methodology have a closer connection with classical statistics than classical sampling theory (the design-based approach), in that both are based on similar stochastic models (Brus & de Gruijter, 1997).

The global indices depend upon three factors under control of the experimenter: the number of nearest neighbors used in the estimation procedure, the spatial pattern of the sample points, and the density of the points across the mapped area. The indices decrease slowly and monotonically by increasing the density of sample elements. The index level must be selected according to the cost of data gathering, the further uses of collected information, and the amount of uncertainty that is acceptable in the study (Olea, 1984).

Oliver and Webster (1986) continue the study of regionalized variables in the context of nested sampling. Pollutants vary continuously and randomly in space, but the pattern and scale of the variation is not readily apparent. They postulate that the semi-variogram of regionalized variable theory provides a precise solution to identify the scale and pattern of variation of continuous spatial variables once the approximate scale of spatial variation is known.

The semi-variogram is a key tool of modern geostatistics. It can provide a concise and unbiased description of the scale and pattern of spatial variation. The semi-variogram can be

estimated from the sample values, and once a suitable mathematical model has been fitted to the values of the experimental semi-variogram, its parameters can be used for local estimation by kriging, and for optimizing sampling. It is estimated at regular intervals of spatial lag, preferably from a regular systematic sample; however, this procedure limits the range of spatial variation that the semi-variogram can reveal. Unless investigators already roughly know the spatial scale of the major source of variation they may sample either too sparsely to identify it if the range is short or unnecessarily intensively if only long-range variation is present (Oliver & Webster, 1986).

Barnes (1988) presents a study to minimize the kriging variance for secondary sample designs[6] during geologic site characterization. He offers alternatives to the use of *global kriging variance*[7] as a heuristic criterion in sample network design. These alternatives minimize the *average local kriging variance*[8] and the *maximum local kriging variance.*[9]

The objective of initial sample design is to collect enough information to assemble an original model for the site under investigation. After building the model the sampling objective changes to minimizing the chance of surprises, or, essentially, to minimize the probability of the existence of unknown features which would trigger a radical modification to the current model.

---

[6] Secondary sample designs are applied after some data and other information has been obtained. The objective is to refine the model or the decision in a specific way. Secondary sampling designs can be either point (sample) or model (geospatial model) based (SADA Documentation, 2008, Ch. 37).

[7] The global kriging variance is the estimation variance that results from using the available samples to estimate the average of the entire area of interest (Barnes, 1988, 811).

[8] The average local kriging variance is most easily explained by partitioning an area of interest into "L" pieces. Use the available samples to estimate the average value of each individual piece. Associated with each of these L estimates is a kriging variance. Sum all of these L kriging variances and divide by L. The average local kriging variance is then defined as the limit of this arithmetic average of kriging variances as the area of interest is divided into more and more pieces: the average point kriging variance over the area of interest (Barnes, 1988, 812).

[9] The maximum local kriging variance process begins as stated above for the average local kriging variance, however, after averaging the value of each individual piece determine the maximum of the L associated kriging variances. The maximum local kriging variance is then defined as the limit of this maximum of kriging variances as the area of interest is divided into more and more pieces: the maximum point kriging variance over the area of interest (Barnes, 1988, 812–13).

Barnes claims that the maximum local kriging variance criterion is more appropriate than the global kriging variance when the issues at hand are concerned with discovering unidentified extremes across the area of interest that often occurs when investigating a potentially contaminated site. If the objective of the sampling program is to determine extremes, additional samples should be placed in zones with high estimation uncertainty and that have a reasonable probability of being an extreme value (Barnes, 1988).

Englund and Heravi (1993) acknowledge the work of Barnes (1989) and other applications that involve variations on the minimization of the kriging variance to determine ideal sampling patterns, to identify the best locations for one or more proposed additional samples, or to find the minimum number of samples needed to attain a specified maximum level of error. They state that this approach does not incorporate economics, nor does it readily permit evaluation of the consequences of decision-making with uncertainty. The sample design procedure they propose is a Monte Carlo resampling scheme which simulates the remediation operation, including data collection, interpolation, and selection.

Englund and Heravi (1994) continue their investigation into ways of improving the cost-effectiveness of site characterization and remediation by examining the relative effectiveness of three alternate approaches to sampling contaminated soils to determine whether remediation is needed. They conducted experiments in phased sampling, with one-phase, two-phase, and *N*-phase design algorithms used on surrogate models of sites with contaminated soils.

To elaborate on phased sampling:

> A sampling phase, as used here, involves an interruption of the sampling process until the data from all prior sampling is available for interpretation. In one-phase designs, the results of the chemical analyses will only become available after sampling has been completed. In two-phase sampling, preliminary estimates of contaminant concentrations based on data from the first phase will be used to determine locations where additional

> data are needed most. The N-phase approach takes phased sampling to its extreme—every sample is a discrete phase. Estimates will be updated after each sample measurement and used to determine the most critical location for the next sample. In practice, this would require a fast, field-portable measurement system (Englund & Heravi, 1994, 248).

They conclude that optimal two-phase designs are better than one-phase designs, and *N*-phase designs are better yet. Short of advocating *N*-phase designs for every sampling program, it is noted that in practice this is not realistic. The improvements in cost are relatively small, and must be balanced with increased costs, such as logistical costs associated with the delay between phases—i.e., remobilization costs. If the measurements involve laboratory analysis, costs will be prohibitively high. Ultimately, the optimal, cost-effective design comes from a two-phase model, with the optimum total number of samples determined from the one-phase algorithm, and 75% of those samples assigned to the first phase (Englund & Heravi, 1994).

Kravchenko and Bullock (1999) conducted a study to evaluate three interpolation techniques in order to find the optimal method for mapping soil properties. They evaluated inverse distance weighting, ordinary kriging, and lognormal ordinary kriging. They concluded that lognormal ordinary kriging can be expected to produce better estimations for lognormally distributed data than ordinary kriging. However, for some data sets lognormal ordinary kriging can result in biased estimations, with relatively high negative mean error between measured data and estimates. They note that ordinary kriging seems to be a safer choice than lognormal ordinary kriging for data sets with more than 200 data points. Kriging with the optimal number of neighboring points, a carefully selected variogram model, and appropriate transformation of data produces more accurate estimations than the inverse distance weighting method.

Investigation of the progress of the kriging algorithm suggests that initial gains are made by moving apart points that are too close together, moving some points into regions where the

initial random design is sparse, and making some adjustments for boundaries. The algorithm usually makes substantial improvement over the initial design with the first few iterations, but quickly reaches a broad valley in the search space and subsequently takes an extremely long time to converge to the optimal design (Cox et al., 1995).

## C. Classical Site Sampling Theory versus Geostatistical Methodology

Many environmental contaminants originate from a point source, and rather than forming a uniformly distributed pattern, their concentrations are often very high near the source, but fall off rapidly as the distance from the source increases. It would make sense, then, to sample more intensively in those highly concentrated regions. However, it is not easy to do so and obtain an unbiased estimate of the total. Even kriging does not necessarily give an unbiased estimate from a preferential sample (McArthur, 1987).

Brus (2010) outlines the differences between sampling approaches:

> In a design-based approach sampling locations are selected by probability sampling, and the statistical inference (e.g., estimation of spatial mean) is based on the sampling design. In a model-based approach there are no requirements on the method for selecting sampling locations, and typically are selected by purposive (targeted) sampling, for instance on a centered grid. In statistical inference a model for the spatial variation is introduced, e.g., an ordinary kriging model, assuming a constant (unknown) mean, or a universal kriging model in which the mean is modelled as a linear function of one or more predictors. Besides the deterministic part for the mean, a kriging model contains a stochastic part describing the variance and covariance of the residuals of the mean. Note that the source of randomness is different in the two approaches. In the design-based approach the selection of the sampling locations is random, whereas in the model-based approach randomness is introduced via the model of spatial variation. In the design-based approach no such model is used. This has important consequences for the interpretation of measures of uncertainty about estimates, e.g., the variance of the estimation (prediction) error. (32)

McArthur (1987) uses simulated data to evaluate several methods for sampling an area and estimating the total amount of a pollutant known to be concentrated in one area of a region. He concludes that stratified random sampling and importance sampling[10] emerge as the best methods for estimating the mean level of a contaminant that is known to be locally concentrated around a source. Both methods give unbiased and reasonably precise estimates of the mean and its sampling variance, while the other sample designs either give biased or imprecise estimates. Often, however, we are seeking extreme values that surpass a specified threshold. His methods might best be suited to the initial site or hot spot characterization and not to an environmental risk assessment or human health evaluation.

De Gruijter and ter Braak (1990) compare the design-based approach (classical) with the model-based approach (geostatistical):

> In the model-based approach, locations need not be selected at random. They typically aren't. The only source of stochasticity is then the postulated underlying process. In this approach inference is therefore primarily based on the model formulated. The nature of the stochasticity involved in the model-based approach is thus fundamentally different from that in the design-based approach where, as we have seen, it originates from a physical sampling process. The latter is in our hands. The design-based approach thus requires fewer assumptions than the model-based approach. It is therefore advantageous with respect to robustness to use the design-based approach whenever possible. (408)

They insist that a design-based approach is inapplicable if probability sampling is impracticable, and a model-based approach is inapplicable if reliable identification of a model is prevented by lack of data. However, geostatistical models may form a natural basis for inference in situations where part of the region is inaccessible, or measuring has been censored or impaired. If data are

[10] Importance sampling uses a Monte Carlo method for computing a multiple integral to formulate a probability density function (McArthur, 1987, 745–46).

available from the vicinity of the region these can be used via the model-based approach, as in kriging.

De Gruijter and ter Braak (1992) go on to refute Barnes' (1988) publication on the sample design for geologic site characterization:

> This strategy is inferior to the basic design-based strategy of Simple Random Sampling combined with Eq. (1) for the following reasons.
>
> 1. Barnes' strategy would normally require a considerably larger sample size for the same coverage probability. Under Simple Random Sampling not more than $\log (1 - P) / \log(\beta)$ samples are needed (Eq. 2), where $P$ denotes the required probability of coverage. This is the lower bound of the number required by Barnes' strategy.
> 2. Barnes' strategy is approximative only and liable to impairment by model errors, whereas Simple Random Sampling with Eq. (1) is exact and independent of any model.
> 3. Barnes' strategy is much more complicated. (863)

Brus and de Gruijter (1997), in a comparison of design-based and model-based sampling strategies, came to the conclusion that both the model-based approach and design-based approach are valid for spatial sampling and estimation. Many studies have declared that independence in the design-based approach is not met due to spatial autocorrelation,[11] however, Brus and de Gruijter insist that independence is not assumed but created by the sampling design. The model-based approach is not necessarily optimal if only one realization is considered. The authors outline an approach to choosing between the two methods given certain factors using a decision-tree.

[11] The idea of autocorrelation, which implies that variables are spatially dependent (i.e., variables closer together have more in common than those farther apart), is central to the theory of geostatistics (McBratney et al., 1981).

D. **Hybrid Designs**

Pettitt and McBratney (1991) consider sampling designs and estimation procedures for the spatial variogram (the semi-variogram) when no information on magnitude or scale of variation is available. Conventionally, two approaches to designing an exploratory (or any other) spatial survey are available: a purely design-based approach with some kind of random design or a model-based approach based on some systematic design. The former may be difficult to implement in the field and the latter runs the risk of "superpopulation"[12] model misspecification.

They propose a hybrid approach to the problem:

> There are advantages and disadvantages with the design-based variance components and the model-based geostatistical approaches. We have devised a statistically and practically efficient novel scheme which is a hybrid between the two standard approaches. This consists of linear transects with exponentially spaced sampling locations uniformly oriented in three directions (Pettitt & McBratney, 1991, 205–206).

They acknowledge that the variance components model implies a non-decreasing variogram which in some circumstances may be an unrealistic assumption. They did not explore this possibility further.

Cox et al. (1995) state:

> Sources and effects of bias in environmental sampling need to be identified and studied. Techniques for examining correlation structure to determine the effective sample size of a design are needed. The issues surrounding combining design-based (e.g., regular designs) and model-based (e.g., conditional simulation) approaches in spatial design and analysis need to be stated and examined. For example, can soil pollution concentrate data collected on different designs be combined across all or a sample of the hazardous waste sites in a region to provide meaningful regional pollution characterization and remediation cost information? (24)

---

[12] A superpopulation is a hypothetical infinite population from which the finite population is a sample and is integral to the model-based approach and geostatistical theory (De Gruijter & ter Braak, 1990).

Brus and de Gruijter (1997) point out that there are no examples in soil science applying the model-assisted approach, which utilizes design-based sampling strategies that make use of a model. The role of a model in design-based inference differs from that in the model-based approach as the latter describes a process by which the data have been generated, whereas in the former it describes the population itself.

## E. **Conclusions**

As outlined above there are many theories and methods to address sample design. Other than random sampling, there does not seem to be an agreed upon methodology. Random sampling lends statistical validity to the sampling design, and may be cost effective; however it is not highly useful when needed to detect hotspots or plumes of contamination created by geospatial variables, and does not provide a complete picture of the spatial distribution of contamination without addressing secondary sampling methods. Most secondary sampling designs built on top of random samples, such as adaptive sampling, however, fail to take into account the prohibitive costs of sampling in a real world situation. Grid designs like systematic sampling are effectively implemented when little or no information exists for the site, but may overestimate or underestimate a characteristic of the population if the grid is aligned with properties of interest. These designs also do not consider the spatial information in earth science data sets. Geostatistical methods incorporate the idea of spatial autocorrelation, that variables closer in proximity are more alike than those farther away, and kriging is effective for interpolating from areas where data exist to areas where they do not, however the models may be biased due to lack of information and may not be optimal if only one realization is considered.

The underlying correlation model that is based on the semi-variogram operates under many assumptions and errors can vastly affect the outcome of the model.

## III. PROBLEM STATEMENT

The issue is that given a particular site and/or data set, it is not clear which LISA secondary sampling design should be used to provide the most rational outcome—that of optimal sample placement to characterize the spatial distribution of the contaminants or the identification of hot spots. Each of these secondary sampling designs are based on a set of general parameters that the user inputs. These manual inputs often generate a wide variety of outcomes. While they allow SADA to be a versatile tool that can adapt to what the user wants to accomplish, they also create a wide margin of error and uncertainty when misused or poorly understood. Furthermore, the LISA methods (Ripley's K, Moran's I, and Geary's C) give a sense of spatial sampling density and spatial variability by specifying a search window around individual observations that assesses the amount of neighbors or the local variance/correlation of the data points found there. They can be potent tools when used correctly, however, the manual user input required for the search radius once again opens up these algorithms to a large amount of uncertainty when used to develop a secondary sampling design.

The objective of this study was to demonstrate the use of SADA to identify secondary sampling locations by taking into account data from a previously sampled site. In addition, I wanted to assess the limitations of the existing LISA parameters in SADA, and to identify guidelines that allow SADA users to utilize the secondary sample designs, in particular the LISA designs, effectively. Lastly, I wanted to assess the applicability of the LISA methods and to identify and potentially optimize the critical parameters to gain a cost effective, practical, reliable, and defensible method of characterizing the spatial distribution of contaminants.

# IV. SITE SELECTION

## A. Site Background

The data set used was sample data from an actual USEPA Superfund site located in the Midwest where a zinc smelter operated. The identity and location of the Superfund site needs to remain confidential, and certain landmarks have been removed from any images of the site to conceal the identity of the study area. The site is in close proximity to neighborhoods and contains elevated levels of arsenic and lead contamination due to the smelter operations. The amount of samples initially placed were 50, and each sample point contains a median arsenic and a median lead concentration in parts per million (ppm). Historically, the samples were collected by USEPA. Site-specific observations were recorded in a Microsoft Excel (Excel) spreadsheet with the following attributes: x, y coordinates, decimal degrees, global positioning system (GPS) identification, date of sampling, median lead concentration in ppm, and median arsenic concentration in ppm.

The data from the spreadsheet were imported in to ArcGIS, a geographical information systems software. The x, y coordinates of each sample was plotted with this software to obtain a spatial representation of the Superfund site samples. An aerial image was added to visualize the site. A boundary was drawn around the area the USEPA targeted for sampling. Figure 1 shows the area, which appears to be a residential zone and will be referred to as Parcel 1.

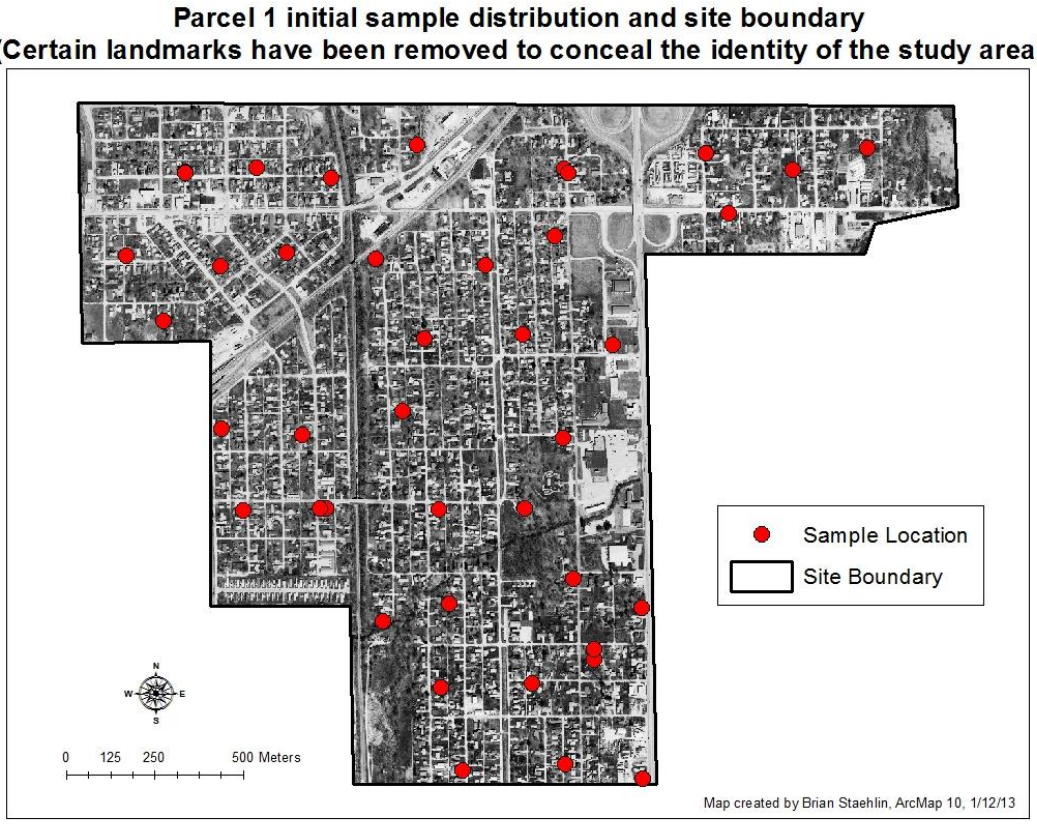


Figure 1. Sample placement and visualization of the Superfund site.

The data set was selected for this study for the following reasons:

- The site fulfills the criteria of a contaminated site.
- The site had historical investigation data that could be used as a base for secondary sampling methods.
- The sample size was small, reflecting the realities of a sampling program.
- The initial sample design was nearly systematic in nature, providing a crude estimate of the spatial distribution of contaminants.

B. **Site Characterization**

The 50 samples in Parcel 1 are arrayed in a grid-like pattern and appear to have been an unaligned systematic sample or a stratified random sample, potentially including judgment sampling or phased sampling around samples of concern (i.e., higher contaminant values). The phased sampling theory is also reinforced by the fact that some coordinates are sampled twice, indicating a desire for more information. In SADA, these locations are symbolized by a dotted circle around the stacked samples. Figures 2 and 3 below show the scale of contamination for both arsenic and lead.

Table I shows the descriptive statistics for arsenic and lead in Parcel 1. Figures 4 and 5 show the histograms for arsenic and lead, respectively. As the histograms show, the data is skewed; however, normal properties were calculated.

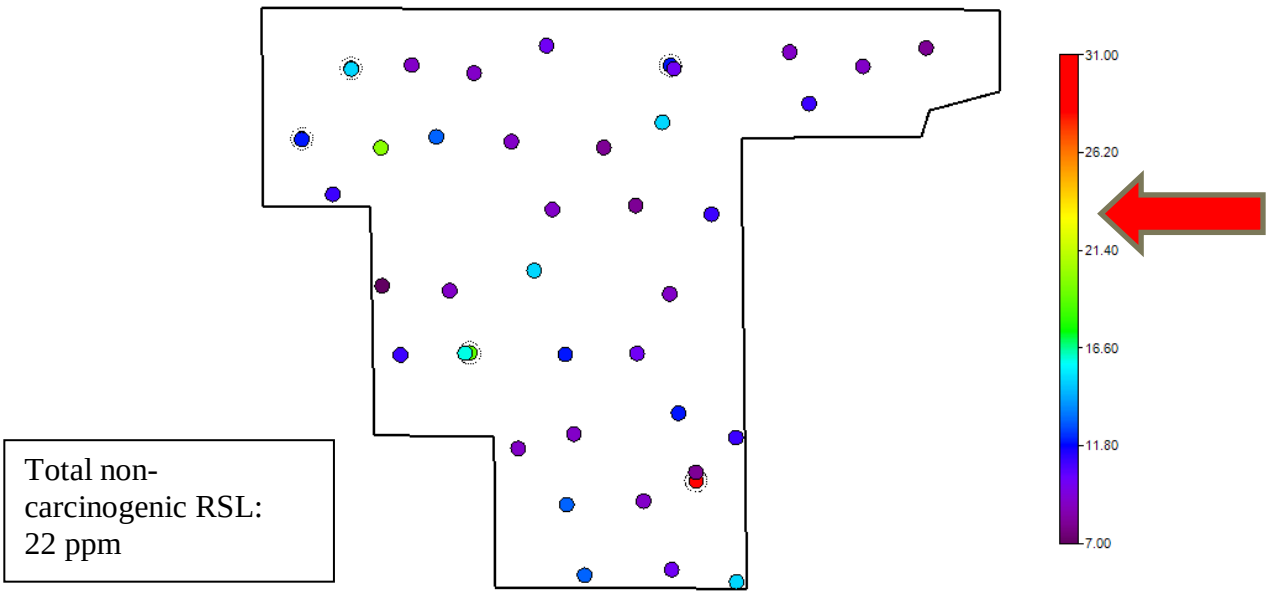


Figure 2. Arsenic values for Parcel 1 in ppm.

