# Chapter 4:  Methods

Any empirical or statistical approach to groundwater vulnerability analysis proceeds from the assumption that high concentrations of contaminants are found more often where vulnerability is high than where vulnerability is low. If a water supply contains a detectable concentration of a man-made pesticide, for example, then that water supply must be vulnerable to contamination, because it has become contaminated.  If many water samples are taken from two supplies, and contaminants appear very frequently in the samples from one supply and much less frequently in samples from the second, one might reasonably conclude that the first supply is more vulnerable to contamination than the second.  Given a large body of water quality measurements from different water sources, it should be possible to gauge the vulnerability of those sources to contamination based on the frequency that contaminants are found in samples from those sources.

This study attempts to form a generally applicable method for inducing the relative vulnerability of groundwater supplies from a large body of contaminant concentration measurements.  The method is spatial and statistical in its approach.  Measurements of contaminant concentration are grouped by their location in specified regions of the subsurface, statistical descriptions of the groups of measurements are formed, and the variation of these statistics from region to region is mapped.  Finally,  to relate the vulnerability of the regions to indicator parameters, the variation of the statistics is compared with variations in hydrologic, soil, and contaminant loading parameters mapped over the same regions.

This chapter describes the mathematical methods used in the study and the assumptions that underlie their use. The chapter is organized along the lines of the six-step outline presented in the last section of Chapter 2. Section 4.1 describes the rationale behind the use of nitrate as a surrogate for vulnerability. Section 4.2 describes the criteria used to select the study regions. Section 4.3 describes the use of GIS and database management systems to form the data into groups for statistical analysis. Section 4.4 describes the calculation of statistical descriptions of the grouped data, and the assumptions underlying the use of those statistics. Section 4.5 describes the use of GIS and stepwise multiple linear regression to form a predictive model from the statistical descriptions of the data and a series of potential indicators. Section 4.6 describes the use of two additional data sets to support the use results based on one body of nitrate measurements to make more general statements about groundwater vulnerability.

## 4.1 NITRATE AS A SURROGATE FOR VULNERABILITY

*Susceptibility*, *vulnerability*, and *probability of contamination* are related, but distinct, ideas. For the purposes of this study, a groundwater supply is said to be susceptible to contamination if it is possible for a contaminant to reach it, even if no source exists for that contaminant. The supply is vulnerable to a particular contaminant if it is susceptible and a source of the contaminant is present. The risk of contamination is the likelihood or probability that the contaminant is actually present in the groundwater. Probability, unlike susceptibility and vulnerability can be described by a number. In other words, probability of contamination is quantifiable, while susceptibility and vulnerability are not.

Although probability of contamination is quantifiable, it is not directly measurable. Water quality measurements describe the degree to which chemical constituents are present in water—that is, their concentration—not risk or probability. How, then, is it possible to conduct an *empirical* investigation of groundwater susceptibility or vulnerability, which cannot be quantified, or of probability of groundwater contamination, which cannot be measured?

**Threshold Concentrations.** This study estimates probabilities of contamination by calculating the frequency with which threshold concentrations of constituents are exceeded in groups of groundwater measurements. These probability estimates serve as surrogates for susceptibility and vulnerability. Four thresholds, in mg/l nitrate as nitrogen, were chosen. The lowest is 0.1 mg/l, the detection level described in Section 3.1. The highest is 10 mg/l, the maximum concentration permissible in public water supplies. Another threshold was chosen at 5 mg/l, which is one-half the MCL, and triggers increased monitoring requirements in public water supplies. The fourth threshold was selected at 1 mg/l to indicate the range at which human influences may be suspected. This last threshold is lower than the level used by Madison and Brunett (1985) as indicative of human influence, but falls in the range they call "transitional," possibly indicating human influence. Since this work examines groups of samples in regions, rather than single wells, it is appropriate to use this lower value; consistent exceedences of this threshold are more indicative of vulnerability than a single exceedence.

**Nitrate as Surrogate Constituent.** Measurements of the groundwater concentrations of solvents, herbicides, PCBs, and other industrial and agricultural chemicals are very scarce in Texas. Because of this scarcity, it is not possible to base a Statewide study on the measurements of the chemical constituents, like atrazine or tolulene, for which monitoring waivers can be granted. Instead, the study is based on roughly 46,000 measurements of nitrate concentration in Texas groundwater. Although waivers cannot be granted for nitrate monitoring, nitrate is a potential surrogate indicator of contamination by agricultural chemicals, a major group of regulated constituents.

Nitrogen fertilizers are very frequently applied to the same crops as pesticides, so it is reasonable to assume that if nitrate can migrate from the crops on the surface to the water in the subsurface, so can the pesticides. The presence of elevated nitrate levels in groundwater is assumed, for purposes of this study, to indicate that a viable pathway exists from the surface, where most nitrate sources are located, to the groundwater. The regulations themselves include elevated nitrate levels in the list of factors that can be considered in a vulnerability assessment for pesticides. Because nitrate has been widely measured for many years (the first nitrate measurement in the database on which the study is based was taken in 1896) a sufficient body of measurements exists to form the basis of an empirical study.

Nitrate is not a perfect indicator of vulnerability to agricultural chemicals, however. Natural mineral sources exist, as do other anthropogenic sources not necessarily related to chemical application, such as septic systems and cattle production. Although this study assumes a relationship between vulnerability to nitrate contamination and vulnerability to contamination by agricultural chemicals, its main task is one of identifying areas vulnerable to nitrate contamination. If a successful methodology for identifying areas vulnerable to nitrate, then the same methods can be applied to other chemicals as monitoring results become available.

## 4.2 IDENTIFICATION OF ANALYSIS REGIONS

The selection of analysis regions defines the study. As following sections will show, the methods used in this study treat the regions as homogeneous bodies, lumping all data and all results by their association with the regions selected in the first step of the process described in Section 2.6. Comparisons are made between regions, but not within them.

A frequently overlooked part of the DRASTIC pollution potential evaluation system (Aller et al., 1987) is the authors' recommendation that the numerical rating system be applied to *hydrogeologic settings*, which they define as "mappable unit[s] with common hydrogeologic characteristics." In other words, the DRASTIC rating system should be applied only to regions that can properly be characterized by a single rating. The four studies cited by the General Accounting Office (GAO, 1992) as attempts to validate DRASTIC with field data use counties as the mapping unit (one also uses smaller units in some cases). Three of these studies find little correlation between DRASTIC ratings and groundwater contamination. The poor correlation may be due in part to the inappropriateness of counties for use as mapping units. The use of counties as mapping units may also account for the lack of correlation between fertilizer sales and the occurrence of nitrate in groundwater shown in the example in Chapter 1 of this report.

In this study, the principal analysis regions are 7.5' quadrangles. Each quadrangle is characterized by descriptive statistics calculated on the results of all measurements collected from wells in that quadrangle, and no distinction is made between different parts of a single quadrangle. Maps of the analysis results show the variation of exceedence probabilities from one quad to another, essentially using a single number for each quad to characterize the results of the analysis.

It follows, then, that in selecting a set of regions for analysis, the designer of the study should have some reasonable expectation that each region is homogeneous. At least there should be less variation in water quality and indicator parameter values within regions than between them. Because of their spatial compactness, 7.5' quadrangles are assumed to meet this requirement.