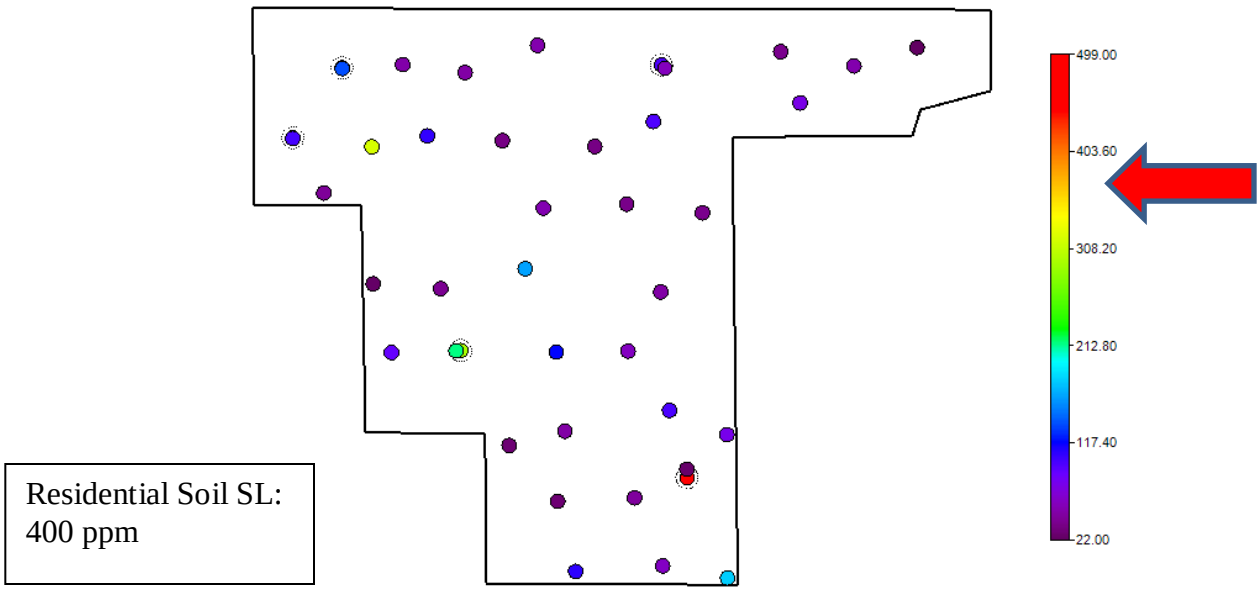


Figure 3. Lead values for Parcel 1 in ppm.

**TABLE I**

STATISTICAL ANALYSIS FOR ARSENIC AND LEAD FOR PARCEL 1

| **PARCEL 1** | | |
|---|---|---|
| | **Arsenic (ppm)** | **Lead (ppm)** |
| **Count** | 50 | 50 |
| **Arithmetic Mean** | 11.44 | 89.25 |
| **Median** | 10 | 52 |
| **Min** | 7 | 22 |
| **Max** | 31 | 499 |
| **Range** | 24 | 477 |
| **Variance** | 17.0294 | 7958.0128 |
| **Standard Deviation** | 4.1264 | 89.2077 |
| **Standard Error** | 0.5836 | 12.6159 |

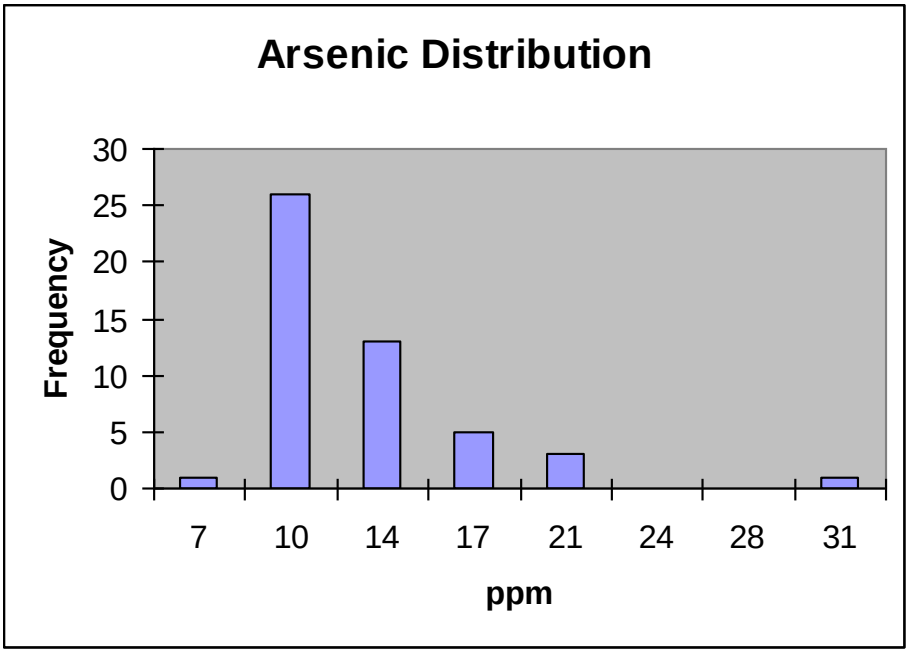


Figure 4. Arsenic histogram.

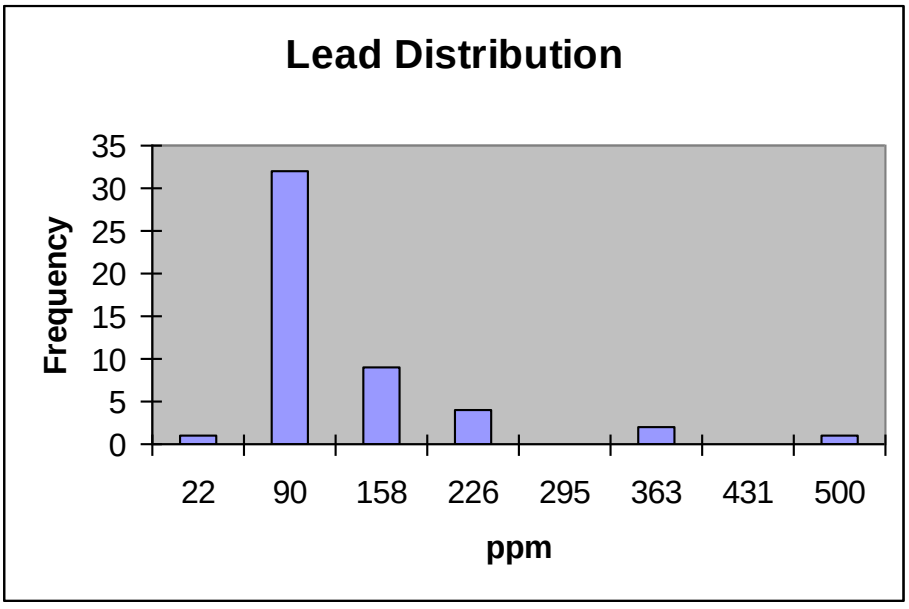


Figure 5. Lead histogram.

# V. SPATIAL ANALYSIS AND DECISION ASSISTANCE

## A. **Background**

Spatial Analysis and Decision Assistance began as an effort in 1996 between the University of Tennessee and the Oak Ridge National Laboratory's Environmental Restoration program. The purpose of this collaboration was to develop tools that would integrate human health and ecological risk assessment with geospatial processes in a way that could directly impact environmental restoration decisions. In the late 1990s and early 2000s the USEPA continued to support SADA, followed shortly thereafter by the USNRC. Through its development stages the authors have sought to maintain the original principles of the project: everyday applicability and ease of use (SADA, 2008).

Here is a broad list of SADA's capabilities:

- Data Explorationa and Visualization
- Geographic Information Systems
- Statistical Analysis
- Human Health Risk Assessment
- Ecological Risk Assessment
- Data Screening and Decision Criteria
- Geospatial Interpolation
- Uncertainty Analysis
- Decision Analysis
- Sample Design
- Multi-Agency Radiation Survey and Site Investigation Manual module

This study will focus on the three LISA methods of secondary sampling designs that SADA has to offer.

## B. Secondary Sampling Designs

Secondary sampling designs are applied after some background information or data have already been obtained. The general objective is to further refine the model or the decision in some very specific way. Secondary sampling designs are broken into two categories: point (sample) or model (geospatial) based (SADA, 2008). The geospatial designs can either be based on an interpolant, such as ordinary kriging, or on a LISA algorithm that is based on a search window that also factors neighboring points into the equation.

A LISA is any statistic that satisfies the following requirements:

- the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation, otherwise known as hot spots;
- the sum of LISAs for all observations is proportional (or equal) to a global indicator of spatial association.

A LISA can be expressed for a variable $y_i$, observed at location $i$, as a statistic $L_i$, such that:

$$L_i = f(y_i, y_j),$$

where $f$ is a function (possibly including additional parameters), and the $y_j$ are the values observed in the neighborhood $J_i$ of $i$ (Anselin, 1994).

### 1. Ripley's K design

The Ripley's K design locates samples in areas with the lowest sampling density. The Ripley's K statistic is a measure of neighborhood sampling density, and the value is

evaluated at nodes in the grid by specifying a simple search neighborhood about the nodes assessing the number of data points found there. In principle, it is similar to the Adaptive Fill design; the number of new samples is chosen by the user. But rather than locate the sample based on the furthest distance from the closest neighbor, the location is based on the node with the lowest sampling density in the nearby vicinity. Ripley's K requires the use of a minimum distance constraint to counteract the likelihood that nodes of low sampling density will be clustered together (SADA, 2008).

In addition, a LISA search radius must be specified. The LISA methods, which include Ripley's K, Moran's I, and Geary's C, are a set of functions that give some sense of spatial sampling density and spatial variability. For Ripley's K, a window of size $h$ is centered about each sample point and the number of sample points found within the window is computed. The window is then moved to every other sample point and the number is recomputed. The values are averaged producing an average value for the distance window $h$. The estimator for $K(h)$ is:

$$\hat{K}(h) = \frac{\sum_{i=1}^{N} \sum_{j=1(j \neq i)}^{N} w_{ij} I(h_{ij} \leq h)}{\lambda N}$$

Where $\lambda = N/|A|$, $N$ is the number of samples, $A$ is the area of the site, and $w_{ij}$ is a spatial weight used to account for edge effects near the boundary. The SADA produces a moving window of sample counts over an extent of grid nodes. For a specific distance of $h$, SADA creates a continuous map of count data using a defined base grid. For each grid node, the number of points is computed within a distance $h$. This provides a sense of the spatial distribution of clusters for a given distance $h$ (SADA, 2008).

2. **Moran's I design**

The Moran's I design is another LISA method that seeks to place new sample points in areas of high local sample variance. The idea is to collect more data in locations where greater uncertainty or variability exists (SADA, 2008). As in the Ripley's K design, a moving window of radius *d* is positioned at sample points around the site and the weighted variance of sample points within the window is computed. The estimator for *I(d)* is:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where *N* is the number of spatial units indexed by *i* and *j*; *X* is the variable of interest; $\bar{X}$ is the mean of *X*; and $w_{ij}$ is a spatial weight used to account for edge effects near the boundary (SADA, 2008).

3. **Geary's C Design**

The Geary's C Design is the final LISA method and seeks to place new sample points in areas with greater negative correlation among samples found in the search neighborhood. The idea is to collect more data in locations where greater uncertainty or variability exists. This design differs from Moran's I in that uncertainty is measured not by local variance but by local correlation. The more negative the correlation among data within the neighborhood, the more they are unalike (SADA, 2008). In Geary's C, semivariance is computed as follows:

$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(X_i - X_j)^2}{2W \sum_i (X_i - \bar{X})^2}$$

where $N$ is the number of spatial units indexed by $i$ and $j$; $X$ is the variable of interest; $\bar{X}$ is the mean of $X$; $w_{ij}$ is a spatial weight used to account for edge effects near the boundary; and $W$ is the sum of all $w_{ij}$ (SADA, 2008).

## C. General Spatial Analysis And Decision Assistance Parameters

Some sampling parameters are offered by SADA that are generally used across all secondary sampling designs. Many require manual inputs which provide SADA with much versatility when creating secondary sampling designs.

### 1. Number of samples

There are three options for choosing the number of samples to place:

- You Pick—a blank field allows the user to manually input the amount of samples they would like to be placed.
- Based on Sign Test—a non-parametric method based on the median value of the site.
- Wilcoxon Rank Sum Test—another non-parametric method, but with the difference that the Wilcoxon Rank Sum Test tests against another distribution instead of a single decision criteria.

### 2. Minimum distance constraint

Utilizing SADA, the user can input a minimum distance constraint in all of its secondary sampling designs. By specifying a minimum distance constraint, the user can create

"no sample zones" around existing samples and potential new sample locations (SADA, 2008). In some sampling designs samples may inherently cluster together in order to meet an optimization criteria. In many cases, however, the user may want to spread new samples throughout a particular region to provide good spatial coverage.

3. **Tie break options**

When multiple new sample locations fulfill the goal of a secondary sampling design a decision must be made so that only one of the new sample locations is chosen. Three types of tie break methods are provided by SADA (SADA, 2008):

- Random—a random number generator is used to select one of the sample locations. A seed is used to feed the random number generator. Using a seed allows the user to recreate a sample design by entering the same value.
- Maximize spatial coverage—the value that will maximize the spatial coverage of the site is chosen as a new sample.
- Closest to center of the site—the value that is closest to the center of the site is chosen as a new sample.

The random seed is an optional parameter. If kept blank and ties occur, then the tie breaker is decided each time the design is reapplied.

4. **Grid specifications**

Spatial Analysis and Decision Assistance allows the user to choose grid specifications based on number of cells in the Easting and Northing directions. The corresponding spacing between the nodes of these cells is also shown. The blocks formed by this

grid become the basis for the secondary sampling designs. Their purpose is two-fold: for geospatial based designs concentration values are modeled at each block across the site, and the nodes formed by the blocks provide sample locations.

## VI. IDENTIFYING GUIDELINES FOR GENERAL PARAMETERS

### A. Number of Samples

While it is useful to utilize the non-parametric methods offered by SADA to control Type I or Type II errors, the goal of this study is to achieve a methodology for selecting secondary sampling designs based on a more common occurrence. From a practical standpoint, sampling budgets will often dictate the number of samples that can be afforded. Typically, this is a smaller number than may be desired. Given the initial sample design and taking into account budget constraints that may be faced in the field, it was decided that 10 samples would be placed when exploring the secondary sampling designs.

### B. Random Tie Breaker

Out of the three options available—random, maximize spatial coverage, and closest to center of the site—it would appear that the random option offers the least amount of impact while exploring other, more crucial variables. This is also SADA's default option. Unless there is a reason for the user to maximize the spatial coverage of the site or to sample closer to the center of the site, the random tie breaker offers the least amount of bias. The random seed will be left blank by default.

### C. Grid Resolution

Default grid specifications will be calculated by SADA in order to get the user started with the secondary sampling design process. When this option was chosen for Parcel 1, SADA chose a grid of 50 nodes in the Easting direction and 50 nodes in the

Northing direction, as seen in Figure 6. The cell size is calculated by dividing the length of the site boundary in each direction by the number of cell blocks, in this case 50. The cell size in the Easting direction is 49.49 meters, and in the Northing direction is 38.80 meters.

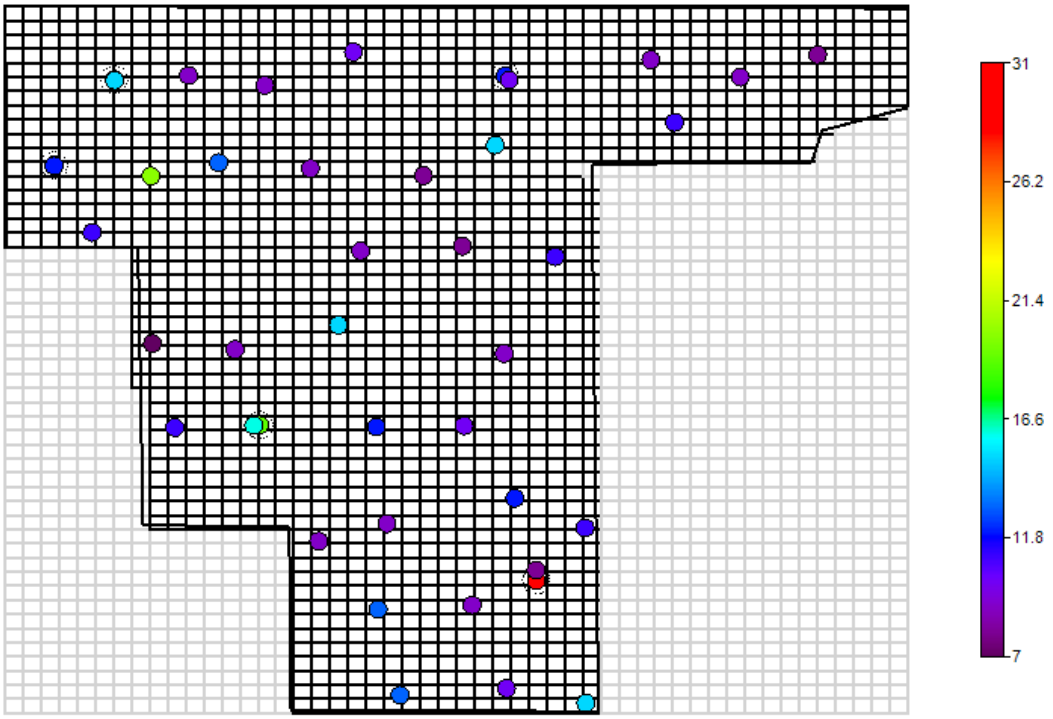


Figure 6. 50 x 50 grid in SADA.

The nodes outside of the Parcel 1 borders are not eligible for sample placement and are therefore grayed out. Doubling the number of cell blocks in each direction to provide a 100 x 100 grid derives the result shown in Figure 7.

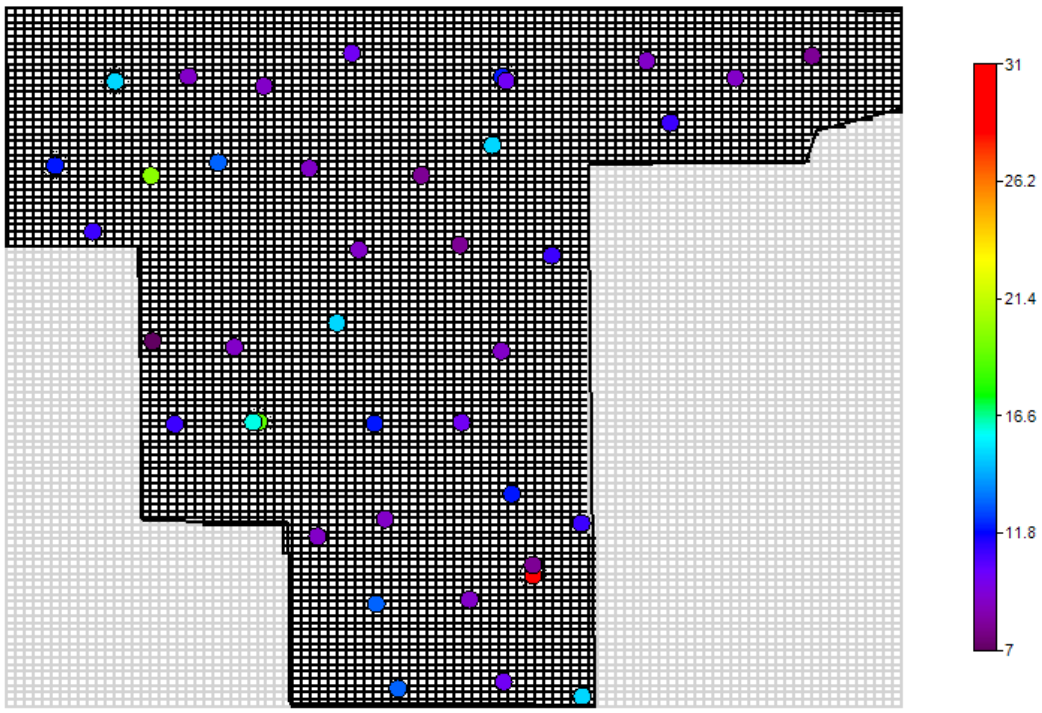


Figure 7. 100 x 100 grid in SADA.

The finer resolution allows the grid to align better with the uneven lengths of the Parcel 1 boundary.

To see the impact upon secondary sampling design a comparison must be made keeping all other variables constant. Ten samples are each placed on a 50 x 50 grid and on a 100 x 100 grid using the Moran's I method with an arbitrarily chosen LISA search radius. Triangles denote new sample locations. Figure 8 shows a side by side comparison of the results.