Although the regions should be internally homogeneous, there should also be a reasonable expectation that there will be significant variations between regions. The scope of the study should be sufficiently large that comparisons of the descriptive statistics from region to region will yield meaningful variations. Because this study includes the entire state of Texas, it is reasonable to assume that 7.5' quadrangles from widely disparate parts of the state will show significant differences in summary statistics of water quality measurements. Certainly, differences in climate, geology, and human activities are great enough that they can be detected in 7.5' quadrangles across Texas.

Since the study method is statistical, there should be enough measurements available in the regions to make meaningful statistical calculations possible. This requirement must be balanced against the requirement that regions be homogeneous. Small regions will be more homogeneous, but will contain fewer measurements, reducing the confidence in the values of statistics calculated from those measurements. 2.5' quadrangles were considered and rejected as study regions after the number of measurements in the two sizes of quadrangles were compared.

For reasons that will be explained in Section 4.4, quadrangles with fewer than 12 measurements were not included in the maps or the regression analyses. As the histograms in Figure 4.1 show, about 1.5% of the 2.5' quadrangles (597 of 38,523) have 12 or more measurements. More than 26% of the 7.5' quadrangles (1,158 of 4,407) have 12 or more measurements. Selecting 7.5' quadrangles over 2.5' quadrangles increased the number of measurements included for mapping and regression analysis, and included a much larger proportion of the area of the state in the study.

Figure 4.2 shows a 7.5' quadrangle (number 5740, which has already been used as an example throughout Chapter 3), the nine 2.5' quads it contains, and the locations of the wells in those quads that were included in this study. 51 nitrate measurements recorded in the TWDB database were taken from 37 wells located in this quadrangle. Only one of the 2.5' quads in 5740 has as many as 12 measurements, and if the measurements were more evenly distributed, none would.

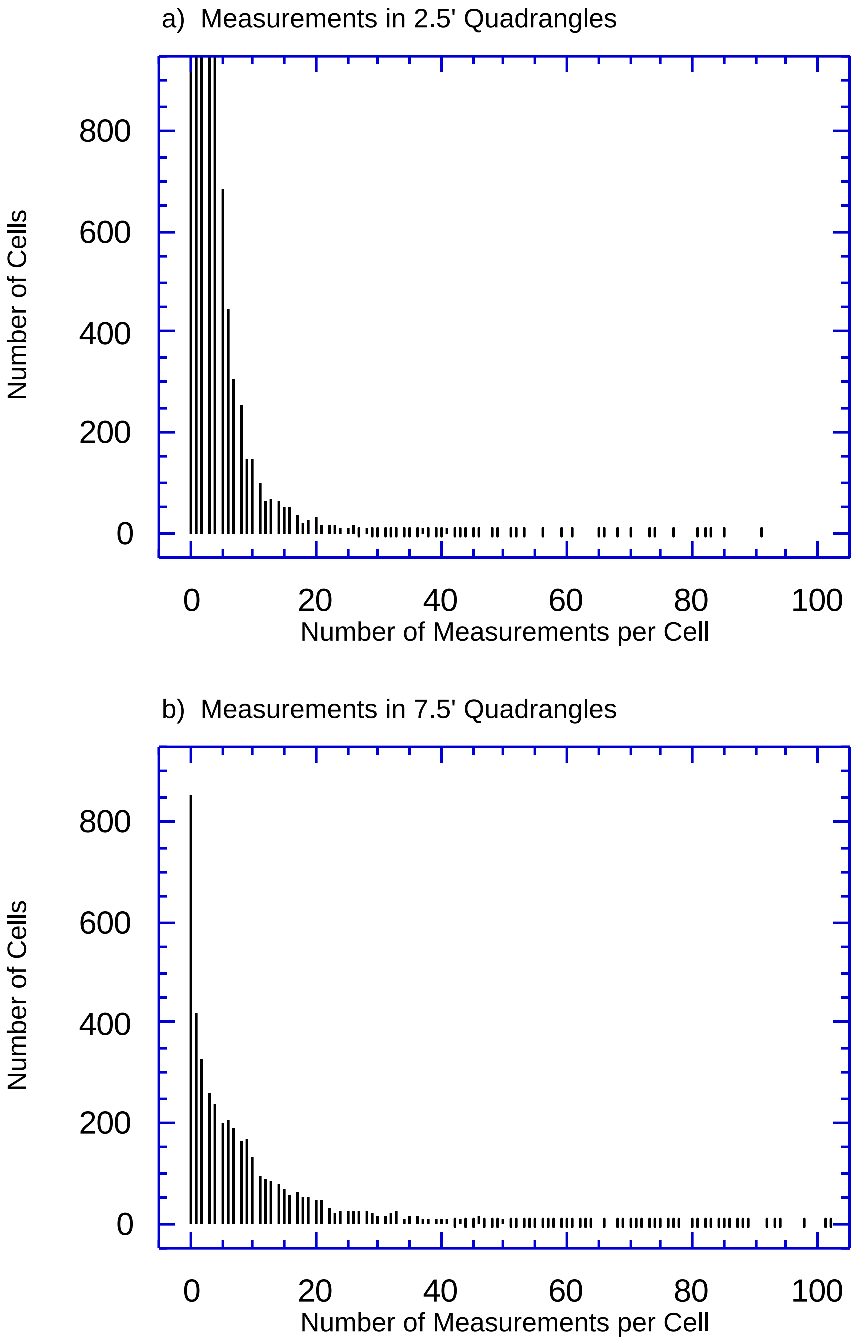


Figure 4.1  Measurement Histograms for 2.5' and 7.5' Quadrangles

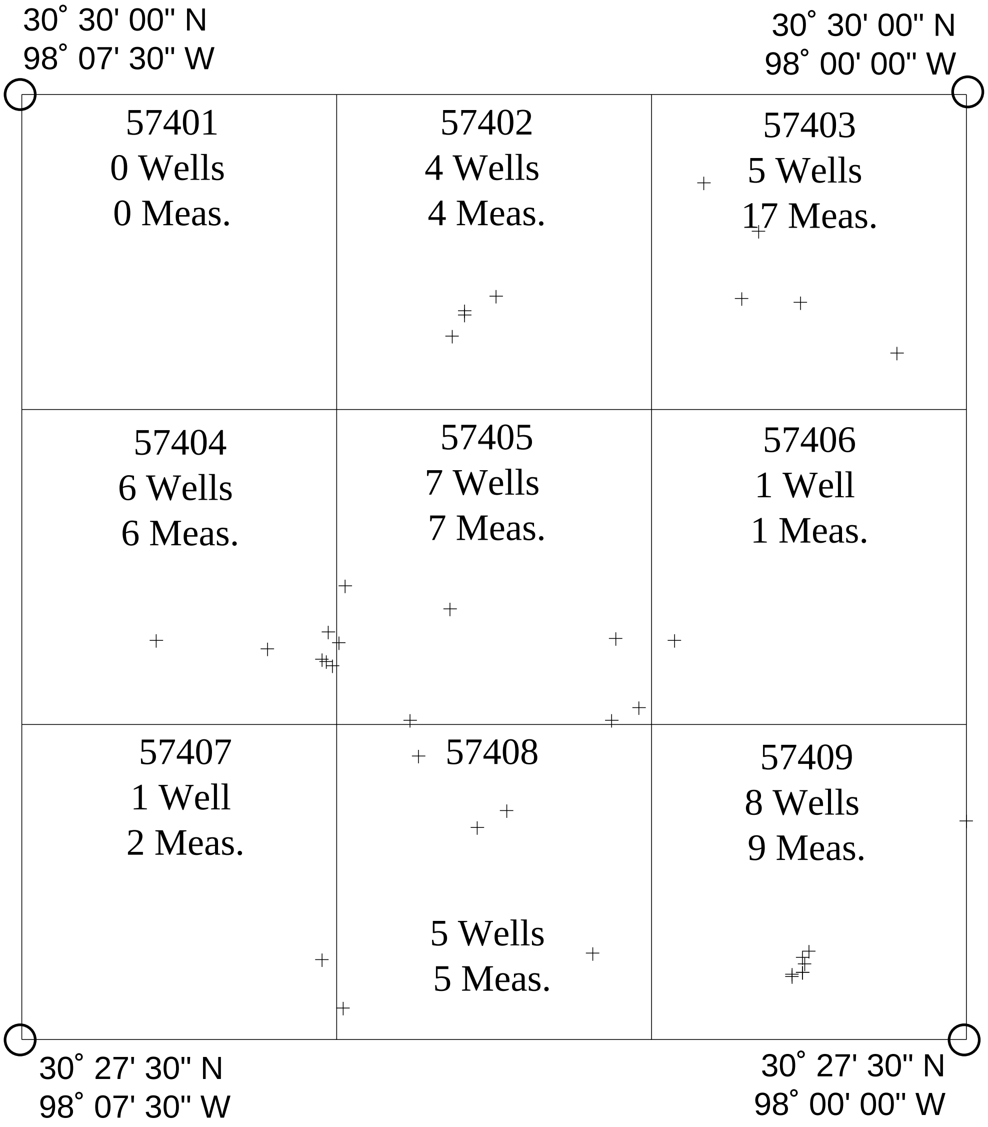


**Figure 4.2 Well Locations in Quadrangle 5740**

The five aquifers selected to form the second set of analysis regions are assumed to be homogeneous because geologic characteristics vary more between the aquifers than within them, and because the water in the aquifers mixes internally much more than between aquifers. The internal homogeneity of the aquifers will be discussed further in Chapter 6, where the results of the analyses are presented. The differences in their geologic structure and the separation of their spatial extents assure that discernible differences can be found between them. The selected aquifers are classified as major aquifers by the TWDB, and nitrate measurements from wells in each of them are plentiful. The aquifers thus meet the same requirements for selection as analysis regions that the 7.5' quadrangles do.

## 4.3 Grouping Data for Analysis

Once the data set has been chosen, and a set of analysis regions has been selected, the data must be sorted into groups for statistical analysis. The formidable task of forming thousands of records of nitrate measurements into meaningful groups is made feasible by database management systems and geographic information systems. This section describes the principles of these technologies that are important to this study, and the application of those principles to the tasks of organizing Texas groundwater data.

### 4.3.1 Database Management Systems

The database management systems used in this study are described in terms of the relational model. Other models for database management systems exist, including entity-relationship, network, hierarchical, and object-oriented models. The relational model is the basis for Structured Query Language (SQL), a widely used system for building, maintaining, and using databases. Although INFO, the database management system used in this study, does not use SQL, the INFO operations carried out in this study can be described in terms of the relational model. Doing so makes this discussion more general, by eliminating references to commands and syntax meaningful only in INFO.

A relational database is a group of tables, each with a unique name. Each row in a table corresponds to an entity of interest to users of the database, and contains a fixed number of attributes, which describe that entity. A simple table of nitrate measurement data might consist of rows containing an ID number for the well where a water sample was collected, the year, month, and day the sample was collected, and the nitrate concentration measured in the sample. The list of attributes in the rows of a database table is called the *scheme* of the table. A table called "*meas*" will be used as an example. The scheme of *meas* is

*meas-scheme* = (well-ID, year, month, day, nitrate).

The scheme of *meas* defines the way that nitrate measurements can be described in this database. Mathematically, the scheme describes the Cartesian product of a set of *domains*, where a domain is a set of possible values. The domain of month, for instance, might be the integer values 1 through 12. Any combination of valid values for all five attributes fits the scheme, whether or not the values correspond to an actual nitrate measurement. To be included in the table however, the combination of values must correspond to an actual nitrate measurement. The table *meas* is thus a subset of the Cartesian product of the domains well-ID, year, month, day, and nitrate. Mathematicians call a subset of the Cartesian product of a set of domains a *relation* . This is the origin of the name "relational" for this database model. An individual element of a relation is an *n-tuple*, or simply a *tuple*. See Korth and Silberschatz (1991) or any number of other database textbooks for a more complete discussion of the relational model.

Operations on relational databases can be described in many ways. This discussion will use the tuple relational calculus. A query in the tuple relational calculus takes the form

$$\{r \mid \mathrm{P}(r)\}$$

and returns the set of tuples *r* such that the predicate P is true for *r*. Predicates are statements about tuples and their attributes, which are evaluated as true or false. Some of the mathematical notations used in the predicates are shown in Table 4.1.

Table 4.1 Predicate Symbols for Relational Calculus

| **Symbol** | **Definition** |
|---|---|
| $\in$ | "Is a member of" |
| $\exists$ | "There exists" |
| $\forall$ | "All" |
| $\wedge$ | "And" |
| $\vee$ | "Or" |

Attribute values are indicated with notation of the form *r*[year], meaning "the value of the attribute year for tuple *r*." For example, the query

$$\{r \mid r \in \mathit{meas} \wedge r[\text{well-ID}] = 5740304\} \qquad (4\text{-}1)$$

reads "all tuples that are members of the relation *meas* and have the value 5740304 for the attribute 'well-ID.'" In more concrete terms, it returns every record of a measurement collected from well number 5740304. The results of this query, applied to data from the TWDB database, are shown in Table 4.2. (Nitrate is given as nitrate-N.)

Table 4.2 Results of Query 4-1

| **Well-ID** | **Year** | **Month** | **Day** | **Nitrate** |
|---|---|---|---|---|
| 5740304 | 1966 | 4 | 2 | 0.10 |
| 5740304 | 1966 | 12 | 14 | 0.10 |
| 5740304 | 1967 | 6 | 20 | 3.17 |
| 5740304 | 1968 | 6 | 7 | 2.71 |
| 5740304 | 1968 | 7 | 26 | 3.05 |
| 5740304 | 1971 | 6 | 4 | 1.81 |
| 5740304 | 1972 | 5 | 0 | 1.81 |
| 5740304 | 1974 | 3 | 11 | 1.33 |
| 5740304 | 1976 | 8 | 5 | 1.06 |
| 5740304 | 1980 | 3 | 24 | 0.88 |
| 5740304 | 1986 | 6 | 10 | 0.48 |
| 5740304 | 1991 | 8 | 26 | 0.10 |

A group of queries can be used to provide data for the comparison of data selected by different criteria. The following queries, for example, show that a greater proportion of samples collected in 1990 contained nitrate in excess of 1 mg/l than those collected in 1964. This point is examined in more detail in following chapters.

$$\{r \mid r \in \mathit{meas} \wedge r[\text{year}] = 1964\}$$

$$\{r \mid r \in \mathit{meas} \wedge r[\text{year}] = 1964 \wedge r[\text{nitrate}] > 1.0\}$$

$$\{r \mid r \in \mathit{meas} \wedge r[\text{year}] = 1990\}$$

$$\{r \mid r \in \mathit{meas} \wedge r[\text{year}] = 1990 \wedge r[\text{nitrate}] > 1.0\}$$

The first query returns all records of nitrate measurements taken in 1964; the second returns all records of nitrate measurements taken in 1964 that report concentrations greater than 1 mg/l. The third and fourth queries return similar records for the year 1990. By counting the number of records returned with each query, it can be found that 400 of 1,324 measurements (30%) in 1964 and 608 of 1,166 measurements (52%) in 1990 showed nitrate concentrations above 1 mg/l.