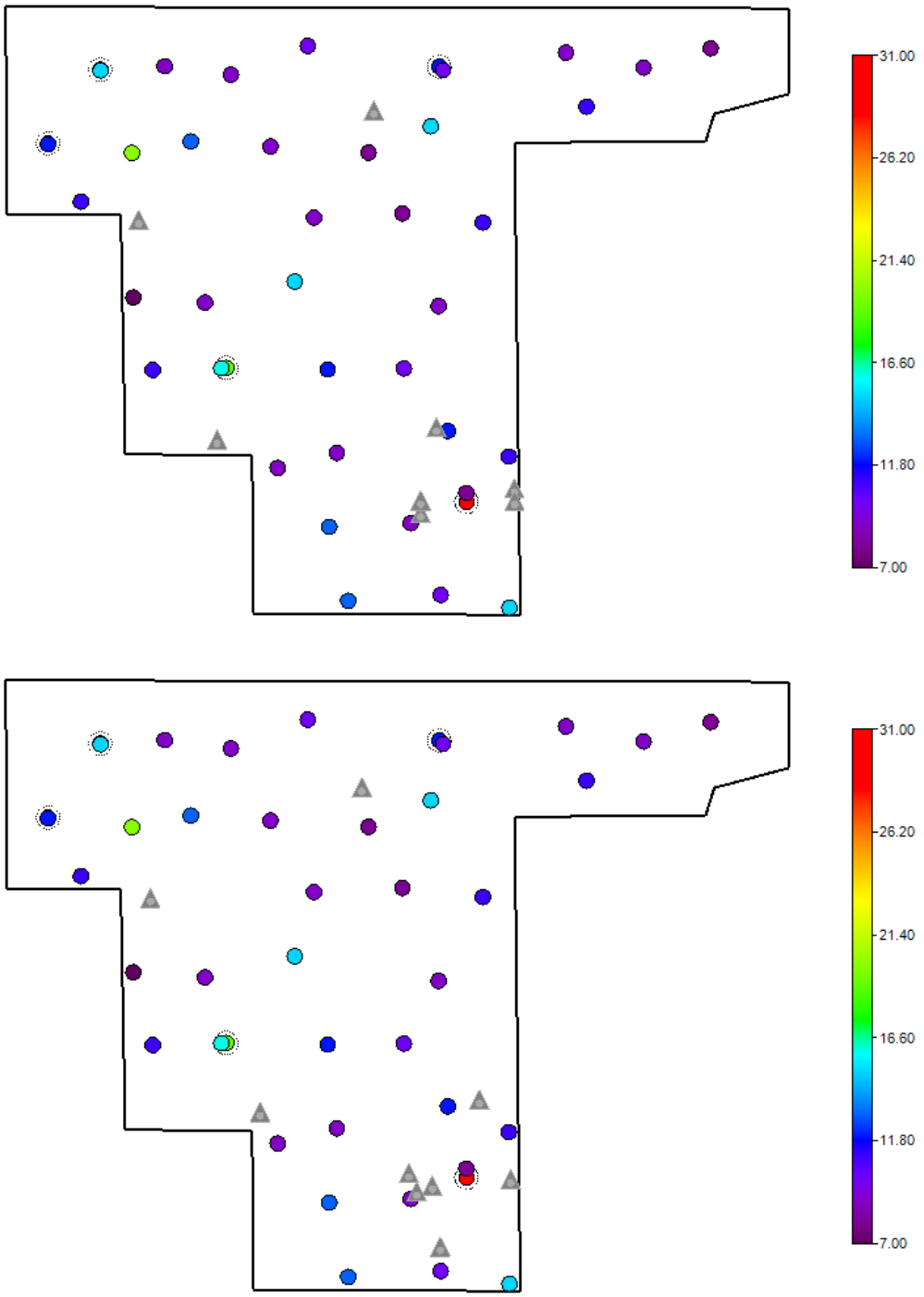


Figure 8. Ten samples placed with Moran's I algorithm using a 50 x 50 and a 100 x 100 grid.

For some points the difference between the two designs is slight. However, the biggest discrepancies revovle around the 31 ppm sample point in the bottom right corner of the site where there will be high local sample variance which is what the Moran's I algorithm measures.

In addition to this visual comparison it stands to reason that a coarser grid is not very optimal as there will be fewer nodes hence fewer locations to sample. Unless the user has pertinent information on a site that justifies the use of a coarser grid, such as representing a remedial unit, or an exposure unit, increasing grid density and using a finer grid makes practical sense. Doubling the grid specs once more derives a 200 x 200 grid. At this grid resolution it is difficult to see the individual cell blocks, which is apparent in Figure 9. Grid resolutions greater than 200 x 200 will also be difficult to distinguish from one another.

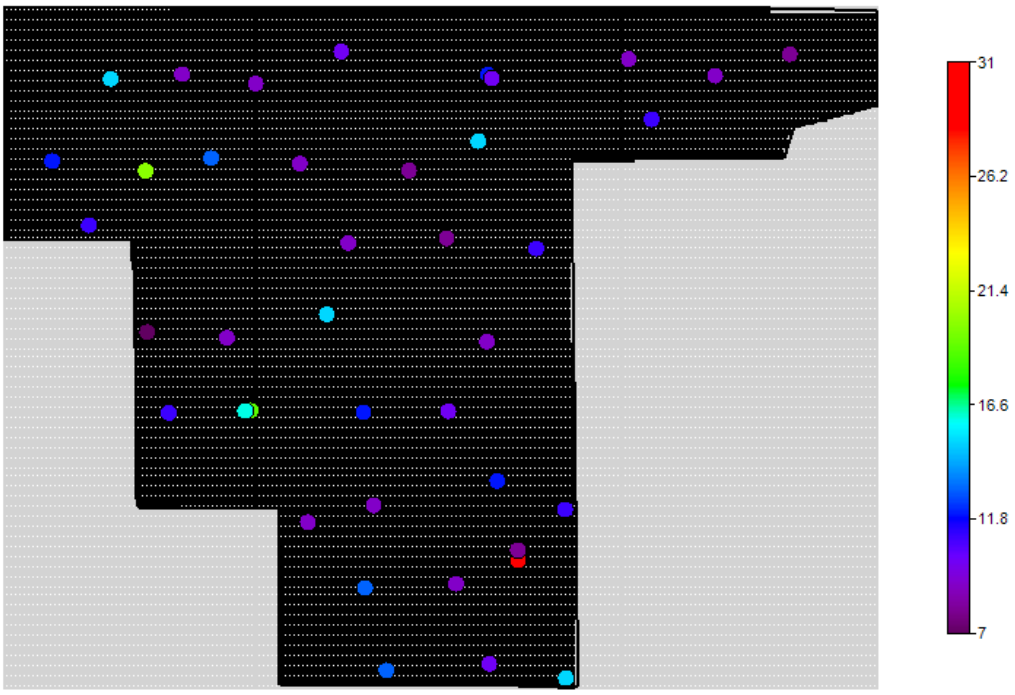


Figure 9. 200 x 200 grid in SADA.

As shown in Figures 10 and 11, with more locations available to sample the optimality of the Moran's I algorithm changes and begins to focus more on other areas of the site other than the bottom right corner. The finer the grid, the more samples will cluster at optimal locations. What then is the best grid resolution to use? With manual inputs the combinations are endless and the only real impediment to running designs with higher resolution grids is the amount of

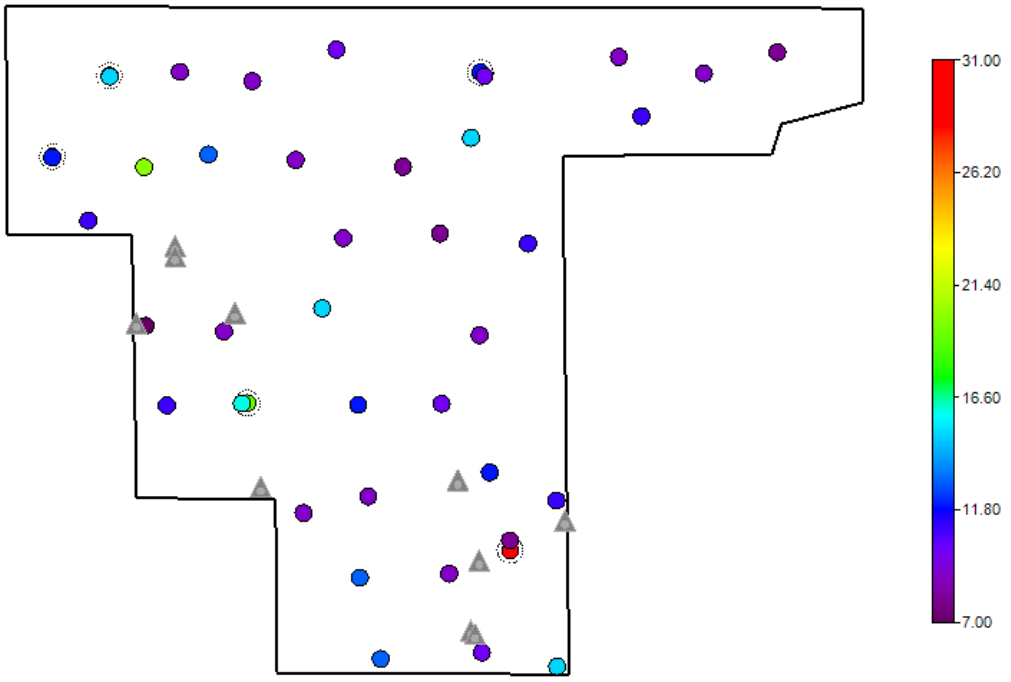


Figure 10. Ten samples placed with Moran's I algorithm using a 200 x 200 grid.

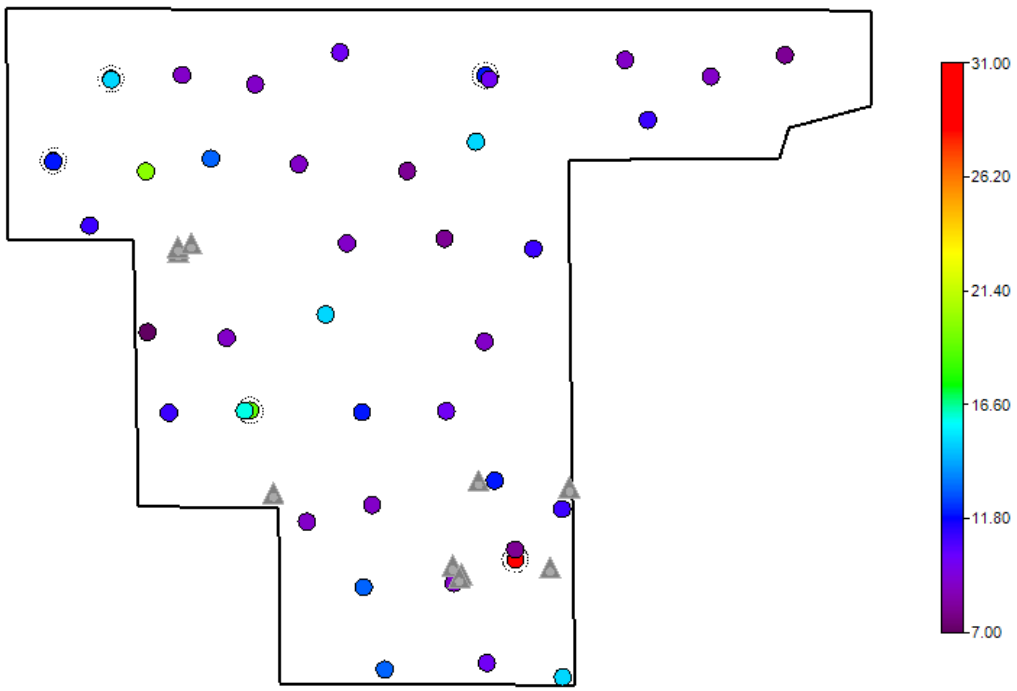


Figure 11. Ten samples placed with Moran's I algorithm using a 400 x 400 grid.

time it takes to place samples. The algorithm must consider each node for each sample to find the optimal location.

The question becomes, how does one practically choose grid specifications that mirror real world conditions? What are the constraints? Typically, sample designers use GPS devices to detect spatial coordinates. These GPS devices used by sample designers at USEPA can accurately detect spatial coordinates to within a certain distance, and in conversations with USEPA samplers, it was determined that the average distance GPS devices commonly used for sampling can accurately measure to was about +/- three (3) meters (+/- ten (10) feet). These devices are not high end devices which typically run in the range of thousands of dollars, but rather cost a few hundred dollars. To simulate this on the SADA grid, the user can input the size

of the cell blocks, i.e., the distance from node to node, in the Easting and Northing directions. The number of cell blocks in both directions would then be calculated by SADA. In the case that the cell size selected does not divide evenly into the site boundaries, SADA can either expand or contract the site boundaries so that the requested cell size works. This happens to be the case when a cell size of 3 x 3 meters is inputted into the grid specifications, but rather than change the site boundary, the user can play around with the cell block numbers and see the corresponding cell block size in order to achieve the right specifications. After some experimentation the grid specifications for Parcel 1 are set at 800 cells in the Easting direction and 650 cells in the Northing direction, which translates into nodes spaced out at 3.09 meters and 2.98 meters, respectively (see Figure 12).

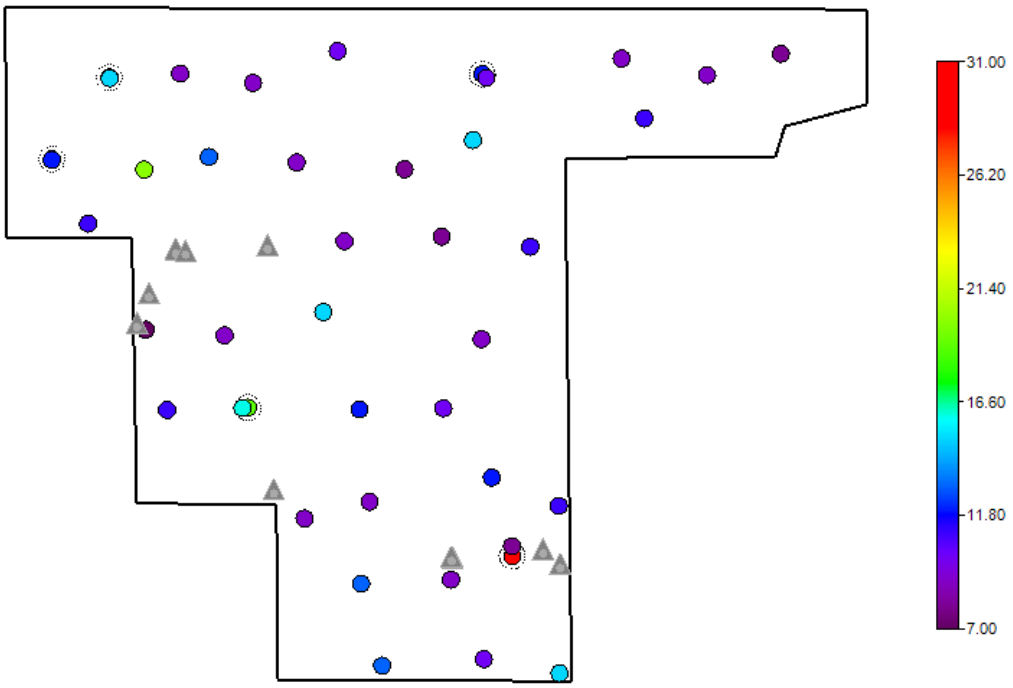


Figure 12. Ten samples placed with Moran's I algorithm using a 800 x 650 grid.

## D. **Minimum Sample Distance**

So far the minimum distance constraint has not been considered. In order to observe the outcomes and the feasibility of each secondary sampling design it becomes apparent that a very minimal sample distance restraint is required. The value chosen was three meters. This value is the most minimal, reliable distance that two samples can be distinguished using the common GPS devices mentioned in the previous section. It is worth briefly considering the impact of the minimum distance constraint. Figures 13, 14, and 15 show a Moran's I design with an arbitrarily chosen LISA search radius placing ten samples using an 800 x 650 grid with a minimum distance constraint of 0, 150, and 250 meters respectively. An attempt to place samples with a minimum distance constraint of 300 meters yielded zero samples placed; the initial sample design was too dense to permit the placement of new samples in this case. In the case of a 250 meter minimum distance constraint only three out of ten samples could be placed.

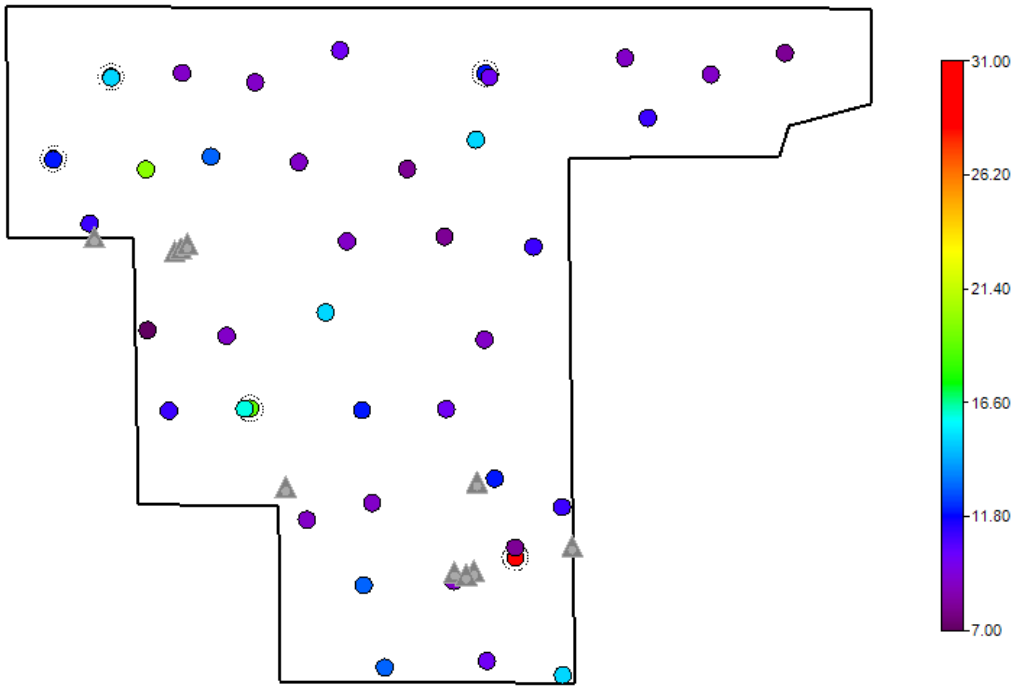


Figure 13. Ten samples placed with Moran's I algorithm at a minimum distance of 0 meters.

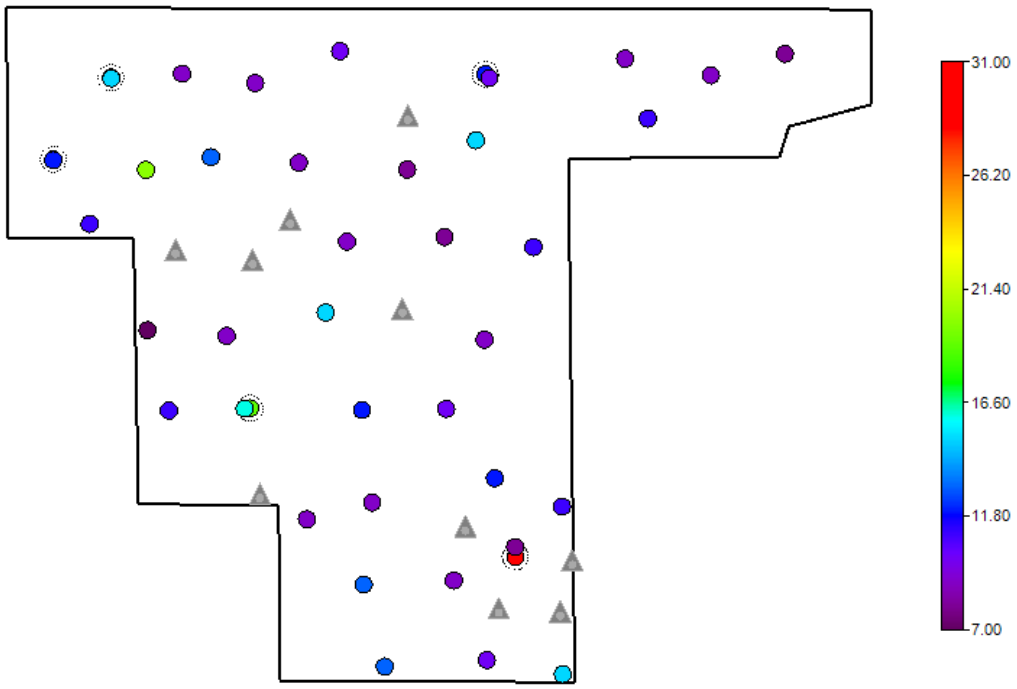


Figure 14. Ten samples placed with Moran's I algorithm at a minimum distance of 150 meters.