The real power of relational databases comes from their ability to combine information from multiple tables. If a second scheme is defined as

*well-scheme* = (well-ID, depth),

sets of wells can be selected on the basis of their depth and, more importantly, sets of measurements can be selected on the basis of the depth of the well from which they were collected, as well as the year in which they were collected. The attribute well-ID, which is common to both tables, provides a means for linking the two tables. Such linking attributes are called "keys." The query

$$\{r \mid r \in meas \wedge \exists\, s \in well \ \ (r[\text{well-ID}] = s[\text{well-ID}] \wedge s[\text{depth}] < 100)\} \quad (4\text{-}2)$$

reads "all tuples that are members of the relation *meas* for which there exists a tuple in the relation *well* with the same value for the attribute well-ID and with a value less than 100 for the attribute depth." More intuitively, the query returns all nitrate measurement records for which the corresponding well record indicates a well depth less than 100, where "corresponding" means "having the same well number." More practically, it returns all records of samples collected from wells less than 100 feet deep.

The earlier queries about 1964 and 1990 can be modified to include only samples collected from wells less than 100 feet deep, like this

$$\{t \mid t \in \textit{meas} \wedge t[\text{year}] = 1964 \wedge \exists\, s \in \textit{well}\ \ (t[\text{well-ID}] = s[\text{well-ID}] \\ \wedge\ s[\text{depth}] < 100)\}$$

$$\{t \mid t \in \textit{meas} \wedge t[\text{year}] = 1964 \wedge t[\text{nitrate}] > 1.0 \wedge \exists\, s \in \textit{well} \\ (t[\text{well-ID}] = s[\text{well-ID}] \wedge s[\text{depth}] < 100)\}$$

$$\{t \mid t \in \textit{meas} \wedge t[\text{year}] = 1990 \wedge \exists\, s \in \textit{well}\ \ (t[\text{well-ID}] = s[\text{well-ID}] \\ \wedge\ s[\text{depth}] < 100)\}$$

$$\{t \mid t \in \textit{meas} \wedge t[\text{year}] = 1990 \wedge t[\text{nitrate}] > 1.0 \wedge \exists\, s \in \textit{well} \\ (t[\text{well-ID}] = s[\text{well-ID}] \wedge s[\text{depth}] < 100)\}$$

The first two queries of this group return records showing that in 1964, 304 of 517 measurements (59%) taken from wells less than 100 feet deep showed nitrate concentrations greater than 1 mg/l. The last two queries return records showing that in 1990, 210 of 272 measurements (77%) taken from wells less than 100 feet deep showed nitrate concentrations greater than 1 mg/l.

Relational databases are capable of carrying out much more complicated queries than the examples given here, involving more tables, and returning values for any subset of the attributes those tables contain. The examples here illustrate the most important features used in this study.

Because the well-numbering system used by the TWDB includes in the well ID the numbers of the 1_, 7.5', and 2.5' quadrangles where each well is located, queries of the type shown here are sufficient to group nitrate measurements by quadrangle. Similarly, since the well-description data provided by TWDB includes the names of geologic formations from which the wells draw water, queries of the same type will also group measurements by aquifer.

In general, however, locating wells and water-quality measurements in regions defined by maps requires operations that cannot be performed by database management systems alone. Grouping and querying of data by spatial categories will usually require a geographic information system.

### 4.3.2 Geographic Information Systems

A geographic information system (GIS) stores data about the world in thematic maps or data layers, called coverages, which contain different kinds of features and information. A coverage of Texas, for instance, could show political features, such as counties, or physical features such as rivers. These features would be stored in different data layers, with different information, although they occupy the same space on the earth's surface. A GIS coverage may incorporate database tables, which describes the attributes of the features mapped in the coverage.

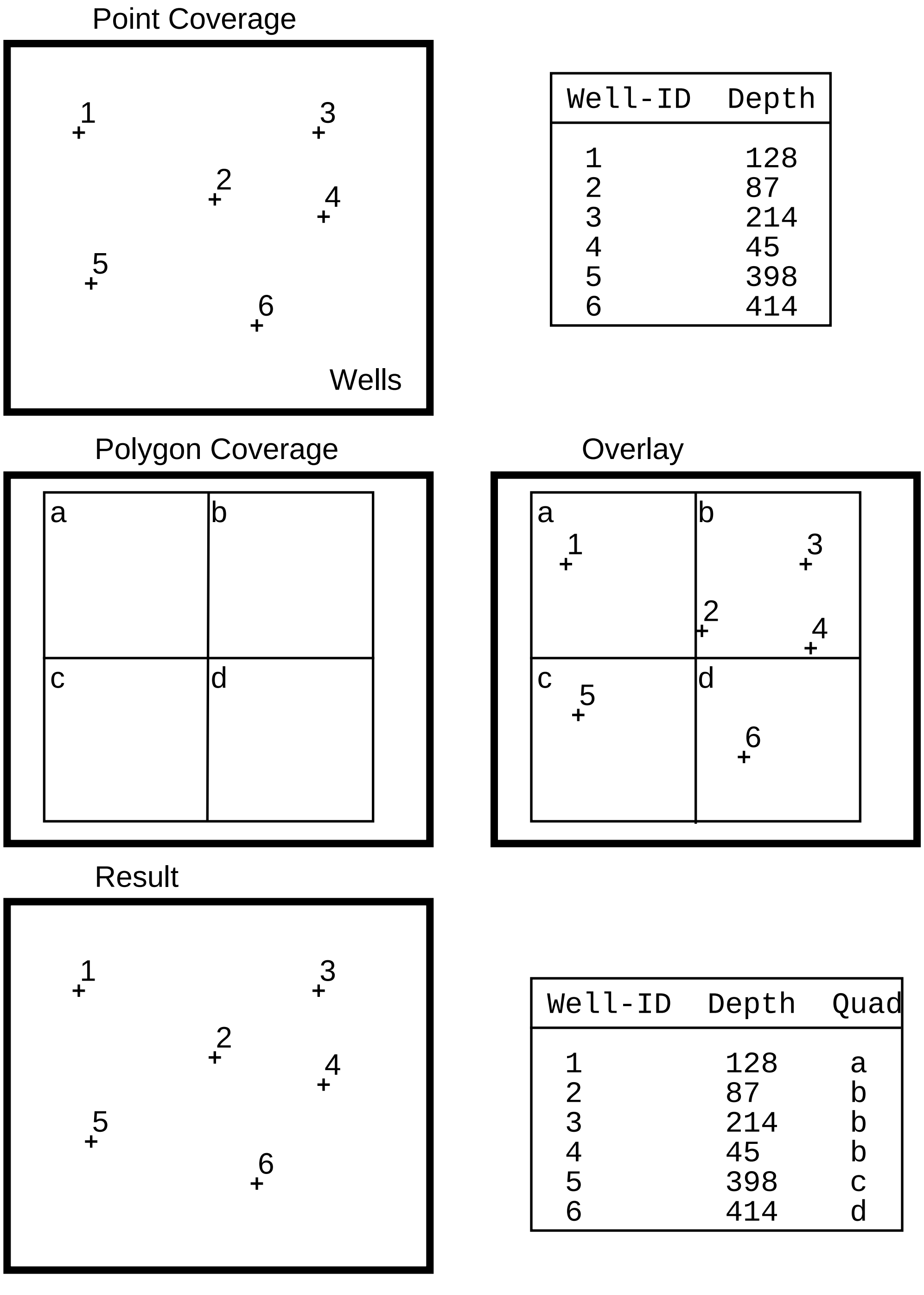
GISs fall into two broad categories, vector and raster. Arc/Info, the GIS used in this study, has modules for representing features in both vector and raster systems (ESRI, 1991). The quadrangles used as analysis regions are constructed in the vector system. Raster systems will be discussed further in Sections 4.5 and 4.6.