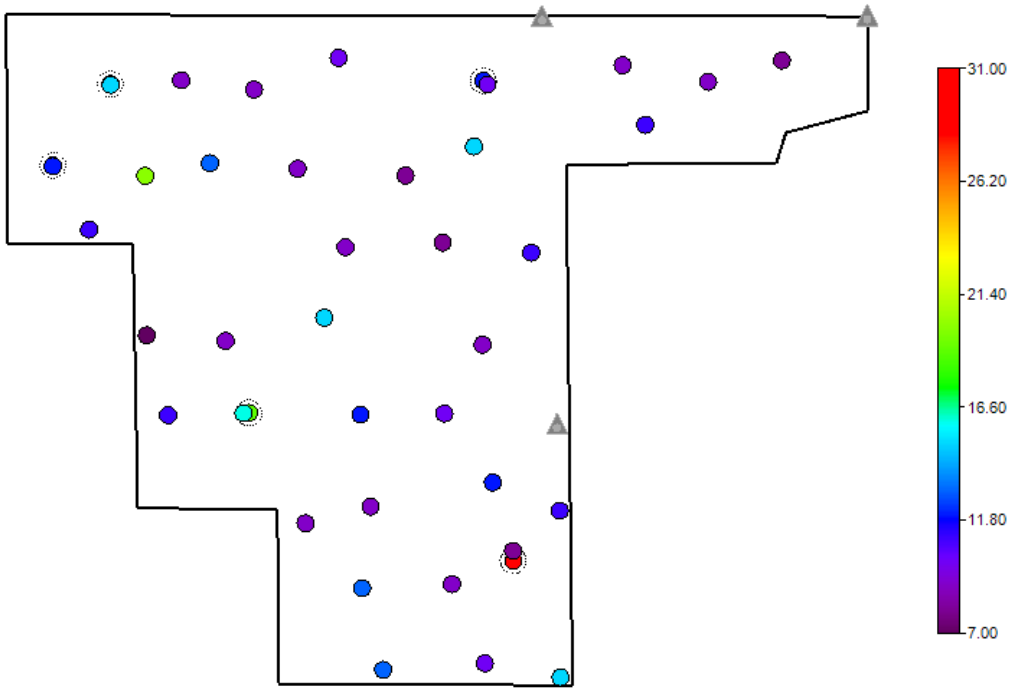


Figure 15. Three samples placed with Moran's I algorithm at a minimum distance of 250 meters.

The general trend is that as the minimum sample distance increases, samples diffuse into less densely sampled areas and towards edges and corners until the design can no longer place samples.

## VII. EXPLORING THE LOCAL INDEX OF SPATIAL ASSOCIATION PARAMETER

### A. **Background**

A LISA is any statistic that satisfies the following requirements:

- the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation, otherwise known as hot spots;
- the sum of LISAs for all observations is proportional (or equal) to a global indicator of spatial association.

A LISA can be expressed for a variable $y_i$, observed at location $i$, as a statistic $L_i$, such that:

$$L_i = f(y_i, y_j),$$

where $f$ is a function (possibly including additional parameters), and the $y_j$ are the values observed in the neighborhood $J_i$ of $i$ (Anselin, 1994).

In typical exploratory spatial data analyses, the predominant approach to assess the degree of spatial association is based on global statistics such as Moran's I or Geary's C, and it generally ignores local patterns of association (hot spots) and local instabilities in overall spatial association (Anselin, 1994). In SADA, by centering a LISA search window around each individual observation the weighted variance of the neighboring points within the search window are computed and samples are placed in areas of high local sample variance (Moran's I) or in areas of local negative correlation (Geary's C). Essentially, additional samples are placed in areas with the greatest sample variability in order to potentially test for local pockets, or hot spots. The search radius of the window determines the amount of neighbors factored into each individual observation and influences the placement of each additional sample.

For Ripley's K, which measures the average number of points within a certain distance of each other, the LISA search window dictates the number of neighbors factored into each individual observation and therefore determines if there is any clustering. Here clustering is referring to the sample density; it is independent of the concentration values. In contrast, the Moran's I and Geary's C methods are concentration dependent. The Ripley's K method then places additional samples in areas of lower sampling density. The radius of the search window will dictate how many neighbors are within a certain distance from each individual observation thereby affecting the program's definition of clustering and the location of additional samples.

## B. **Search Radius**

The search radius (i.e., the moving search window) is another parameter that the user must manually input. For those with no experience with the LISA tools there seems to be no intuitive radius that the user should implement. Different radii tend to create vastly different designs for all three LISA tools.

### 1. **Ripley's K**

This method is often employed in epidemiological studies to determine if there is any clustering in disease events. In SADA the method is used to determine if there is any clustering in the sample data (SADA, 2008). In different terms, the Ripley's K method is used to determine where areas of low sampling density are found. The nodes in these areas may be candidates for new sample locations. This is an improvement over the adaptive fill secondary sampling design which only considers gaps defined by nearest neighbors (SADA, 2008).

Figures 16 through 21 depict samples placed by the Ripley's K method at increasing LISA search windows.

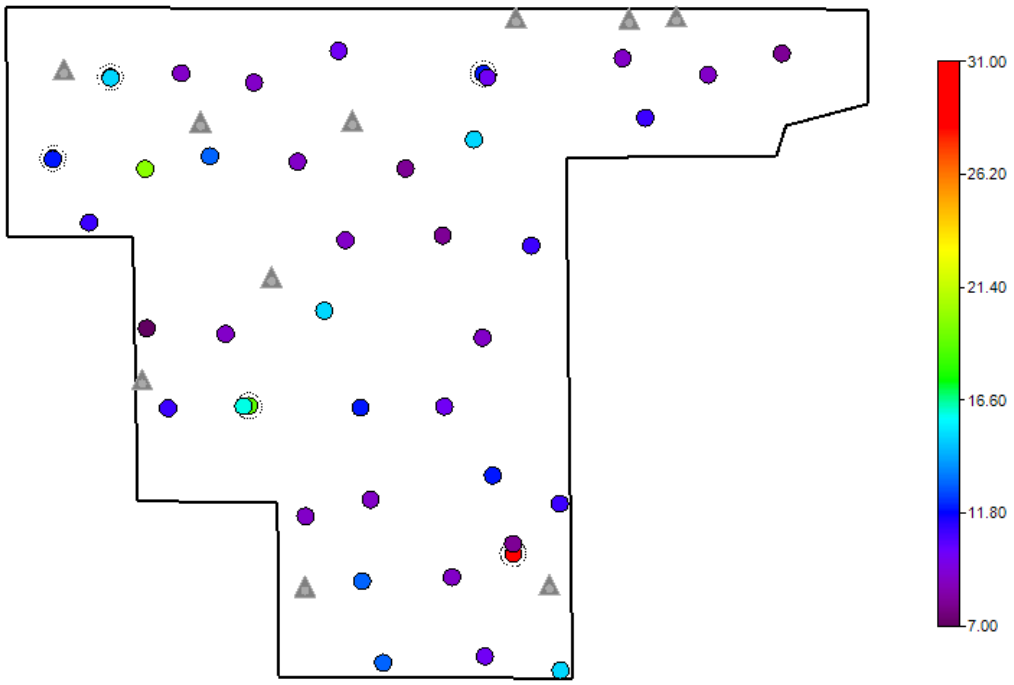


Figure 16.Ten samples placed using the Ripley's K method and a 100 meter LISA search radius.

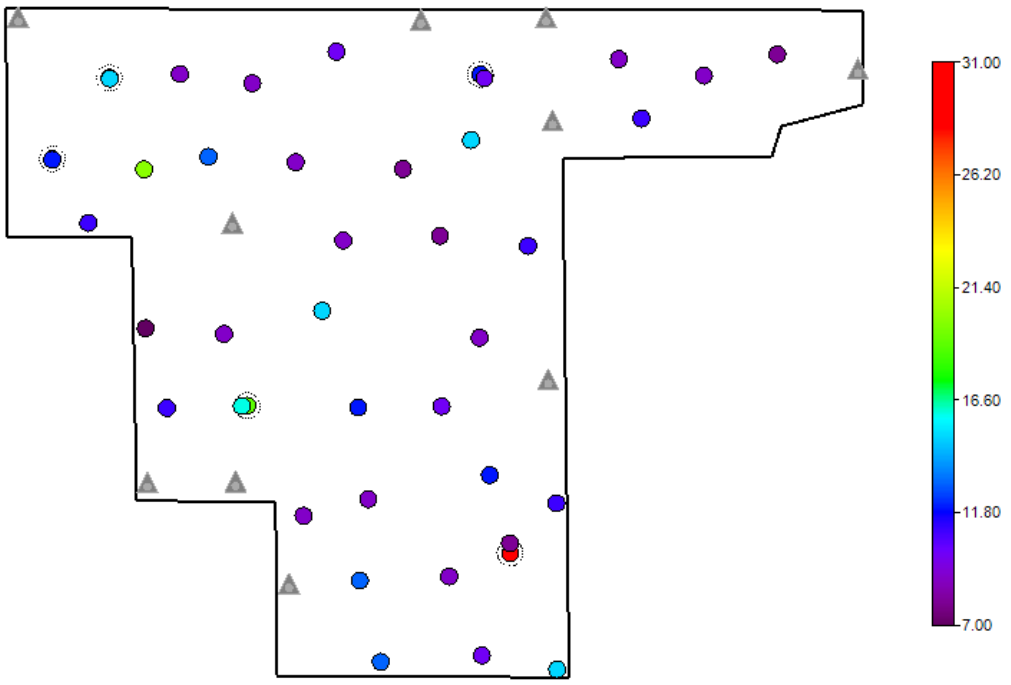


Figure 17. Ten samples placed using the Ripley's K method and a 200 meter LISA search radius.

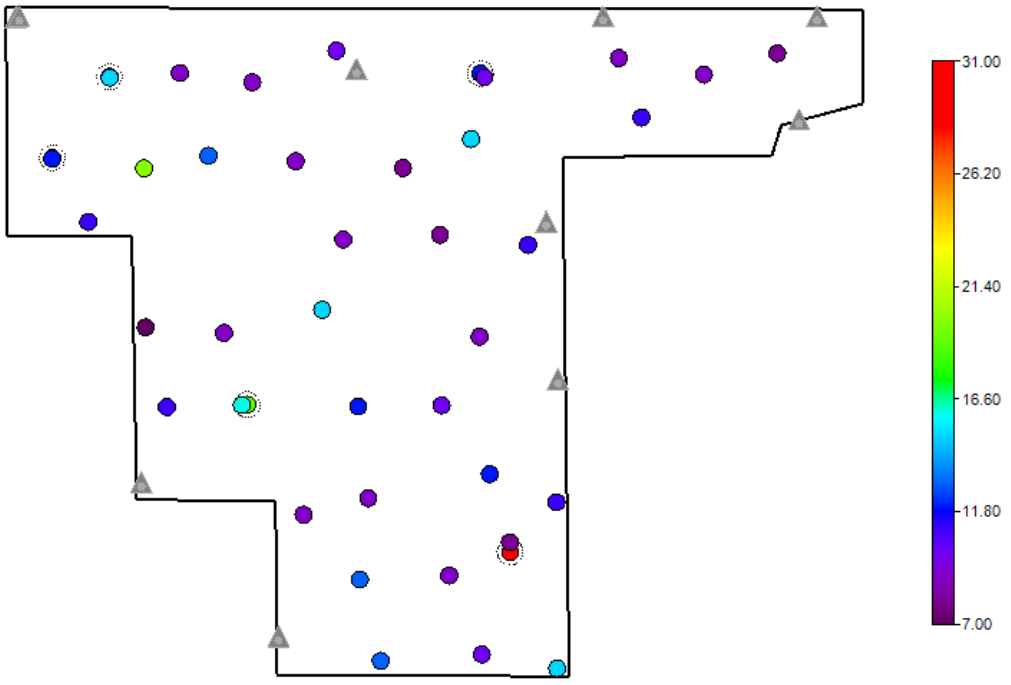


Figure 18. Ten samples placed using the Ripley's K method and a 300 meter LISA search radius.

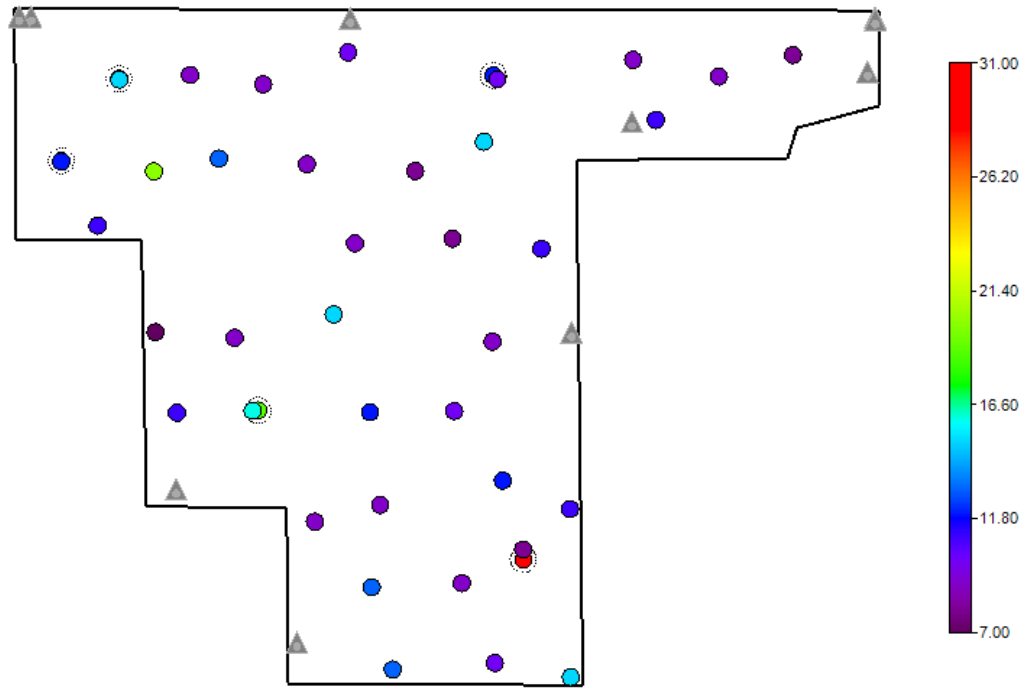


Figure 19. Ten samples placed using the Ripley's K method and a 400 meter LISA search radius.

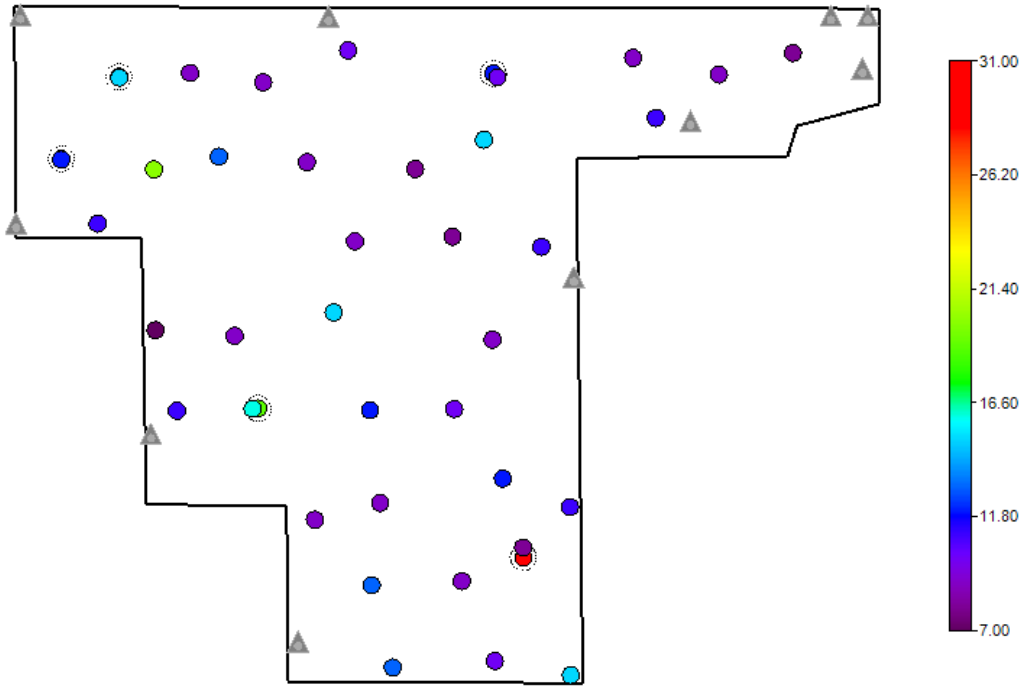


Figure 20. Ten samples placed using the Ripley's K method and a 500 meter LISA search radius.

As the LISA search radius increases for the Ripley's K method, the new sample points tend to be located away from the middle and other sample points. At about a 300 meter search radius and beyond (Figures 18 through 21), virtually all of the samples are located at the boundaries of the site. When sampling according to density these "edge effects" will occur because typically samplers will sample away from the boundaries, meaning that sampling is less dense on the edges of the site.

At a 500 meter search radius (Figure 20) four new sample points are located at the top right corner of the site, including three that cluster in the corner at the edges. Given the shape of the site it would stand to reason that as the search radius becomes larger and the definition of clustering changes, the top right corner is going to repeatedly become the area of lowest sampling density. This may cause some bias in these sampling designs. Indeed, at a LISA search radius of 1000 meters, all but one of the sample points is located at the top right corner, as seen in Figure 21.

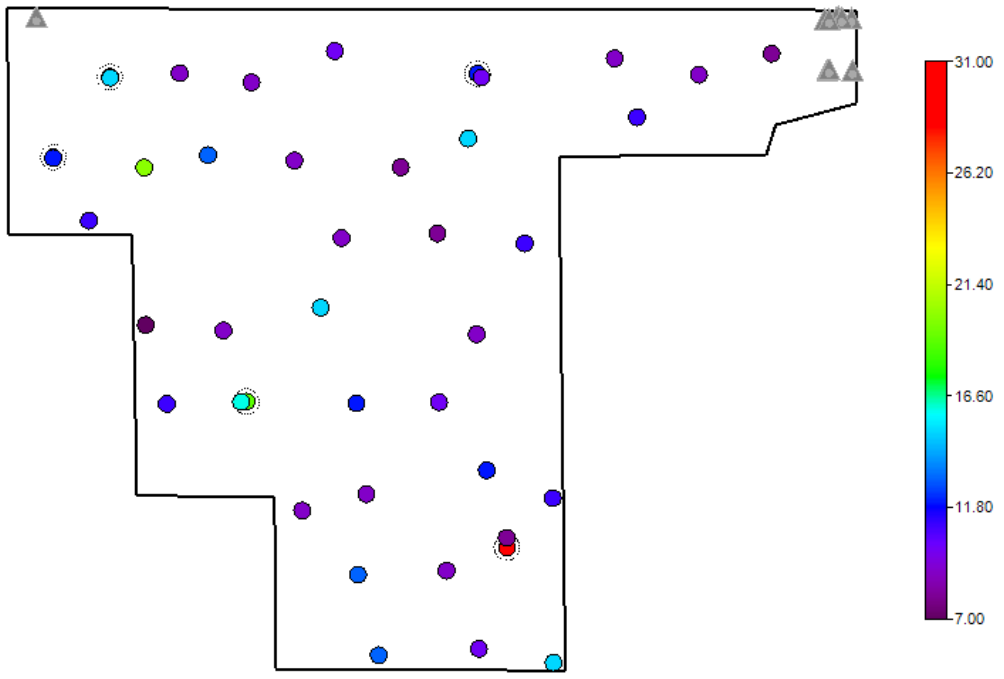


Figure 21. Ten samples placed using the Ripley's K method and a 1000 meter LISA search radius.

2. **Moran's I**

Moran's I is a measure of spatial autocorrelation. Autocorrelation is the concept that variables are spatially dependent. It implies that variables that are clustered together have more in common than variables that are farther apart. Typically Moran's I is used in exploratory spatial data analyses. Much like Ripley's K, the statistic is calculated by positioning a moving window of radius d at data points around the site. The weighted variance of data points within the window are then computed. New samples are placed in areas of high sample variance (SADA, 2008).

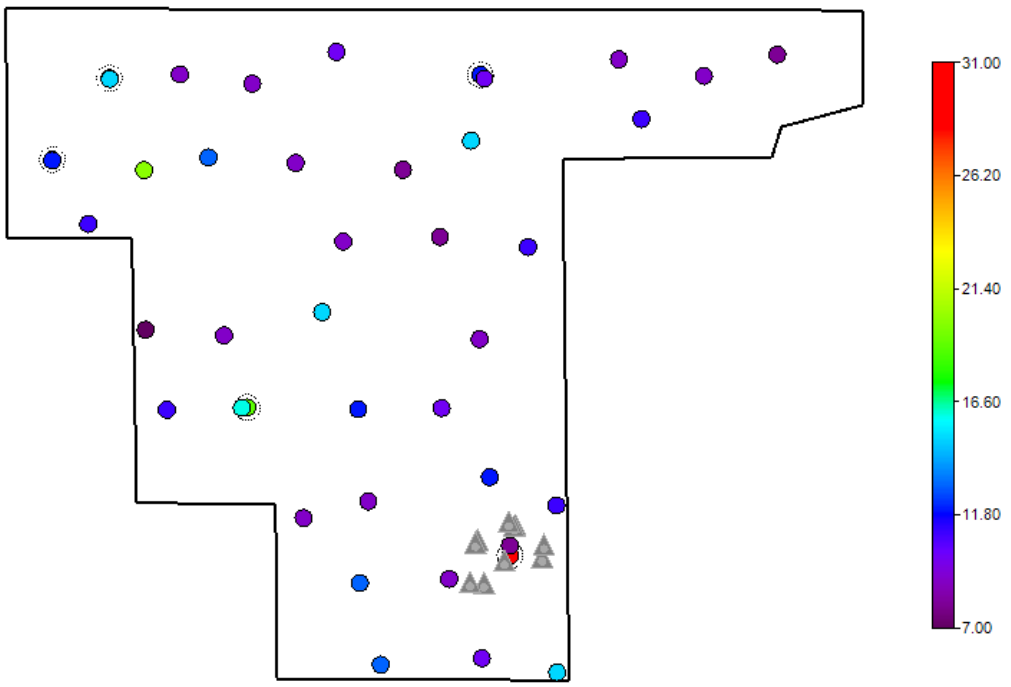

Figure 22. Ten samples placed using the Moran's I method and a 100 meter LISA search radius.