A vector GIS represents features as points, lines, or polygons. Points are represented by a single pair of coordinate values, lines by series of points, and polygons by closed sets of lines. Lines and polygons can take any shape, and descriptive data can be linked to features of any type. A vector GIS coverage can contain points only; points and lines; or points, lines, and polygons. Attribute data can be stored for all types of features present in a coverage, but is often associated only with the highest-order features. Typically, a coverage is classified by its highest-order feature as a point coverage, line coverage, or polygon coverage.

Features in a coverage can be thought of as elements of a set, like the records in a database table. Subsets of objects can be formed on the basis of location, attribute values, or a combination, and set operations such as union or intersection can be performed on these subsets.

Since attribute values are stored in database tables, subsets of features can be formed on the basis of attribute values by database queries of the type described in the last section. Grouping data by location requires special operations unique to GIS.

Figure 4.3 illustrates one such operation, the overlaying of polygons on points. In a vector GIS, a point is a single location, and can be used to represent features like wells; a polygon is a contiguous, bounded area on the surface of the earth, and can be used to represent quadrangles. Because the GIS can represent the topology of points and polygons and their relative locations, it is able to identify the polygons that points lie within. At the top, the figure shows a point coverage containing six points representing wells, and the data table associated with that coverage—called a *point attribute table*. Below the point coverage is a polygon coverage containing four quadrangles. The corresponding polygon attribute table is omitted from the figure. The two coverages are combined in an *overlay* operation, and the result is shown at the bottom of the figure. Because the topology of the point coverage is unchanged, the result of the overlay is the addition of a new attribute in the point attribute table identifying the quadrangle in which the wells are located. Wells can now be grouped by quadrangle using ordinary database queries.



| Well-ID | Depth |
|---|---|
| 1 | 128 |
| 2 | 87 |
| 3 | 214 |
| 4 | 45 |
| 5 | 398 |
| 6 | 414 |

| Well-ID | Depth | Quad |
|---|---|---|
| 1 | 128 | a |
| 2 | 87 | b |
| 3 | 214 | b |
| 4 | 45 | b |
| 5 | 398 | c |
| 6 | 414 | d |

**Figure 4.3  Locating Points in Polygons**

If the polygons have attributes of interest, these can be linked to the wells by using the quadrangle number as a key to link the point attribute table of the well coverage to the polygon attribute table of the quadrangle coverage. If the quadrangle coverage has an attribute called "thick" equal to the average soil thickness (in inches) in the quadrangle, the following query would return all records for wells located in quads where the average soil thickness is greater than 60 inches.

$$\{t \mid t \in \mathit{wells} \wedge \exists\, s \in \mathit{quads}\ (t[\text{quad}] = s[\text{quad}] \wedge s[\text{thick}] < 60)\}$$

A more complex query, incorporating a third table, could similarly produce all records of nitrate measurements collected from wells located in quadrangles where the average soil thickness is greater than 60 inches. The linkage between the topology of a coverage and the database tables containing the attributes of features in that coverage lies at the heart of GIS. The ability to represent the results of spatial operations like point-in-polygon overlays in database tables greatly increases the value of those tables to investigators trying to understand the influence of spatially distributed processes.

Polygon-on-polygon overlays, and their use in describing the co-incidence of different spatially distributed parameters will be discussed in a later section.

Given a database consisting of two tables, one of nitrate measurements and one of well descriptions, and a GIS coverage consisting of 7.5' quadrangles, the methods described in this section are sufficient to extract from the database all records of nitrate measurements from any quadrangle in the coverage. If the well description table also includes the names of the aquifers that the wells tap, the same methods can also extract all records of measurements from those aquifers. The statistical analysis used to summarize those measurements is described in the next section.

## 4.4 STATISTICAL MODEL OF VULNERABILITY

In this study, it is assumed that the concentration of a chemical constituent in groundwater is a random function of space and time,

$$C = C_R(x, y, z, t) \tag{4-3}$$

where C is a concentration value, x and y are coordinates parallel to the surface of the earth, z is a vertical coordinate, t is time, and the subscript R denotes a random function. The randomness of the function means that it is impossible to predict an exact value for the concentration, and that a prediction of concentration can properly be described only as a probability function. This impossibility can be interpreted as the result of a process governed by chance, or as a statement of the limits of human knowledge. These two interpretations are not mutually exclusive, but the latter fits this study better because the state of knowledge about groundwater is very limited, and that limitation motivates the study.

If the concentration of a constituent at a point is described by a random function, then the concentration of the constituent in any finite volume of groundwater, such as a sample drawn from a well for analysis, is also described by a random function. At any given moment, a larger volume of the subsurface, such as an aquifer or the volume underneath a 7.5' quadrangle of the earth's surface, contains an infinite number of sample-sized volumes. The concentration values associated with this infinite collection of potential water samples make up a *population*, which can also be described by a probability function.

If complete knowledge of the population were somehow available, that is, if the concentration in every possible sample-sized volume could be known, the probability function could be calculated directly. If $P(C_t)$ is the probability that the concentration in a single sample-sized volume selected at random from the population is less than or equal to a threshold concentration $C_t$, then

$$P(C_t) = \frac{N_e}{N_l + N_e} , \qquad (4\text{-}4)$$

where $N_l$ is the number of sample-sized volumes of water in which the concentration is less than or equal to the threshold, and $N_e$ is the number of such volumes in which the concentration exceeds the threshold. More simply, this is the number of exceedences in the population divided by the total population. Since the population is infinite in number, both $N_e$ and $N_l$ are infinite, but their ratio is finite. Rewriting equation 4-4 as

$$P(C_t) = \frac{N_e / N_l}{1 + N_e / N_l} , \qquad (4\text{-}5)$$

avoids the difficulty of expressions involving infinite numbers. For any water-bearing volume of the subsurface, Equation 4-5 maps any concentration value (any number greater than or equal to zero) to a monotonically increasing number between zero and one, defining a cumulative probability function. If the function is differentiable, its derivative is the probability density function (pdf) for the concentration values in the population.

Although the discussion above describes a population as a body of concentration values determined over a finite region of space at an instant, the same argument would apply as well to a finite region of space over a finite period of time. As time passes, water moves in and out of the region carrying different levels of the constituent with it and changing the concentrations inside the region. From a mathematical standpoint, this is no different from the variation from point to point over the region at a fixed time, the concentration simply varies in four dimensions instead of three. The population is enlarged by the addition of a dimension, but the definition of the probability functions is unchanged.

**Parameters and Statistics.** Properties of the cumulative probability function and the pdf are *parameters* of the population. For the purposes of this study, parameters include not only the usual measures of central tendency (mean, median, etc.), spread (standard deviation, interquartile range, etc.), and so on, but also the probabilities associated with concentrations values that are of particular interest (detection limit, maximum contaminant level, etc.).

In ideal version of this study, Texas would be divided into analysis regions at an instant, and the parameters of the populations associated with those regions would be mapped and analyzed. This ideal study, however, requires complete knowledge of the populations in the analysis regions, knowledge that is plainly unavailable.

Instead, the study deals with *statistics*, or estimates of the parameters calculated from a finite (and small) number of actual measurements in the sectors. The actual measurements form a *sample* of the population.

**Two Probability Estimation Methods.** Two sets of statistics, representing two models of exceedence probabilities, are calculated for the 7.5' quadrangles. The first set are non-parametric estimates of the probabilities that a the nitrate concentration at a point selected at random beneath the quadrangle will exceed a selected threshold value. The second set are the two parameters (mean and standard deviation) of the lognormal distribution that best fits the distribution of nitrate concentrations measured in wells in the quadrangle.

### 4.4.1 Discrete Probability Estimates

To calculate a discrete probability, the quadrangle is imagined to be an urn containing a very large number of red and green balls. For example, if 5 mg/l nitrate-N is selected as the threshold, any potential water sample in the population beneath the quad with a nitrate concentration greater than 5 mg/l would be represented as a red ball, and any potential water sample with a nitrate concentration less than or equal to 5 mg/l would be represented as a green ball. A red ball might represent a concentration of 5.5 mg/l or 300 mg/l; no distinction would be made between these two values. If the number of red balls ($N_r$) and the number of green balls ($N_g$) in the urn are known, the probability of drawing a red ($P_r$) ball is given by

$$P_r = \frac{N_r / N_g}{1 + N_r / N_g}, \qquad (4\text{-}6)$$

which is the same as Equation 4-5. If balls are drawn from an urn containing an infinite number of balls, or drawn from a finite supply and replaced, the ratio of red balls drawn to total balls drawn will be described by the binomial distribution. If n balls are drawn from the urn, the most likely value for $n_r$, the number of red balls drawn is the integer nearest $nP_r$.

The probability of drawing s red balls in n trials is equal to

$$e(n, s, P_r) = \binom{n}{s}(P_r)^{s}(1 - P_r)^{n-s} \qquad (4\text{-}7)$$

where $\binom{n}{s}$ is the number of combinations of n trials that contain s successes (Snedecor and Cochran, 1980). The cumulative probability of s *or more* successes in n trials is given by the sum of all $e(n, m, P_r)$ with m greater than or equal to s.

$$E(n, s, P_r) = \sum_{m=s}^{n} e(n, m, P_r) \qquad (4\text{-}8)$$

**Water Sampling as a Bernoulli Process.** If it were possible to test all the analyzable volumes of water in a sample partition, the ratio of measurements exceeding to measurements not exceeding the threshold could be determined in the same way as the ratio of red to green balls in an urn. The probability that a single sampling event would exceed the threshold could be calculated from Equation 4-6 and the binomial distribution would describe the outcomes of a series of measurement events in the same way that it describes balls drawn from an urn.

If we know that an urn contains a mixture of red and green balls but we do not know the ratio of red to green, we can estimate the ratio by repeatedly drawing a ball from the urn and keeping track of the numbers of red and green balls drawn. Again, if the drawn ball is replaced after each trial or if the urn contains an infinite number of balls, the ratio of red to green is unchanged, and the outcome of the trials will take the form of a binomial distribution. The best estimate of the ratio of red to green balls in the urn is simply the ratio of red balls drawn to green balls drawn. The expected accuracy of this estimate increases as more balls are drawn. Similarly, when water is drawn from a region, the best estimate of the underlying probability that a constituent's concentration exceeds a threshold is the number of exceedences divided by the number of measurements.

**Estimating Probability from Trials.** In general, if a series of n trials results in s successes—drawing a red ball, detecting a constituent in a concentration that exceeds a threshold, etc.—the best estimate of the underlying probability of success, P, for a single trial is

$$\hat{P} = \frac{s}{n} \ . \qquad (4\text{-}9)$$

Although this is the best estimate, it is more appropriate to express the probability estimate as a range of possible values and a degree of confidence that the true probability falls in that range. This takes the form of a statement like "The probability of success in a single trial lies between 40% and 60% with a confidence level of 95%," or "There is a 5% chance that the probability of success in a single trial lies outside of the range between 40% and 60%."

To estimate the upper and lower bounds on an estimate of the probability of success in a trial from the results of several trials, the following steps are followed.

1. Select a two-sided confidence level, 1-α, for the range. This is the likelihood the true probability will lie between the upper and lower bounds calculated. The probability that the true value lies outside the range is equal to α.
2. Calculate the lower bound, $P_l$ , on the estimate by the following method.

For s = 0, i.e. no successes,

$$P_l(0) = 0 \tag{4-10}$$

For s = n, i.e. all successes,

$$P_l(n) = \sqrt[n]{\alpha} \tag{4-11}$$

For s = 1, 2, …, n-1, find the value of $P_l(s)$ such that

$$1 - E(n, s, P_l(s)) = 1 - \frac{\alpha}{2} \tag{4-12}$$

where E(n, s, P) is the cumulative binomial probability function, eq. 4-8.

3. Calculate the upper bound, $P_u$, through symmetry, using the relation

$$P_u\,(s) = 1 - P_l(n - s). \tag{4-13}$$

Steps 2 and 3 require inversion of the binomial distribution. This method of finding confidence intervals on binomial probability estimates is described by the Harvard University Computation Laboratory (1955).

If, for example, 2 out of 10 measurements exceed a 5 mg/l threshold, the best estimate of the exceedence probability $\hat{P}_e$(5 mg/l) is 0.2, and the 90% (two-sided) confidence limits on the exceedence probability are approximately 0.037

and 0.507. If twenty out of 100 measurements exceed the same threshold, the best estimate of P$_e$ remains unchanged, but the 90% confidence interval now falls between 0.137 and 0.277.

**Binomial Estimates of Exceedence Probabilities.** Using Equations 4-10 through 4-13, it is possible to calculate the best estimate of the exceedence probability for any threshold, and upper and lower confidence limits on that estimate, from a sample composed of any number of measured concentrations. For example, Table 4.3 lists the 51 nitrate concentration values listed in the study database for measurements taken in wells located in quadrangle 5740. 35 measurements exceed concentrations of 0.1 mg/l. 20 measurements exceed 1 mg/l. 2 measurements exceed 5 mg/l and 10 mg/l. Table 4.4 shows the results of estimating exceedence probabilities from these measurements using the binomial distribution as a basis for calculation.