At smaller radii, such as 100 meters (Figure 22), the new samples cluster around the bottom right corner. This is due to a sample point with the highest concentration of the site (31 ppm) being surrounded by a number of samples with low concentrations. This creates an area of high sample variance. As the radius increases the potential samples radiate out from this location and also begin to be located in other areas of variability. Figures 23–25 show an increasing radius of 200, 300, and 400 meters, respectively.

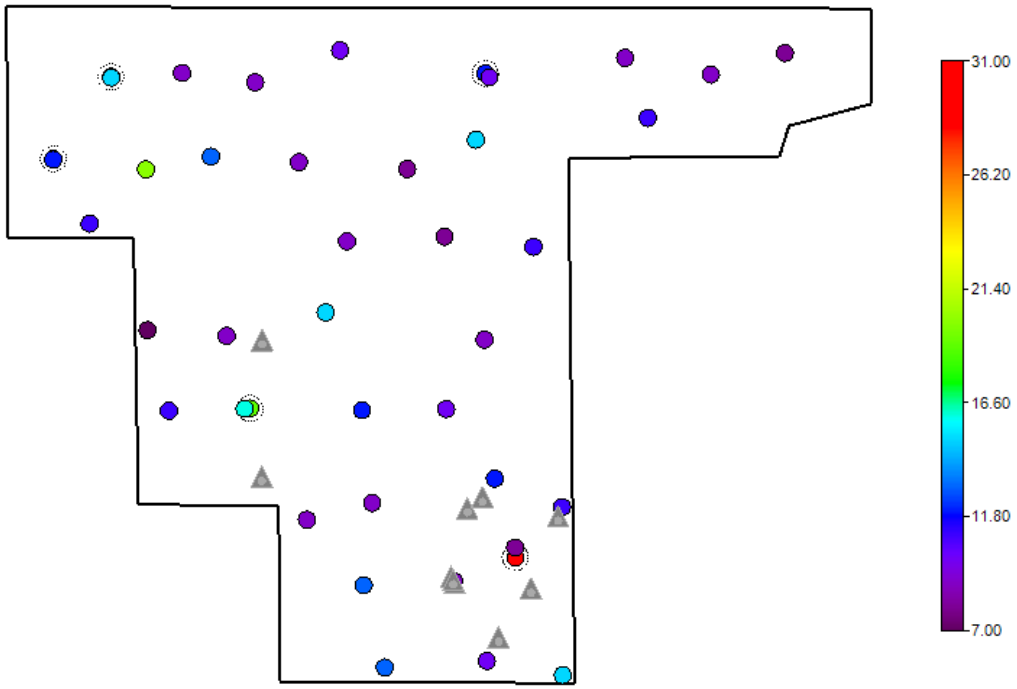


Figure 23. Ten samples placed using the Moran's I method and a 200 meter LISA search radius.

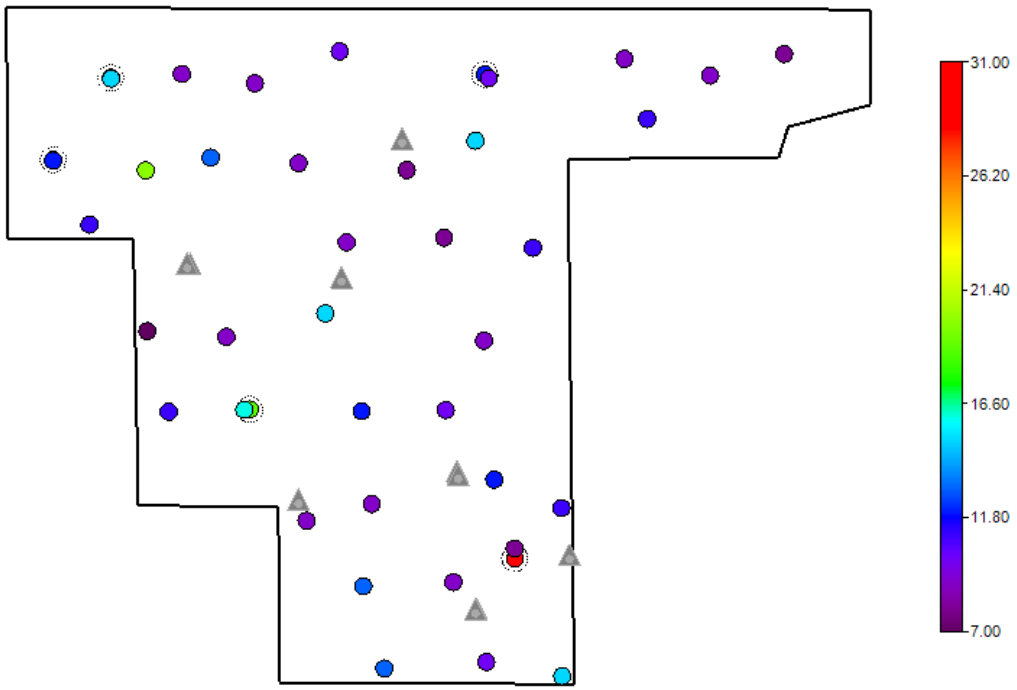


Figure 24. Ten samples placed using the Moran's I method and a 300 meter LISA search radius.

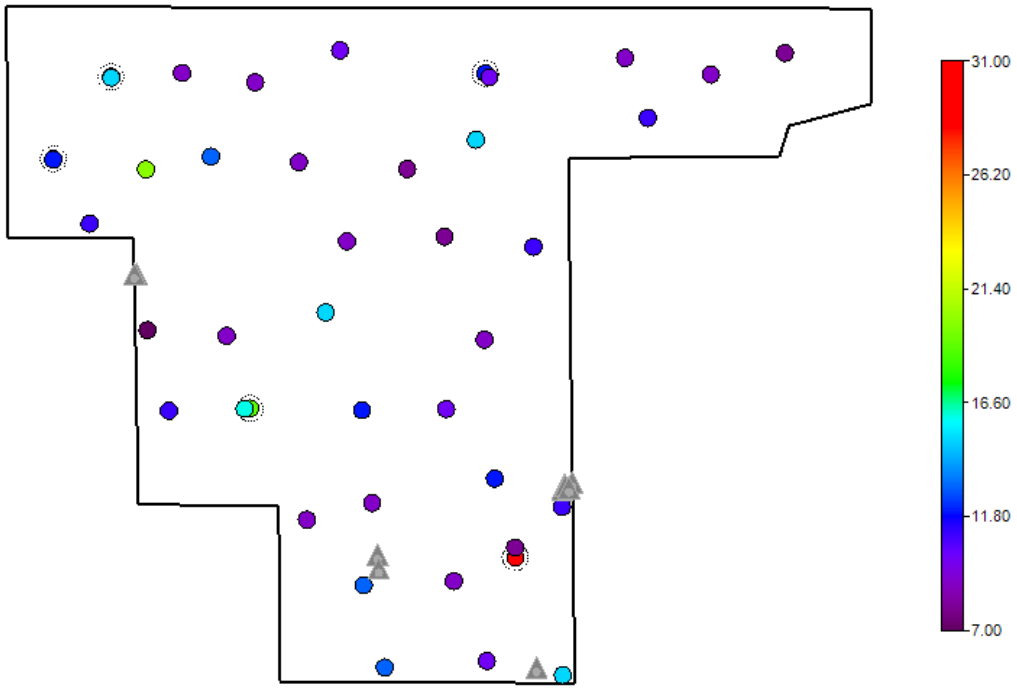


Figure 25. Ten samples placed using the Moran's I method and a 400 meter LISA search radius.

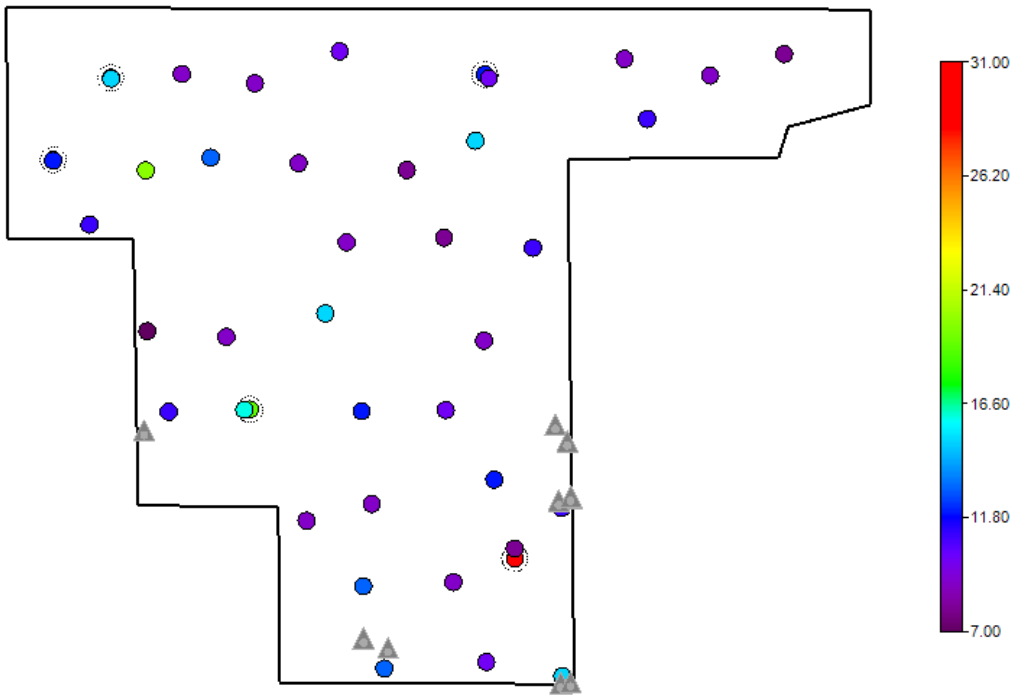


Figure 26. Ten samples placed using the Moran's I method and a 500 meter LISA search radius.

By 500 meters (Figure 26) edge effects start to become more pronounced as more potential sample points are located in the bottom right corner and along the boundaries of the site. At a search radius of 1000 meters (Figure 27) the samples are completely confined to the corners and boundaries of the site.

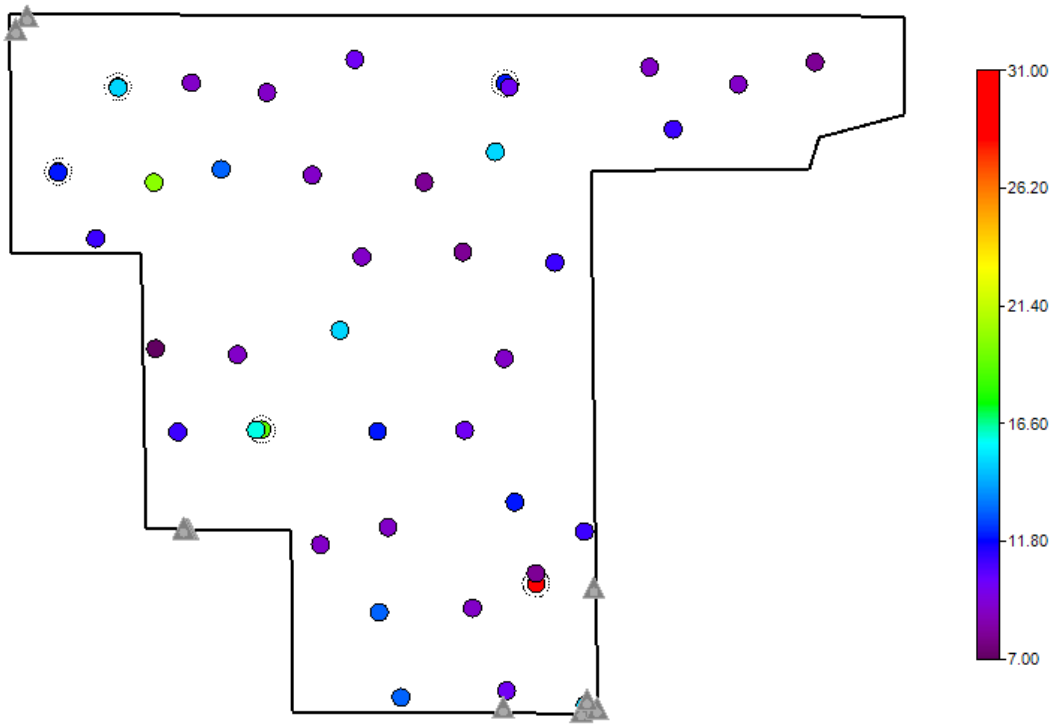


Figure 27. Ten samples placed using the Moran's I method and a 1000 meter LISA search radius.

3. **Geary's C**

Similar to Moran's I, Geary's C measures spatial autocorrelation. However, it does so in terms of correlation instead of variance. Geary's C is also more sensitive to local

spatial autocorrelation while Moran's I is more a measure of global spatial autocorrelation (SADA, 2008).

Also similar to Moran's I, the potential samples cluster around the bottom right area at lower radii distances due to the variability. As the radius increases this variability becomes less pronounced and the samples are spread more throughout the site. Figures 28–32 show an increasing radius of 100, 200, 300, 400, and 500 meters, respectively.

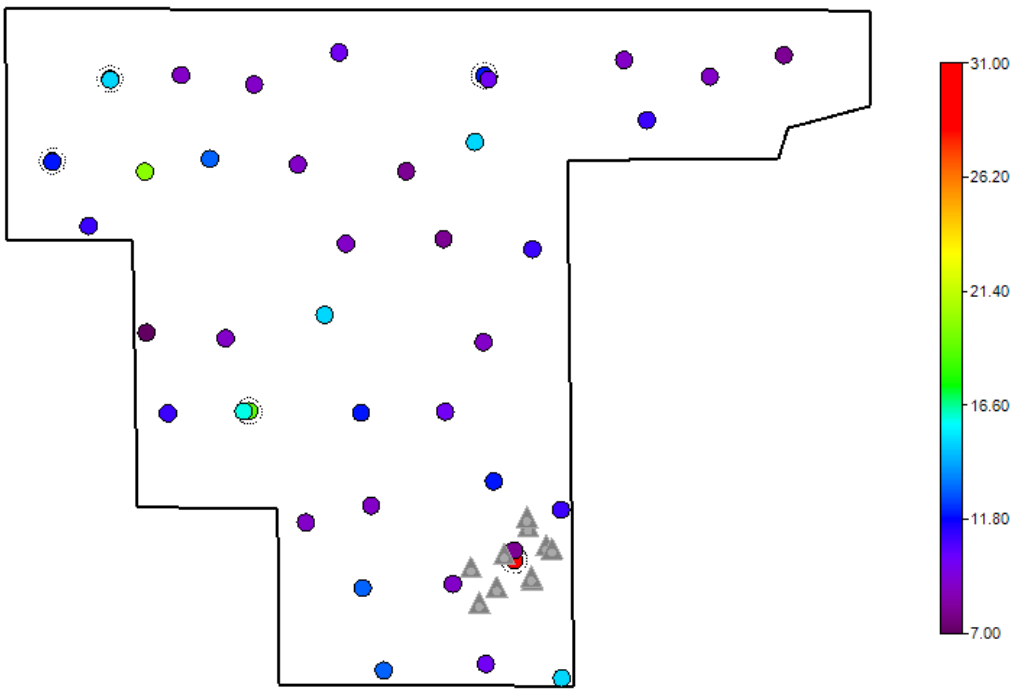


Figure 28. Ten samples placed using the Geary's C method and a 100 meter LISA search radius.

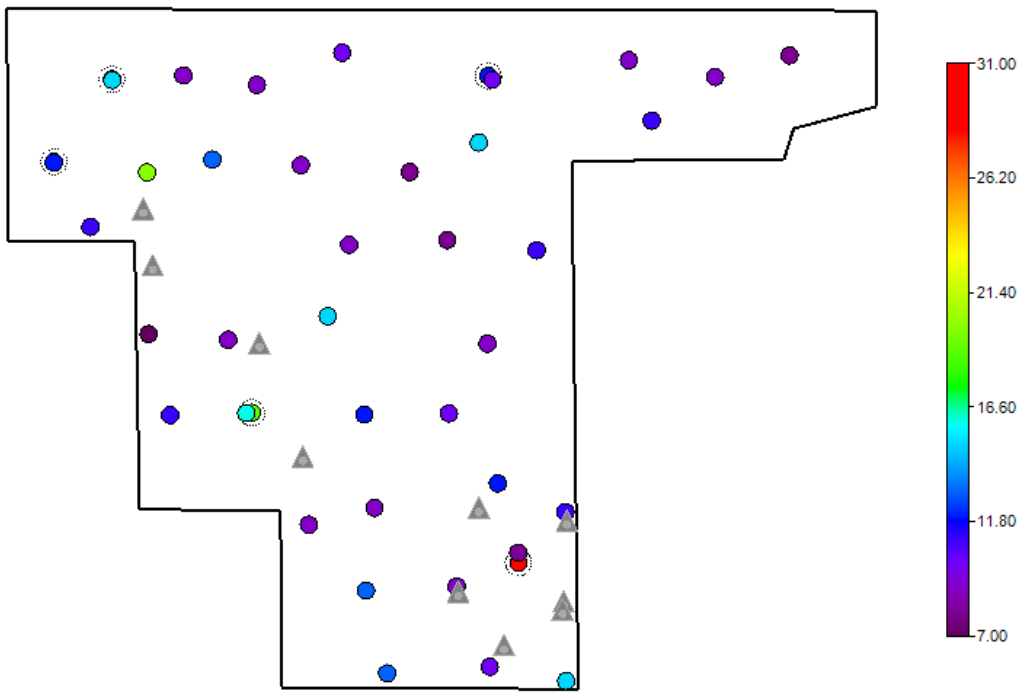


Figure 29. Ten samples placed using the Geary's C method and a 200 meter LISA search radius.

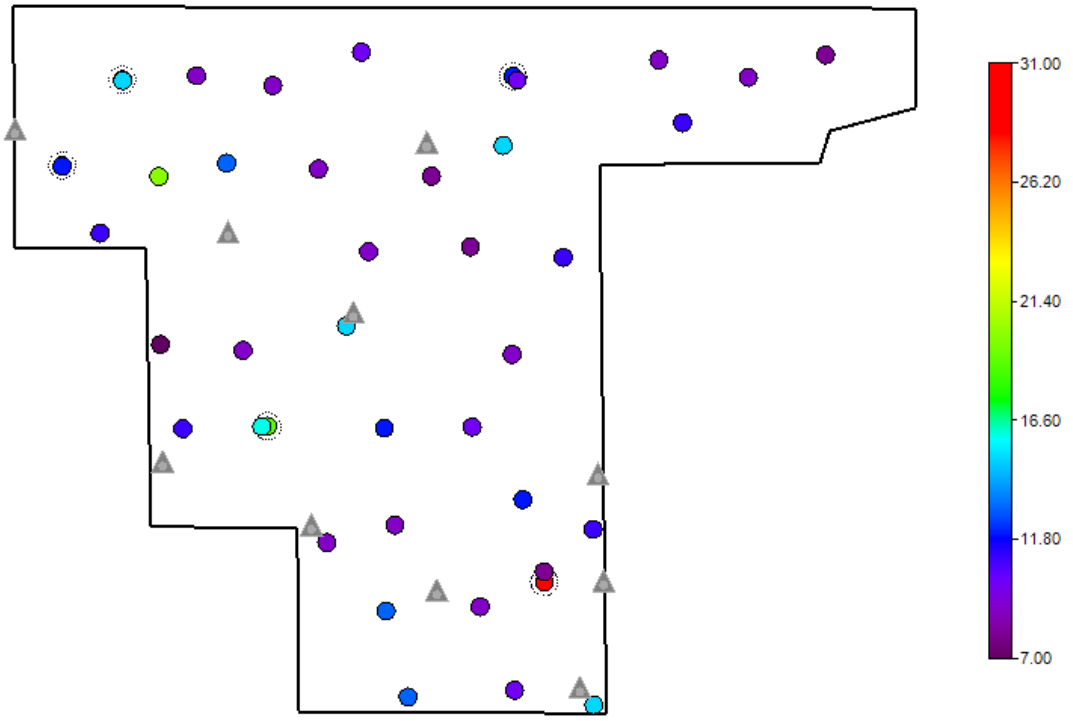


Figure 30. Ten samples placed using the Geary's C method and a 300 meter LISA search radius.

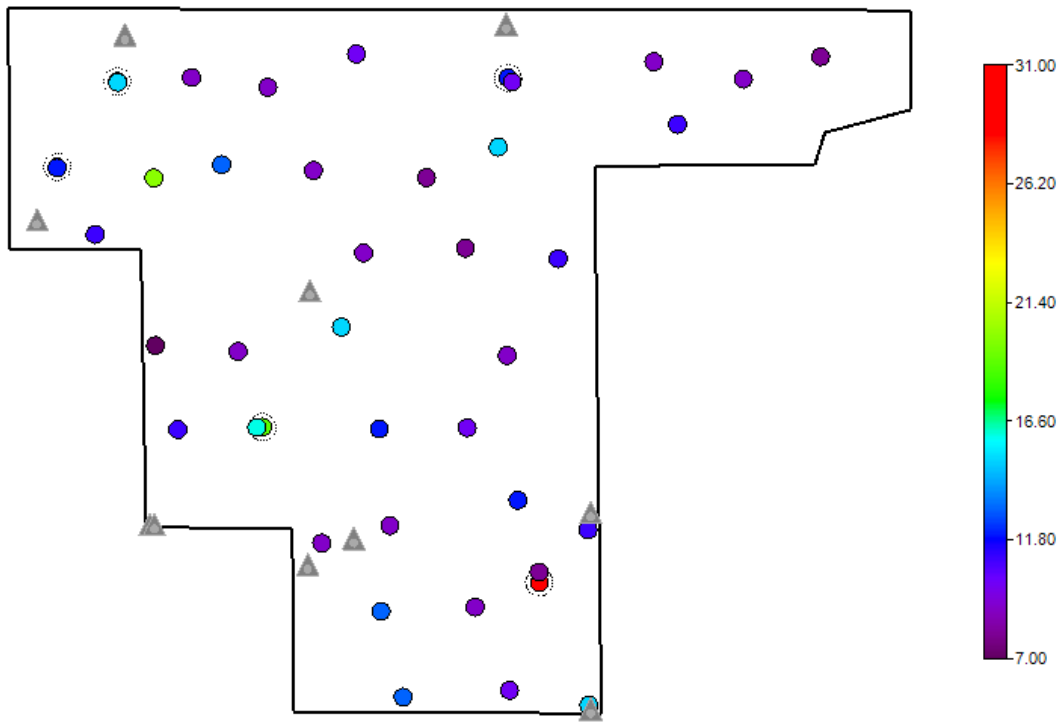


Figure 31. Ten samples placed using the Geary's C method and a 400 meter LISA search radius.