Table 4.3 Nitrate Concentrations in Quadrangle 5740

| **Nitrate Concentration** (mg/l as Nitrogen) | | | | | |
|---|---|---|---|---|---|
| $^2$0.10 | $^2$0.10 | 0.34 | 0.79 | 1.33 | 3.17 |
| $^2$0.10 | $^2$0.10 | 0.34 | 0.81 | 1.58 | 3.17 |
| $^2$0.10 | $^2$0.10 | 0.41 | 0.88 | 1.70 | 4.52 |
| $^2$0.10 | $^2$0.10 | 0.45 | 0.90 | 1.81 | 4.75 |
| $^2$0.10 | $^2$0.10 | 0.48 | 1.06 | 1.81 | 12.67 |
| $^2$0.10 | $^2$0.10 | 0.68 | 1.13 | 2.15 | 15.61 |
| $^2$0.10 | $^2$0.10 | 0.79 | 1.24 | 2.26 | |
| $^2$0.10 | 0.20 | 0.79 | 1.24 | 2.71 | |
| $^2$0.10 | 0.34 | 0.79 | 1.24 | 3.05 | |

Table 4.4 Estimated Exceedence Probabilities for Quadrangle 5740

| **Threshold**<br>(mg/l nit.-N) | $\hat{P}_e$ | $P_l$<br>90% two-sided | $P_h$<br>90% two-sided |
|---|---|---|---|
| 0.1 | 69% | 56% | 79% |
| 1 | 39% | 28% | 52% |
| 5 | 4% | 0.7% | 12% |
| 10 | 4% | 0.7% | 12% |

**Minimum Levels of Confidence.** As more measurements are taken from a population, the degree of confidence in the estimate of an exceedence probability increases—that is, the gap between the upper and lower bounds on the estimate decreases. If the sample of the population consists of a single measurement, and that measurement falls below the threshold, then the estimated exceedence probability is zero (also the lower bound for any confidence interval), but the upper bound of the 90% confidence interval is 0.9. In other words, for nine cases out of ten a single measurement below the threshold comes from a population with an exceedence probability less than 0.9. This is a very weak characterization of the population. If an exceedence probability estimate is to be included in a map or a regression analysis we would like it to make a more definitive statement.

Two possible criteria for including a measurement in the maps and regressions were considered. The first was that an exceedence probability estimate would be included only if it was based on at least a minimum number of measurements. The second was that an exceedence probability estimate would be included if the difference between the upper and lower bounds of the 90% confidence interval was less than a selected value (33%, for example). The two criteria produce different sets of included estimates because the difference between the upper and lower bounds is greater when the probability estimate is close to 0.5 than when it is close to one or zero.

Figure 4.4 shows the 90% confidence intervals on probability estimates calculated from a sample of twelve trials. If six trials are successful, then we can say with 90% confidence that the probability of success in a single trial lies somewhere between 25% and 75%. If no trials are successful, we can say with the same confidence that the probability of success in a single trial is less than 17.5%.

This figure reveals a dilemma in the choice of a method for selecting exceedence probability estimates for inclusion in the maps and regressions. If the selection criterion is a maximum confidence interval, then very few estimates close to 0.5 will pass the test and the maps and regressions will be biased toward the extreme values of exceedence probabilities. If a minimum number of measurements is required, then many estimates with small confidence intervals will be excluded from the maps and regression. Figure 4.4 illustrates this problem

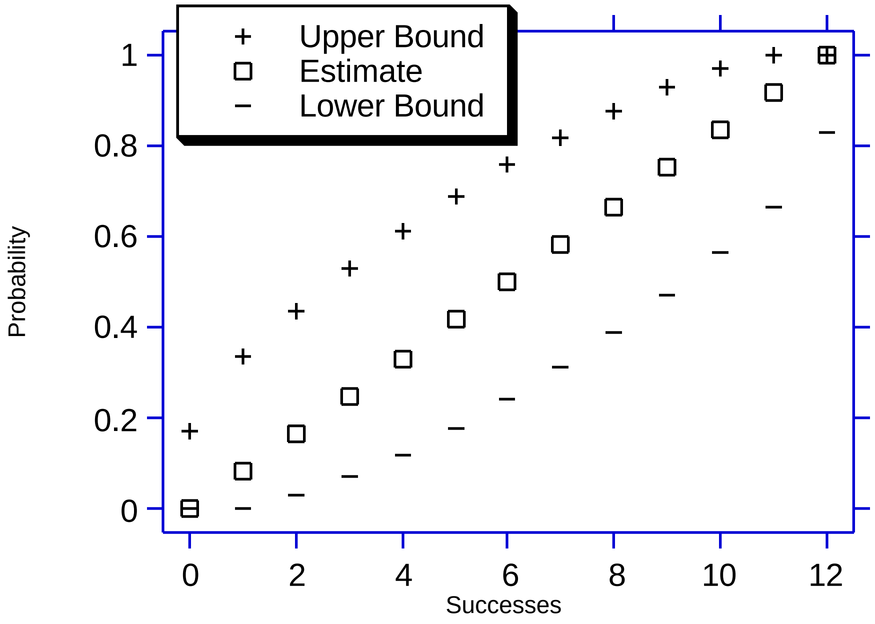


Figure 4.4  Estimating Probability of Success from a Sample of Twelve Trials

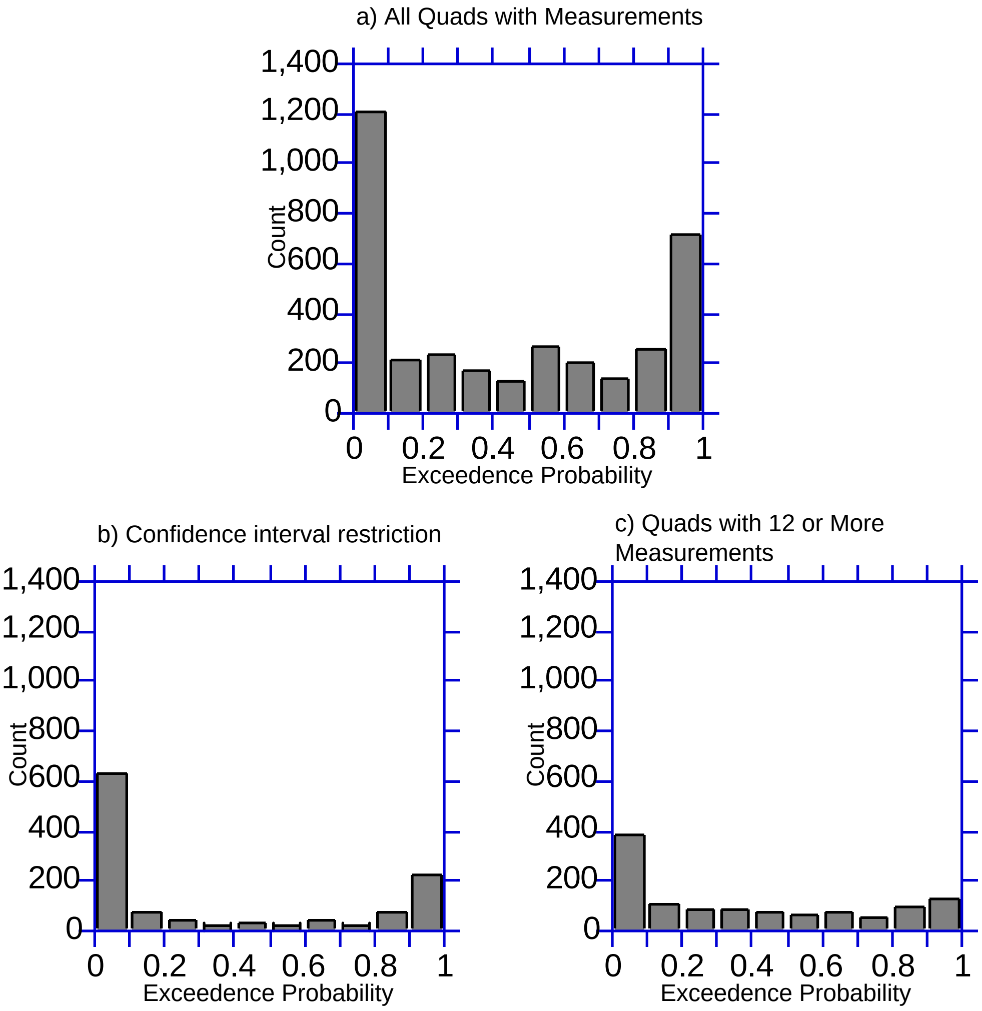


Figure 4.5 Effects of Different Inclusion Criteria

with a series of histograms showing the number of quads falling into bins based on the estimated 1 mg/l exceedence probability for the quads.