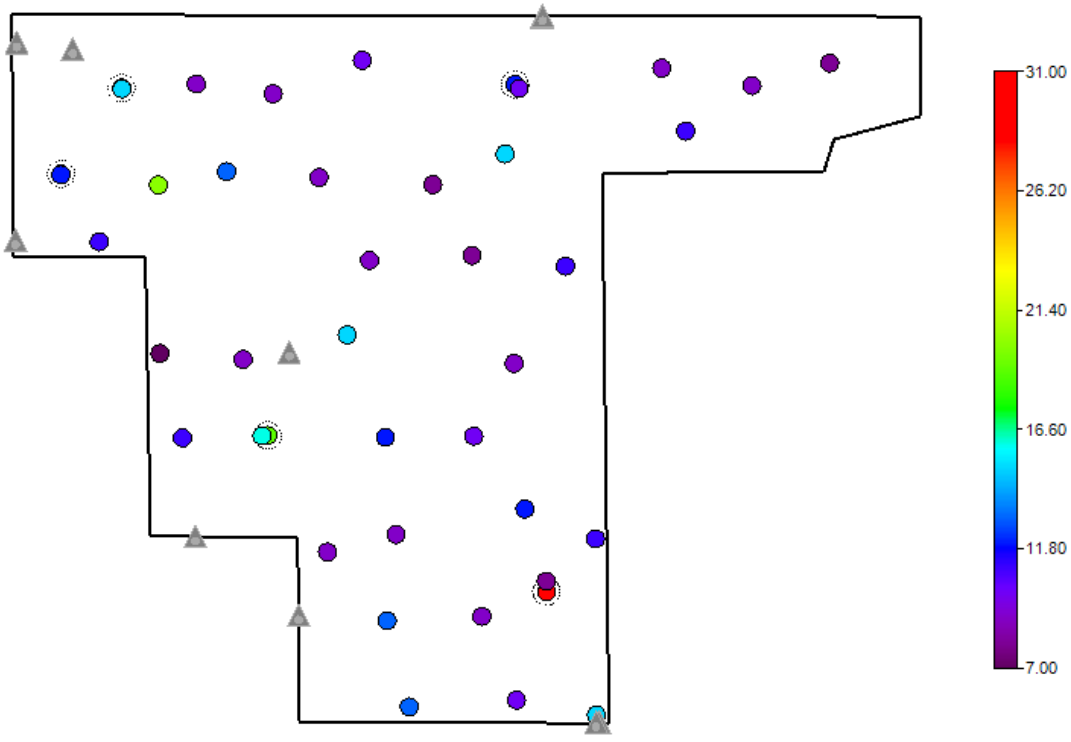


Figure 32. Ten samples placed using the Geary's C method and a 500 meter LISA search radius.

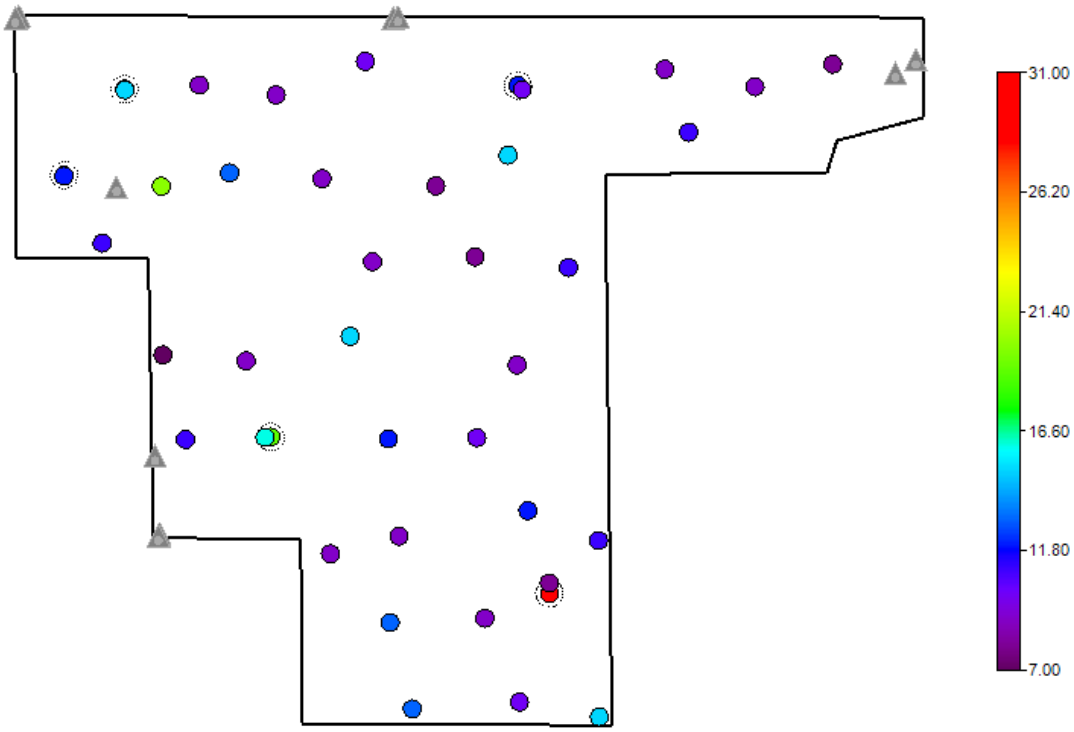


Figure 33. Ten samples placed using the Geary's C method and a 1000 meter LISA search radius.

Unsurprisingly, then, the samples begin to show edge effects and are placed in the corners and boundaries of the site. At a 1000 meter search radius (Figure 33) samples are no longer located in the bottom right area, but are dispersed to other corners of the site. This is not unexpected given the results for Ripley's K and Moran's I at this distance (Figures 21 and 27).

## C. **Conclusions**

Viewing these results it is evident that varying the LISA search radius gives equally variable secondary sampling designs. Unless the user has experience using LISA tools there is no intuitive or set distance to input. These distances and designs are not necessarily wrong, however, the fact that different distances create vastly different designs inherently creates a

problem and can easily lead to misuse. Radii that are too small might not take into account the full extent of spatial autocorrelation, and radii that are too large so that each potential sampling window starts to include most points in the site will give approximately the same variance calculation for all potential sites.

# VIII. IDENTIFYING GUIDELINES FOR THE LOCAL INDEX OF SPATIAL ASSOCIATION PARAMETER

## A. The Case For a Local Index of Spatial Association Distribution

Speaking with professionals familiar with geostatistics and LISA, including the author of the LISA module in SADA, revealed that there was no set search radius distance to use on a given data set. It was suggested by the author to use a distance between one-third and two-thirds of the longest transect distance of the site, and then to reduce the distance incrementally. According to a personal correspondance from the author, Tom Purucker of USEPA:

> The Moran's I secondary sampling design picks the window (with the given radius) centered on the grid node that has the highest sampling variability. So for a very small window (100 m) it selects areas near where the two points are most dissimilar and very close to each other. As you zoom out it picks up different points, and again selects windows where points are most dissimilar within the circular window centered on each potential sample location. However, once it gets very big, then each potential sampling window starts to include almost all the points, so that the variance calculation is practically the same for all potential sites, and only changes based upon the 1–2 samples that are not included in the prospective window. This ends up selecting locations that only have tiny increases in variance relative to other prospective sites, so is a bit random. That is why selection of the radius should not be greater than 2/3 of the longest transect of the site, to avoid results that are not really very meaningful in terms of having highest variance (or whatever) since they all end up being very similar (Purucker, 2012, personal email).

Based on this information the user then would have a potential ceiling radius distance at two-thirds of the longest transect of the site. Even at this distance it is unclear whether or not the variance calculation would be too homogenous for the data. Another problem occurs at the opposite end of the spectrum. Clearly, there will be a minimum distance, such as radii in the single digits, and potentially in the double digits, where not enough information is obtained and samples will not be placed. It is feasible to run the program multiple times at different radii to

pinpoint the distance where samples can be placed and call this the minimum radius distance, but will this be the most optimal design? Indeed, how does the user determine which design between the minimum and the feasible maximum radius distance is the most optimal?

A distribution of potential sample points comprised of potential sample points from this range of LISA radii distances could potentially reduce some of the uncertainty of the LISA search window issue and increase optimality. The user would be able to export these potential sample points from SADA for each search radius into a Geographic Information System, such as ArcMap, to visualize the full distribution. The user would then be able to choose potential sample points randomly to prevent bias and retain statistical significance.

## B. **Distribution Parameters**

The longest transect of Parcel 1 was detemined to be approximately 2500 meters. At the suggestion of beginning with a distance between one-third and two-thirds of the longest transect of the site it was decided to make the maximum radius distance of the search window one-half of the longest transect, which would be 1250 meters. The minimum search radius was set to 100 meters. Each secondary sampling design was run in SADA while incrementally decreasing the search distance radius by 50 meters. This provided, for each method, 24 secondary sample designs with 10 new samples per design, for a total distribution of 240 potential samples. Since contaminants can only be considered individually in SADA, this process was run for both arsenic and lead, for each LISA secondary sampling method.

## C. Results

The coordinates for these potential sample points were imported into ArcMap as point shapefiles. Each potential sample point contained the following information: Easting and Northing coordinates, LISA method used, search radius distance used, and contaminant based on (arsenic or lead). Twenty-four iterations of 10 samples, each at varying search radii, were run for the LISA methods. In the following sections, the first of the pair of figures for each contaminant reflects the distribution of 240 potential sample locations. The second of the pair of figures illustrates a random sample design of 10 samples pulled from each distribution.

### 1. Ripley's K

#### a. Arsenic

The Ripley's K method places potential samples in areas of lower sampling density. The distribution seen in Figure 34 reflects this as many potential sampling points cluster in the corners and on the boundaries of the site where sampling is minimal. There is a large concentration in the upper right corner of the site, which is unsurprising given the relative isolation of the four samples in this location.

#### b. Lead

The distribution (Figure 36) based on lead contamination is very similar to the arsenic distribution. Since the Ripley's K method is not based on variance measurements this is not surprising.

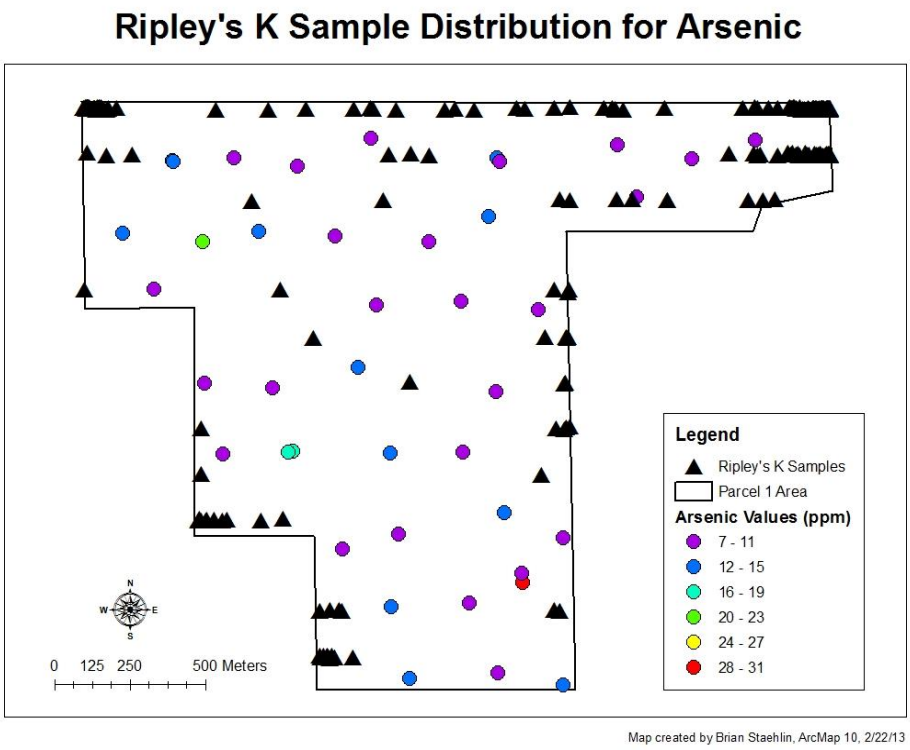


Figure 34. Distribution of potential sample points for arsenic using Ripley's K method.

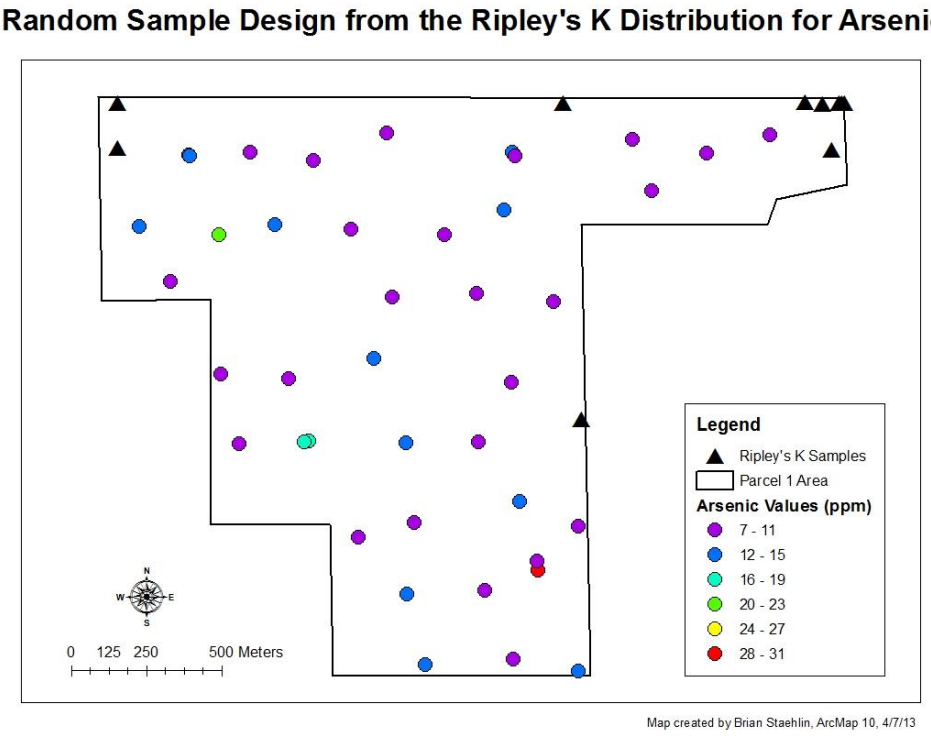


Figure 35. Ten samples chosen randomly from the Ripley's K distribution of potential samples for arsenic.

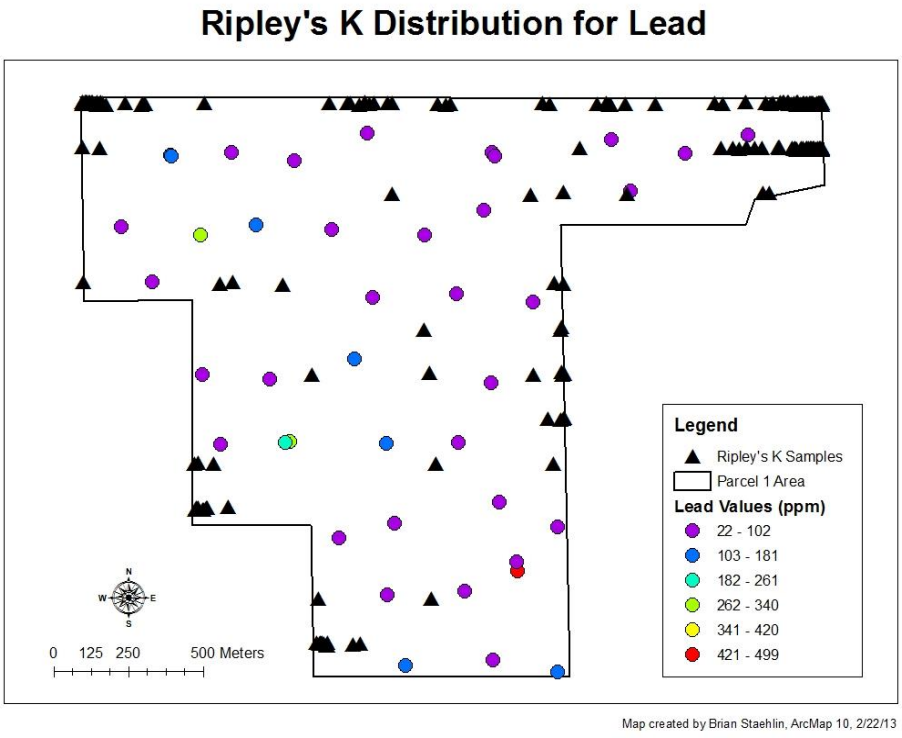


Figure 36. Distribution of potential sample points for lead using Ripley's K method.

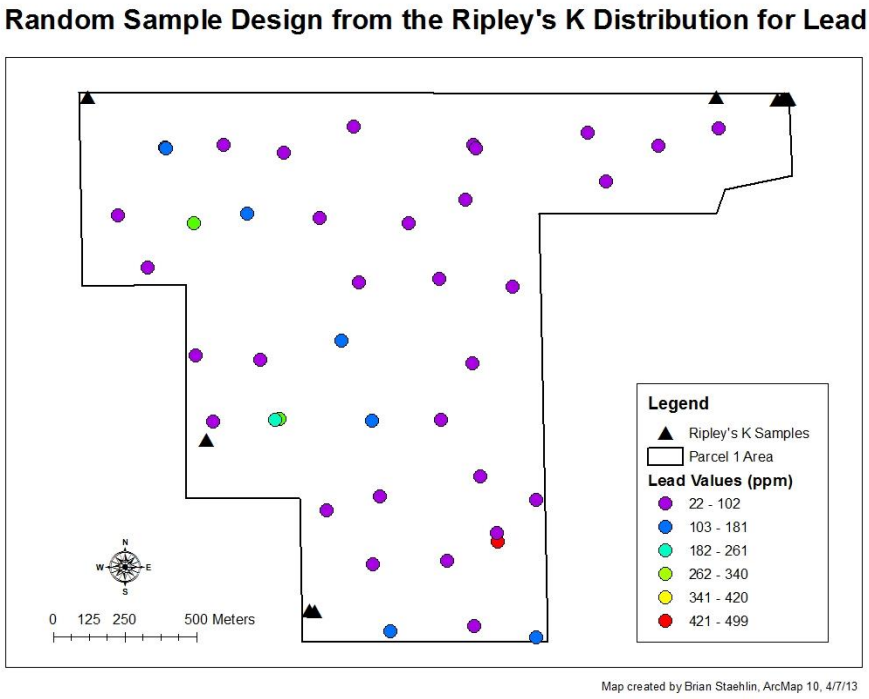


Figure 37. Ten samples chosen randomly from the Ripley's K distribution of potential samples for lead.

2. **Moran's I**

a. **Arsenic**

The Moran's I method places potential samples in areas of high local sample variance. The majority of the distribution focuses on the high variance created by the 31 ppm arsenic sample, which is surrounded by samples of much lower concentration in the southeastern corner of the site. There are a few clusters around samples with slightly elevated concentrations in the western and northwestern parts of the site. Samples that cluster in the corners or on boundaries mostly occurred from the search windows with larger radii. This can be seen in Figures 38 and 39.

b. **Lead**

Much like the arsenic distribution, the lead distribution focuses chiefly around the highest lead concentration of the site, 499 ppm. The two distributions are very similar spatially. This seems to be a testament to the spatial correlation between arsenic and lead sample concentrations. Higher arsenic and lead concentrations are found together, while lower arsenic and lead concentrations are found together. This can be seen in Figures 40 and 41.

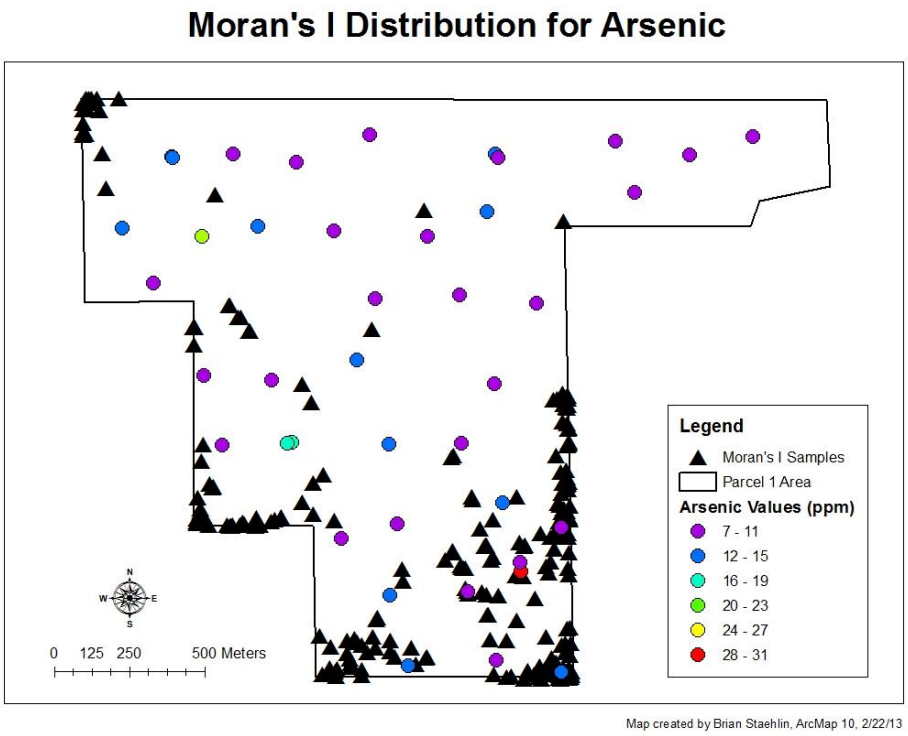


Figure 38. Distribution of potential sample points for arsenic using Moran's I method.

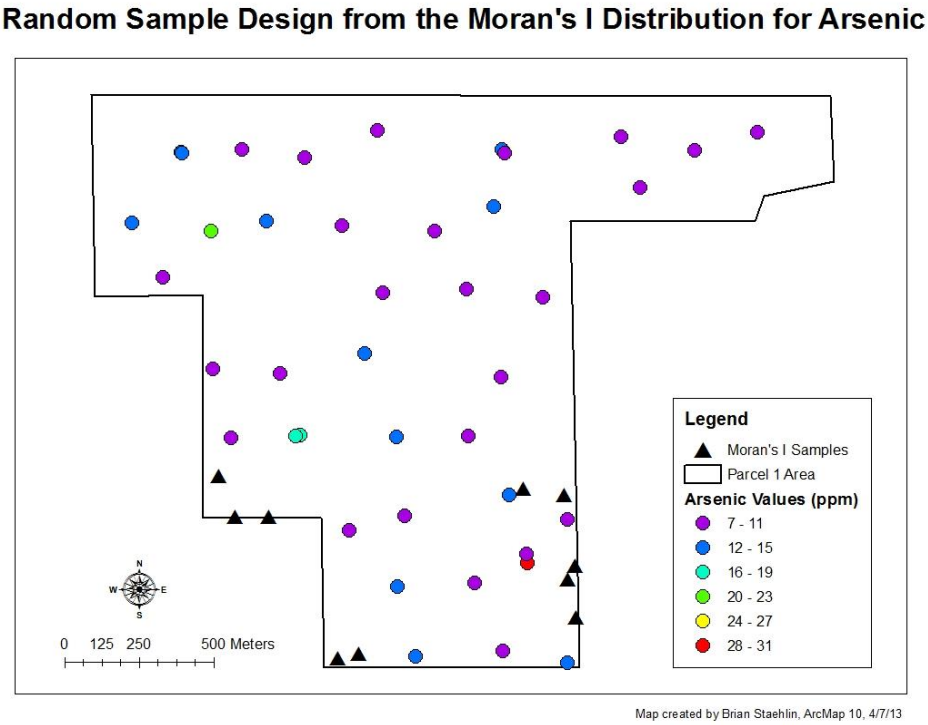


Figure 39. Ten samples chosen randomly from the Moran's I distribution of potential samples for arsenic.

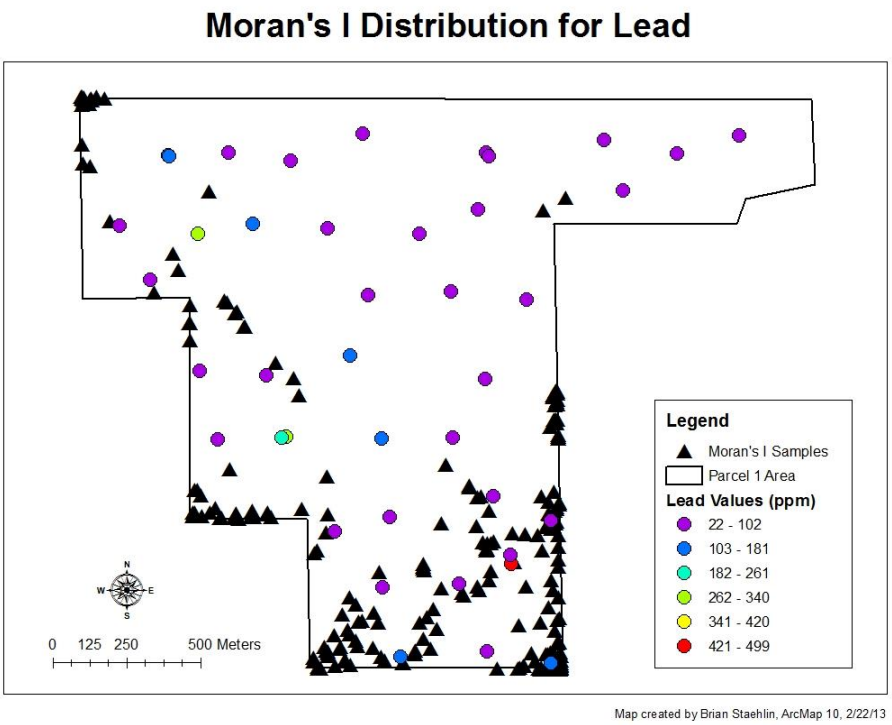


Figure 40. Distribution of potential sample points for lead using Moran's I method.

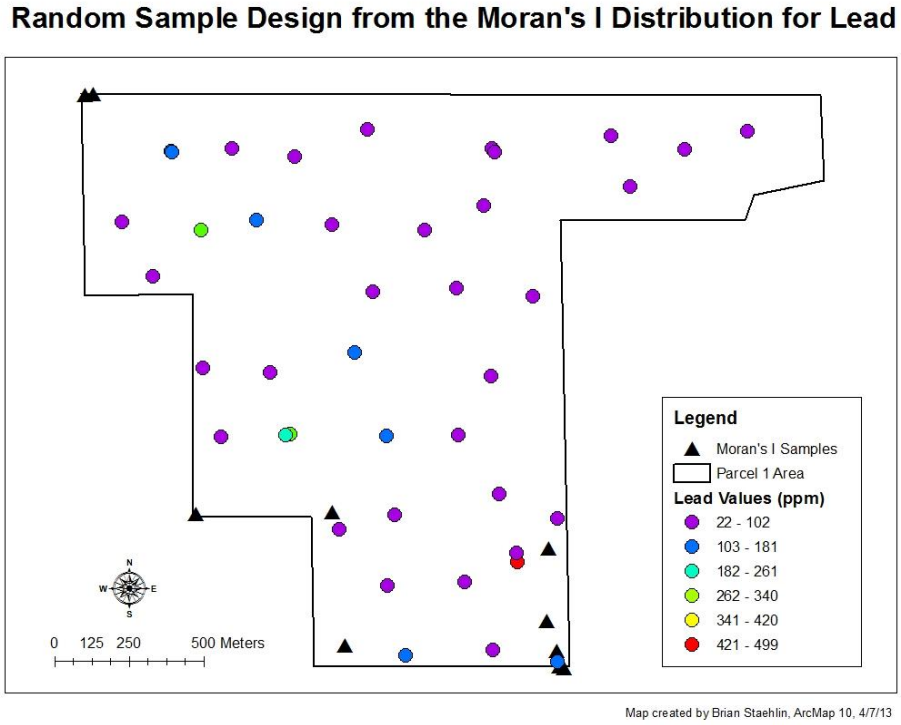


Figure 41. Ten samples chosen randomly from the Moran's I distribution of potential samples for lead.

### 3. Geary's C

#### a. Arsenic

The Geary's C method places potential samples in areas of local negative correlation. While there are similarities to the Moran's I method, such as many samples clustering around the high concentrations in the south, west, and northwest of the site, the Geary's C method seems to disperse the potential samples more evenly throughout these areas as well as into other areas of the site. The 33 sample points in the northeasternmost corner of the site contain potential samples points with LISA search windows as low as 600 meters, but mostly consist of sample points with LISA search windows in the 1000–1250 meters range. This would seem to be a case where once a potential sampling window gets very large and starts to include almost all the points, the correlation calculation is practically the same for all potential sites, and only changes based upon the 1–2 samples that are not included in the prospective window. This ends up selecting locations that only have tiny increases in negative correlation relative to other prospective sites, and these locations may tend to be located on the edges of the site.

#### b. Lead

Figures 42, 43, 44, and 45 show that once again the lead distribution mimics the arsenic distribution in overall trends.

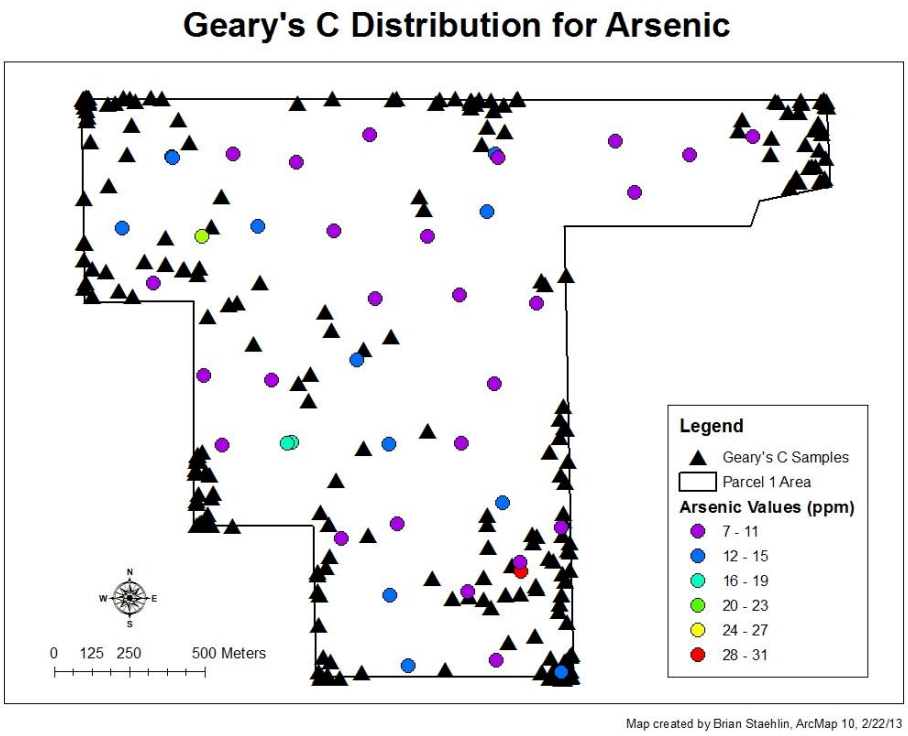


Figure 42. Distribution of potential sample points for arsenic using Geary's C method.

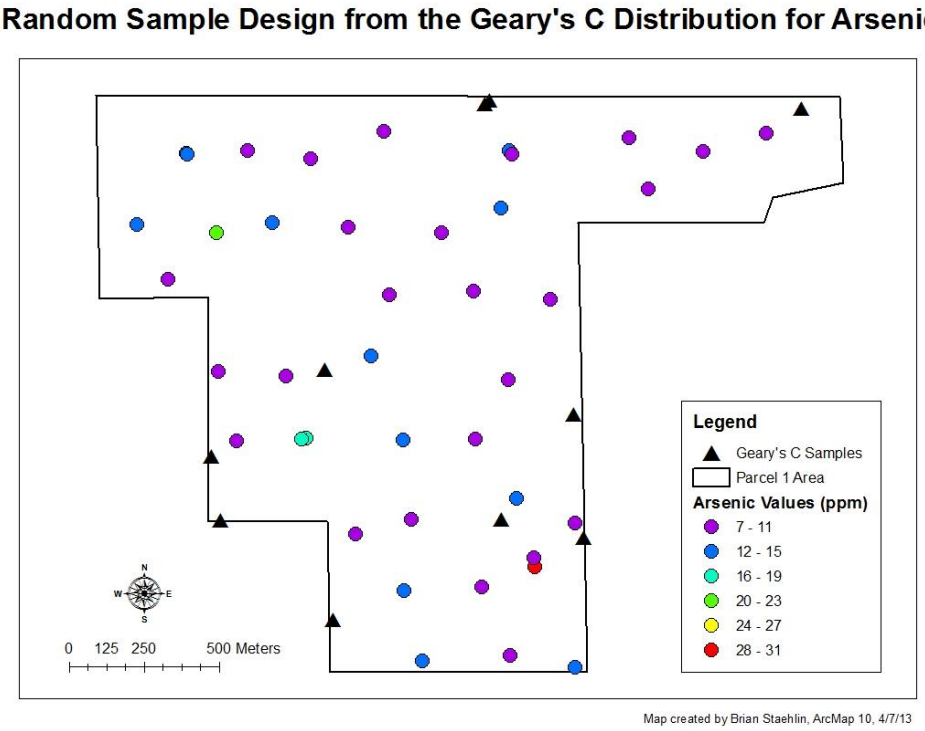


Figure 43. Ten samples chosen randomly from the Geary's C distribution of potential samples for arsenic.

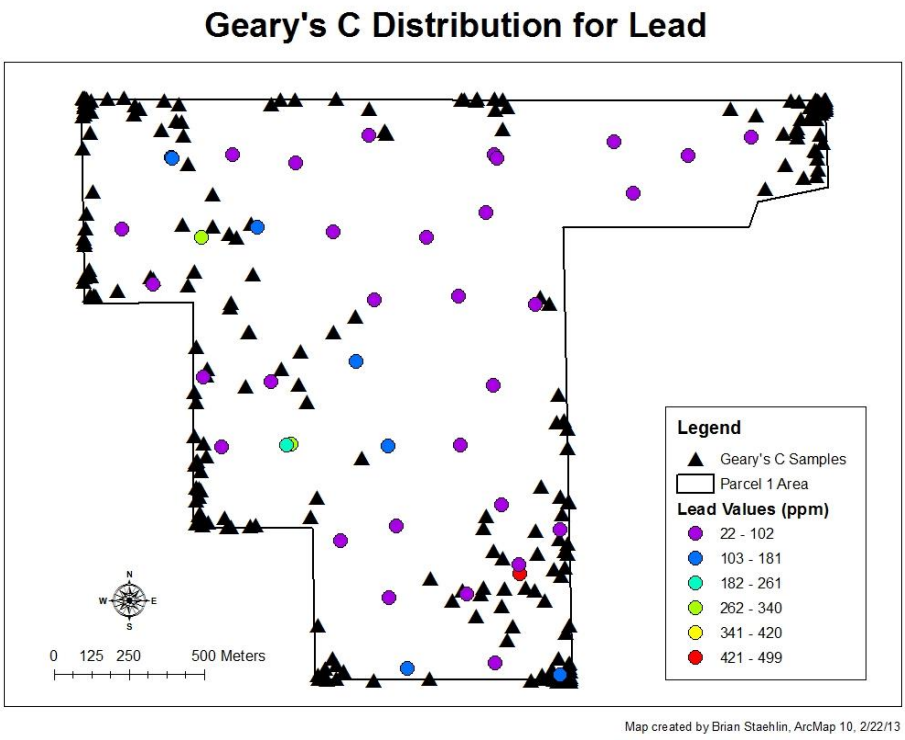


Figure 44. Distribution of potential sample points for lead using Geary's C method.

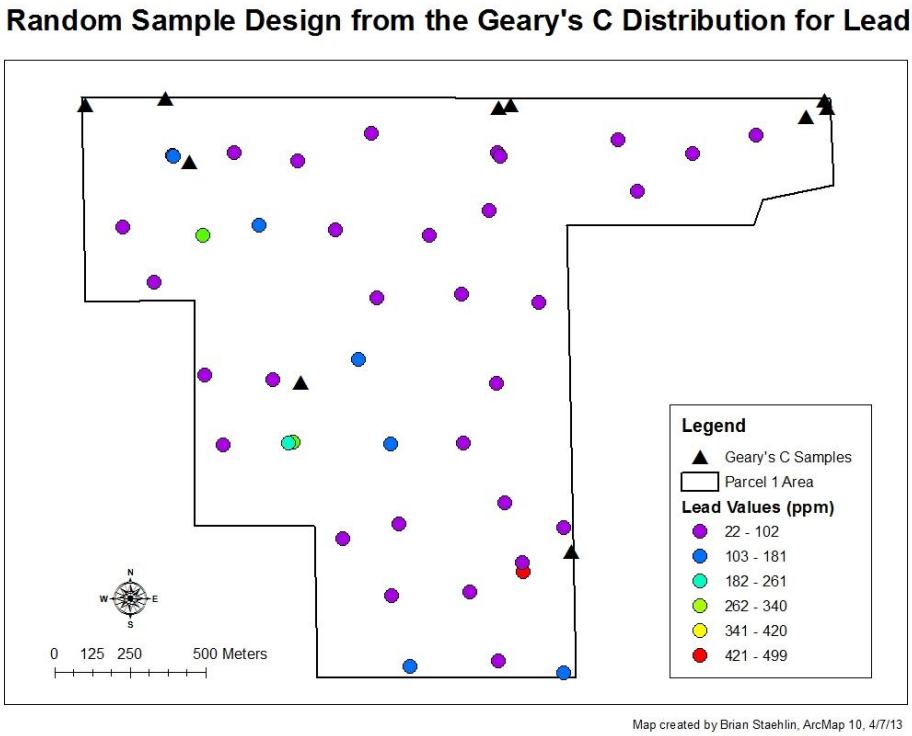


Figure 45. Ten samples chosen randomly from the Geary's C distribution of potential samples for lead.

# IX. DISCUSSION

There are several issues regarding secondary sample design in SADA. The bulk of them deal with parameters left to the discretion of the user: number of samples placed, tie-break options, sample separation distance, grid specifications, and the LISA search radius. In addition, the sample distributions exhibit a degree of edge effects and redundancies. Although manual user input allows the tool a high amount of flexibility, this also creates a situation of uncertainty and potential misuse within the remediation process. While most of these parameters can logically be modeled on real world variables or constraints, the LISA search radius is not intuitive and different distances can vastly affect the outcome of the secondary sampling design. The goal of this study was to take some of the uncertainty out of the LISA algorithm by using a wide range of search radii to create a potential sample distribution; one that the user could use as a pool of potential sample points.

Some uncertainty inevitably remains. The decision to use the range and interval performed in the study is not without question. Though this study provides a substantial amount of potential sample points optimizing the interval would be a logical next step. It is reasonable to expect that this study did not completely optimize the potential sample point distribution—it is likely that increasing the interval one meter at a time would be time consuming and redundant, but a 50-meter interval may not have captured the most optimal distribution; however it is arguable that it is a good basis for future studies.

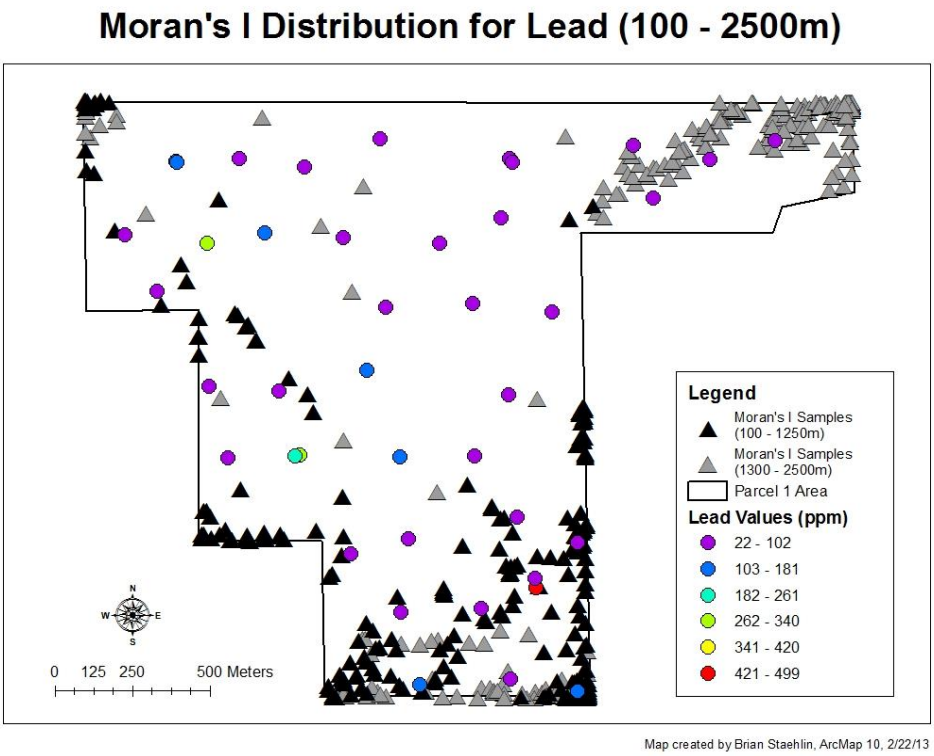


Figure 46. Distribution of potential sample points for lead using Moran's I method at a LISA search window range from 100 meters to 2500 meters in intervals of 50 meters.

An argument can be made, however, for the range, as well as the maximum search radius distance. To explore what occurs when a search distance beyond one-half of the larger transect is used search distances were run in 50 meter intervals from 1300 meters to 2500 meters (the distance of the longest transect of the site), for the Moran's I method using lead contamination. As shown in Figure 46, in addition to being placed in the southern portion of the site, those potential sample points that are placed using a search radius beyond one-half of the longest transect of the site are mostly focused in a cluster in the northeast corner of the site. This is curious behavior given how the Moran's I method operated at lower radii. There is also more dispersion throughout the site in general. This can be explained by only showing the secondary sample design that used a 2500 meter search window, as shown in Figure 47. This dispersive,

seemingly random behavior is atypical to the Moran's I method. The potential explanation for these phenomena once again has to do with variance measurements that include all or practically all of the sample points in the site. In the case of search windows at distances between 1300 meters and 2450 meters, this serves to progressively push potential sample points into the

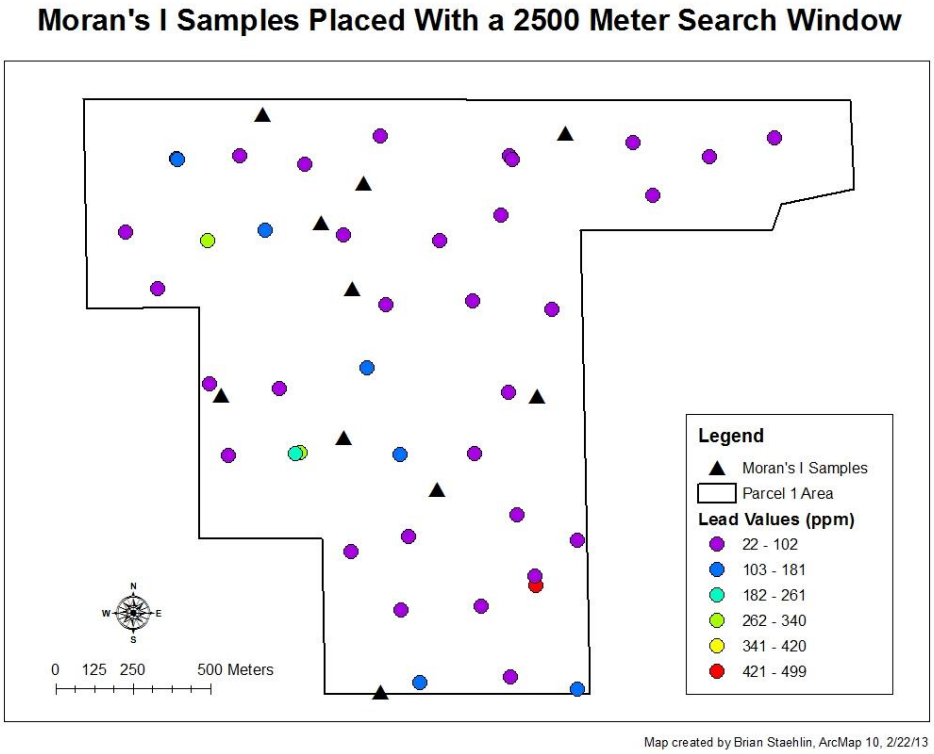


Figure 47. Ten samples placed using the Moran's I method and a 2500 meter LISA search radius.

northeast corner of the site. This could be influenced by the geography of the northeastern section of the site, the variance measurements due to the lead concentrations present in this section of the site, or a combinatory effect of the two. In the case of the 2500 meter search radius, using a search window as large as the longest transect of the site seems to return a

random output and should be discouraged when developing secondary sampling designs such as these.

When a user draws a random sample from these distributions it is worth noting their trends. As seen in Figures 34 and 36, using Ripley's K, many potential sample points in the distribution cluster in the northeast corner of the site due to the properties of the statistic and the nature of the site. Likewise, both random samples (Figures 35 and 37) contain a number of samples in the northeast corner. On the other hand, even though the distributions for arsenic and lead from the Geary's C method (Figures 42 and 44) are very similar (due to contaminant correlation), the random sample designs are quite different, as shown in Figures 43 and 45. There is nothing inherently wrong with this, given the nature of random sampling, but it does emphasize the need for samplers/users to be knowledgable on the various LISA functions and uses and have a clearly defined goal when developing a secondary sample design.

The redundencies and edge effects present in the potential site sample distributions are exhibited in the random samples. The SADA software program should include a methodology for choosing samples from the identified site sample distributions. Ideally this methodology would reduce the redundancies that occur within the site sample distributions and that subsequently occur in the final secondary sample designs. One potential idea is to incorporate a weighting methodology, with samples along the boundaries of the site being weighted to reduce their probability of being chosen. Another method could potentially modify the program so that samples located in boundaries and corners of the site would not be considered in the final secondary sample designs. Perhaps this could be employed by incorporating a polygon at the borders of the site and the potential sample points located within this polygon would be identified as having a zero probability of being contaminated.

Ultimately, the hope is that the developers of SADA may be able to use this study to develop an algorithm that uses site parameters to create a distribution of potential sample points with a range of LISA radii, as done here. The upper bound of this range could be found via iterations, the inflection point being when larger ranges are no longer providing differences in variance.

From this distribution, SADA could randomly, or via another methodology, choose the number of potential sample points requested by the user while also accounting for redundancies in sampling and edge effects. This could streamline the tool further, while alleviating some of the uncertainty inherent in the LISA methods.

# X. CONCLUSION

Current regulations for remediation do not factor in to processes that account for the spatial distribution of contaminants. Spatially defined information that takes into account data gained from previous sample studies would allow site investigators to visualize the extent of the contamination and minimize uncertainty while providing accurate results to reduce costs during data collection and remediation. A competent tool with a comprehensive and cost-effective approach to developing sampling designs, SADA has the ability to be utilized in a range of uses, including Phase III environmental site assessments (ESAs), brownfield redevelopment, or other environmental risk management or site remediation situations.

Secondary sampling designs based on historical investigation data can be created by SADA by utilizing model- (geospatial) based sampling designs. The program is able to mimic real world constraints and parameters and deliver potential secondary sampling designs with minimal costs. Secondary sampling is often performed to define remediation areas and clean areas. However, there are no clearly defined methods to determine both the number and location of secondary samples. The LISA methods as implemented using the SADA software offer a methodology for the placement of these secondary samples. A major drawback of these methods is the uncertainty involved in the range of input parameters. This thesis offers an objective means to reduce the uncertainty inherent in these parameters by employing an iterative function.

The LISA search window greatly affects the outcome of the secondary sample designs. The LISA methods, however, if developed using effective guidelines such as those presented

here, represent powerful tools to be used during the risk assessment and clean-up process. A methodology is recommended to reduce the redundancies that occur within the site sample distribution and that subsequently occur in the secondary site sample design.

## APPENDIX

## DETERMINING A THRESHOLD DECISION CRITERIA

The threshold decision criteria will be based on soil screening levels (SSLs) developed by USEPA. To compute the arsenic screening level, USEPA's Regional Screening Level Calculator for Chemical Contaminants at Superfund Sites (http://epa-prgs.ornl.gov/cgi-bin/chemicals/csl_search) will be used to generate generic SSLs for the chronic ingestion of arsenic in a residential scenario. Soil screening levels are concentrations of contaminants in soil that are designed to be protective of exposure in a residential setting. A residential scenario is used given the site's proximity to a residential area (Osiecki, 2011). Generic SSLs will be used instead of site-specific SSLs (equations given by USEPA) because site-specific SSLs require soil properties, which are not available (USEPA, 1996). Two SSLs will be generated for arsenic from the RSL calculator—one for the carcinogenic endpoint and one for the non-carcinogenic endpoint.

### TABLE II

DEFAULT VALUES FOR ARSENIC RESIDENT EQUATION INPUTS FOR SOIL

| Variable | Value |
|---|---|
| TR (target cancer risk) unitless | 0.000001 |
| $ED_r$ (exposure duration - resident) year | 30 |
| $ET_{rs}$ (exposure time - resident) hour | 24 |
| $ED_c$ (exposure duration - child) year | 6 |
| $ED_a$ (exposure duration - adult) year | 24 |
| $BW_a$ (body weight - adult) kg | 70 |
| $BW_c$ (body weight - child) kg | 15 |
| $SA_a$ (skin surface area - adult) $cm^2$/day | 5700 |
| $SA_c$ (skin surface area - child) $cm^2$/day | 2800 |
| THQ (target hazard quotient) unitless | 1 |
| LT (lifetime - resident) year | 70 |

## TABLE II

DEFAULT VALUES FOR ARSENIC RESIDENT EQUATION INPUTS FOR SOIL

| Variable | Value |
|---|---|
| $IRS_a$ (soil intake rate - adult) mg/day | 100 |
| $IRS_c$ (soil intake rate - child) mg/day | 200 |
| $AF_a$ (skin adherence factor - adult) mg/cm$^2$ | 0.07 |
| $AF_c$ (skin adherence factor - child) mg/cm$^2$ | 0.2 |
| $IFS_{adj}$ (age-adjusted soil ingestion factor) mg-year/kg-day | 114 |
| $DFS_{adj}$ (age-adjusted soil dermal factor) mg-year/kg-day | 361 |
| $IFSM_{adj}$ (mutagenic age-adjusted soil ingestion factor) mg-year/kg-day | 489.5 |
| $DFSM_{adj}$ (mutagenic age-adjusted soil dermal factor) mg-year/kg-day | 1445 |
| $ED_{0-2}$ (exposure duration first phase) year | 2 |
| $ED_{2-6}$ (exposure duration second phase) year | 4 |
| $ED_{6-16}$ (exposure duration third phase) year | 10 |
| $ED_{16-30}$ (exposure duration fourth phase) year | 14 |
| City (Climate Zone) PEF Selection | Default |
| $A_s$ (acres) PEF Selection | 0.5 |
| $Q/C_{wp}$ (g/m$^2$-s per kg/m$^3$) PEF Selection | 93.77 |
| PEF (particulate emission factor) m$^3$/kg | 1359344438 |
| A (PEF Dispersion Constant) | 16.2302 |
| B (PEF Dispersion Constant) | 18.7762 |
| C (PEF Dispersion Constant) | 216.108 |
| V (fraction of vegetative cover) unitless | 0.5 |
| $U_m$ (mean annual wind speed) m/s | 4.69 |
| $U_t$ (equivalent threshold value) | 11.32 |
| F(x) (function dependant on $U_m/U_t$) unitless | 0.194 |
| City (Climate Zone) VF Selection | Default |
| $A_s$ (acres) VF Selection | 0.5 |
| $Q/C_{wp}$ (g/m$^2$-s per kg/m$^3$) VF Selection | 68.18 |
| foc (fraction organic carbon in soil) g/g | 0.006 |
| ρ$_b$ (dry soil bulk density) g/cm$^3$ | 1.5 |
| ρ$_s$ (soil particle density) g/cm$^3$ | 2.65 |
| θ$_w$ (water-filled soil porosity) $L_{water}/L_{soil}$ | 0.15 |
| T (exposure interval) s | 950000000 |
| A (VF Dispersion Constant) | 11.911 |
| B (VF Dispersion Constant) | 18.4385 |
| C (VF Dispersion Constant) | 209.7845 |

Retrieved from - http://epa-prgs.ornl.gov/cgi-bin/chemicals/csl_search (June, 2012)

The regional screening level (RSL) calculator and the SSL documentation does not include SSLs for lead "because EPA has issued separate documents that specify risk-based concentrations for these contaminants in soil." (USEPA, 2002, A-3). Since the SSLs are based in part on reference dose (RfD), and the USEPA has no consensus for an inorganic lead RfD, USEPA evaluates lead exposure based on blood-lead modeling, using the Integrated Exposure-Uptake Biokinetic Model. For the purpose of screening 400 ppm is recommended for residential soils. Caution should be used, however, when both water and soil are being assessed in a risk assessment, as the combination can result in a higher blood-lead level if these values aren't conservative enough (Osiecki, 2011). This is beyond the scope of this paper, however.

Once the parameters are input into the RSL calculator for arsenic (chronic ingestion of soil in a residential scenario), the calculator provides both the assumptions and values used in the equation, and the resident risk-based SLs for soil.

**TABLE III**

RESIDENT RISK-BASED ARSENIC SCREENING LEVEL FOR SOIL

| Chemical | CAS Number | Ingestion SF (mg/kg-day)$^{-1}$ | SFO Ref | Inhalation Unit Risk (ug/m$^3$)$^{-1}$ | IUR Ref | Chronic RfD (mg/kg-day) | RfD Ref | Chronic RfC (mg/m$^3$) | RfC Ref |
|---|---|---|---|---|---|---|---|---|---|
| Arsenic, Inorganic | 7440-38-2 | 1.50E+00 | I | 4.30E-03 | I | 3.00E-04 | I | 1.50E-05 | C |

| GIABS | ABS | Volatilization Factor (m$^3$/kg) | Soil Saturation Concentration (mg/kg) | Particulate Emission Factor (m$^3$/kg) | Ingestion SL TR=1.0E-6 (mg/kg) | Dermal SL TR=1.0E-6 (mg/kg) | Inhalation SL TR=1.0E-6 (mg/kg) |
|---|---|---|---|---|---|---|---|
| 1 | 0.03 | - | - | 1.36E+09 | 4.27E-01 | 4.49E+00 | 7.69E+02 |

| Carcinogenic SL TR=1.0E-6 (mg/kg) | Ingestion SL HQ=1 (mg/kg) | Dermal SL HQ=1 (mg/kg) | Inhalation SL HQ=1 (mg/kg) | Noncarcinogenic SL HI=1 (mg/kg) | Screening Level (mg/kg) |
|---|---|---|---|---|---|
| 3.90E-01 | 2.35E+01 | 2.79E+02 | 2.13E+04 | 2.16E+01 | 3.90E-01 ca[a] |

Retrieved from - http://epa-prgs.ornl.gov/cgi-bin/chemicals/csl_search (June, 2012)

[a]ca=Cancer, nc=Noncancer (Where nc SL < 100 x ca SL)

The screening level recommended by the RSL calculator is 0.4 ppm which is the carcinogenic screening level (SL). This is due to the non-carcinogenic SL being less than 100 times the carcinogenic SL. The noncarcinogenic SL is based on a hazard quotient of 1, and the carcinogenic SL is based on a total risk equal of 1.0E-6 (a risk of acquiring cancer of 1 in 1 million). The problem with this conservative value of 0.4 ppm is that naturally occurring arsenic

exceeds this level in all parts of the country (Osiecki, 2011). For the purpose of the following scenarios the total non-carcinogenic SL of 22 ppm, which is based on the following equation, will be used:

$$SL_{res\text{-}sol\text{-}nc\text{-}tot}\,(mg/kg) = \frac{1}{\frac{1}{SL_{res\text{-}sol\text{-}nc\text{-}ing}} + \frac{1}{SL_{res\text{-}sol\text{-}nc\text{-}der}} + \frac{1}{SL_{res\text{-}sol\text{-}nc\text{-}inh}}}$$

Figure 48. The residential soil land use equation, containing the ingestion, dermal, and inhalation exposure routes for noncarcinogenic ingestion. Retrieved from - http://www.epa.gov/reg3hwmd/risk/human/rb-concentration_table/usersguide.htm (June, 2012)

It is worth noting that the direct ingestion of soils is one of the most common routes of human exposure to contaminants in the residential setting, and inorganic forms of metals are not likely to cross biological membranes as easily, which supports ingestion exposure over dermal absorption (USEPA, 1996). However, arsenic is a soil contaminant that USEPA evaluates for dermal exposures. In addition, USEPA considers ingestion/dermal absorption SSLs for most metals to be adequately protective of fugitive dust exposures SSLs for the residential scenario (USEPA, 2002). While this evidence points to using the non-carcinogenic ingestion SL of 24 ppm, or perhaps a combined ingestion/dermal absorption SL, the more conservative total non-carcinogenic SL of 22 ppm, which combines all the applicable exposure routes, will be used.

## CITED LITERATURE


Anselin, L. 1994. "Local Indicators of Spatial Association—LISA." Research paper 9331. West Virginia University, West Virginia.

Barnes, R. 1989. "Sampling Design for Geologic Site Characterization." *Geostatistics* 2:809–822.

Barth, D. S., and B. J. Mason. 1984. "Soil Sampling Quality Assurance and the Importance of an Exploratory Study." In *Environmental Sampling for Hazardous Wastes*. 97–104. Edited by G. Schweitzer et al. ACS Symposium Series Washington, DC: American Chemical Society.

Brus, D. 2010. "Design-Based and Model-Based Sampling strategies for Soil Monitoring." Soil Science Centre, Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands.

Brus, D. J., and J. J. de Gruijter. 1997. "Random Sampling or Geostatistical Modelling? Choosing Between Design-Based and Model-Based Sampling Strategies for Soil (with Discussion)." *Geoderma* 80:1–44.

Cochran, W. G. 1977. *Sampling Techniques*. New York: John Wiley & Sons.

Cox, D. D., L. H. Cox, and K. B. Ensor. 1995. "Spatial Sampling and the Environment." *Technical Report Number 38*. National Institute of Statistical Sciences, North Carolina.

De Gruijter, J. J., and C. J. F. ter Braak. 1990. "Model-Free Estimation From Spatial Samples: A Reappraisal of Classical Sampling Theory." *Mathematical Geology* 22:407–415.

De Gruijter, J. J., and C. J. F. ter Braak. 1992. "Design-Based Versus Model-Based Sampling Strategies: Comment on R.J. Barnes' "Bounding the Required Sample Size for Geologic Site Characterization." *Mathematical Geology* 24:859–864.

Englund, E. J., and N. Heravi. 1992. "Conditional Simulation: Practical Application for Sampling Design Optimization." In *Geostatistics Troia '92*. 613–624. Edited by A. Soares. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Englund, E. J., and N. Heravi. 1994. "Phased Sampling for Soil Remediation." *Environmental and Ecological Statistics* 1:247–263.

Gilbert, R. O., and B. A. Pulsipher. 2005. "Role of Sampling Designs in Obtaining Representative Data." *Environmental Forensics* 6:27–33.

Holmes, J. H. 1970. "The Theory of Plane Sampling and Its Applications in Geographic Research." *Economic Geography* 46:379–392.

Isaaks, E. H., and R. M. Srivastava. 1989. *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

Johnson, R. L. 1996. "A Bayesian/Geostatistical Approach to the Design of Adaptive Sampling Programs." *Geostatistics for Environmental and Geotechnical Applications* 1283:102–116.

Kravchenko, A., and D. G. Bullock. 1999. "A Comparative Study of Interpolation Methods for Mapping Soil Properties." *Agron. J* 91:393–400.

McArthur, R. D. 1987. "An Evaluation of Sample Designs for Estimating a Locally Concentrated Pollutant." *Communications In Statistics—Simulation and Computation* 16:735–759.

McBratney, A. B., R. Webster, and T. M. Burgess. "The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables—I, Theory and Method." *Computers & Geosciences* 7:331–334.

Milne, A. 1959. "The Centric Systematic Area-Sample Treated as a Random Sample." *Biometrics* 15:270–297.

Olea, R. A. 1984. "Sampling Design Optimization for Spatial Functions." *Mathematical Geology* 16:369–392.

Oliver, M. A., and R. Webster. 1986. "Combining Nested and Linear Sampling for Determining the Scale and Form of Spatial Variation of Regionalized Variables." *Geographical Analysis* 18:227–242.

Osiecki, K. M. 2011. "Background Levels in Soil Remediation." Master's thesis, University of Illinois, Chicago.

Pettitt, A. N., and A. B. McBratney. 1993. "Sampling Designs for Estimating Spatial Variance Components." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42:185–209.

Sambanis, A. 2012. "Assessment of Spatial Analysis and Decision Assistance (SADA) Potential for Clean Up." Master's thesis, University of Illinois, Chicago.

"Spatial Analysis and Decision Assistance," last modified 2008, accessed April 7, 2013. http://www.tiem.utk.edu/~sada/index.shtml.

US Environmental Protection Agency. 1983. "Characterization of Hazardous Waste Sites—A Methods Manual, Volume II, Available Sampling Methods." EPA-600/S4-83-040. Research and Development. Environmental Monitoring Systems Laboratory. US Environmental Protection Agency, Las Vegas, NV 89114.

US Environmental Protection Agency. 1984. "Soil Sampling Quality Assurance User's Guide." EPA-600/S4-84-043. Research and Development. Environmental Monitoring Systems Laboratory. US Environmental Protection Agency, Las Vegas, NV 89114.

US Environmental Protection Agency. 1989a. "Methods for Evaluating the Attainment of Cleanup Standards Volume 1: Soils and Solid Media." EPA 230/02-89-042. Office of Policy, Planning, and Evaluation. US Environmental Protection Agency, Washington, DC 20460.

US Environmental Protection Agency. 1992. "Preparation of Soil Sampling Protocols: Sampling Techniques and Strategies." EPA/600/SR-92/128. Research and Development. Environmental Monitoring Systems Laboratory. US Environmental Protection Agency, Las Vegas, NV 89193.

US Environmental Protection Agency. 2002. "Guidance on Choosing a Sampling Design for Environmental Data Collection, for Use in Developing a Quality Assurance Project Plan." EPA/240/R-02/005. Office of Environmental Information. US Environmental Protection Agency, Washington, DC 20460.

## VITA

| | |
|---|---|
| NAME: | Brian Joseph Staehlin |
| EDUCATION: | BA, Liberal Arts and Sciences, Economics, University of Illinois at Urbana-Champaign, Illinois, 2003 |
| TEACHING: | Environmental and Occupational Health Science Division, University of Illinois at Chicago, 2011–2013 |
| EXPERIENCE: | Graduate Student Intern, Artistic Circles/Waterpressures.org, January 2012 to April 2012 |
| | Science Policy Intern, Environmental Law and Policy Center, May 2011 to October 2011 |
| | Administrative Assistant, Robert Half International, April 2006 to August 2011 |