In Figure 4.5a, all quadrangles with any measurements at all are included, even those with only one measurement. The inclusion of single-measurement quads leads to high counts at the high and low ends of the probability scale.

Figure 4.5b shows the results of restricting the counts to cells with a 90% confidence interval width of 0.33 or less. Again, the counts at the extreme values are high, and most of the quads in the middle range have dropped out. Figure 4.5c shows the results of restricting the counts to cells with 12 or more measurements. This decreases the number of included quads at the extreme values and increases the number in the middle range, producing a cross-section of probability estimates that more closely follows the unrestricted set, but allows middle-value quads to be included when their confidence intervals are greater than those of extreme-value quads that were excluded.

The minimum-number-of-measurements criterion was chosen because it better reflects the unrestricted data set. The minimum number of measurements for a quad to be included in the maps and regression was set at twelve, because the worst case uncertainty (widest confidence interval) was ± 0.25 for an exceedence probability estimate of 0.5. This was judged to be the widest tolerable confidence interval for inclusion.

In summary, the discrete exceedence probability estimates are calculated by the following method.

1. The total number of nitrate measurements are counted.
2. The number of measurements exceeding the selected threshold are counted.
3. An exceedence probability is estimated by dividing the number of exceedences by the number of measurements.

4. If the number of measurements in the quadrangle is greater than twelve, the exceedence probability is included in maps and regression analysis.

### 4.4.2 Lognormal Probability Estimates

If the probability distribution of a population follows a particular function, such as the lognormal distribution, the probability that a measurement will exceed a threshold can be calculated from that function's definition and a small number of parameters. Estimates of the distribution parameters are, like the discrete probabilities in the preceding section, statistics calculated from sample data.

In the case of exceedence probabilities for chemicals in groundwater, there is no reason to believe *a priori* that the true probability density of the population in a sample partition will match the form of an analytical function exactly, so any assumed function is an approximation. The choice of an analytical function is based on three factors: the suitability of the form of the function to the sample data, the number of parameters, and the calculability of the parameters. The ideal function would fit the sample data and have a small number of easily calculated parameters.

In this study, the lognormal distribution is used as an approximate form for the continuous probability distribution of constituent concentrations. This choice is based on both appropriateness to groundwater processes, and pragmatic concerns. In general, processes such as infiltration and percolation, which follow multiplicative rules, tend to produce lognormally distributed results, so lognormal distributions are fairly common in groundwater systems. As a practical matter, fitting more than two parameters is often difficult and tends to produce inconsistent results. Of the commonly used one- and two-parameter distribution forms (exponential, normal, lognormal) the lognormal distribution appears to fit the data in this study the best. The exponential probability density function is monotonically decreasing, and the normal probability density function is symmetrical about the mean; neither of these conditions is true for the distribution of nitrate concentrations.

Estimates of parameters for some distributions, including the lognormal distribution, can be calculated from the moments of the data. However, this method of estimation cannot be applied when the data are *censored*, as are the water quality data used in this study.

Censoring occurs when some of the data are identified as "less than" or "greater than" some limiting value, rather than as exact values. Probability distribution parameters can only be calculated from the moments of censored data if specific values are assumed for data falling in the censored range (i.e. below the detection limit).

Instead of calculating parameters from moments, it is possible to evaluate the parameters by calculating a "best fit" to the data over the uncensored range. For any value of constituent concentration actually recorded for measurements in a sampling region, the number of exceedences can be counted, yielding an estimate of the value of the cumulative probability function at each recorded value. Values of the parameters of the selected distribution form are chosen to minimize or maximize a fitting score, such as the sum of squares of deviations or the likelihood function. This parameter-fitting method is a numerical analog to graphical fitting by plotting the values on probability paper.

The following method was used to estimate the parameters of the lognormal distribution of a group of measurements. The method is illustrated with data from Quadrangle 5740, which is summarized in Table 4.5 and Figures 4.6.

1. The measurements were ranked by concentration from high to low (as in Table 4.4).

2. The common (base 10) logarithm of each unique concentration value was calculated.

3. An estimated cumulative probability for each unique concentration value Blom's formula,

$$p(X\ ^{2}\ X_m) = \frac{m - 3/8}{n + 1/4} \qquad (4\text{-}14)$$

   was used to estimate the probability, with $X = \log_{10}(C)$, the log of a concentration value, n is the total number of measurements and $X_m$ is the mth-ranked concentration value. Blom's formula produces nearly unbiased estimates of probability for normally distributed data (Chow, et al., 1988).

4. The normal variate z corresponding to each cumulative probability value was calculated by inversion of the gaussian normal probability function ($z(0.16) = -1$, $z(0.5) = 0$, $z(0.84) = 1$, etc.). This was calculated from the Blom's formula p using (for $0 < p < 0.5$)

$$w = \left[\ln\left(\frac{1}{p^2}\right)\right]^{1/2} \qquad (4\text{-}15)$$

$$z = w - \frac{2.515517 + 0.802853w + 0.010328w^2}{1 + 1.432788w + 0.189269w^2 + 0.001308w^3} \qquad (4\text{-}16)$$

When p = 0.5, z = 0. When p > 0.5, 1-p is substituted for p in eq. 4-15, and the z value calculated from eq. 4-16 is given a negative sign (Abramowitz and Stegun, 1965).

Table 4.5  Data for Lognormal Fit to Quadrangle 5740

| Rank | C (mg/l - N) | log (C) | Blom's P | z(P) |
|---|---|---|---|---|
| 16 | 0.10 | -1.00 | 0.30 | -0.51 |
| 17 | 0.20 | -0.70 | 0.32 | -0.46 |
| 20 | 0.34 | -0.47 | 0.38 | -0.30 |
| 21 | 0.41 | -0.39 | 0.40 | -0.25 |
| 22 | 0.45 | -0.34 | 0.42 | -0.20 |
| 23 | 0.48 | -0.32 | 0.44 | -0.15 |
| 24 | 0.68 | -0.17 | 0.46 | -0.10 |
| 28 | 0.79 | -0.10 | 0.54 | 0.10 |
| 29 | 0.81 | -0.09 | 0.56 | 0.15 |
| 30 | 0.88 | -0.05 | 0.58 | 0.20 |
| 31 | 0.90 | -0.04 | 0.60 | 0.25 |
| 32 | 1.06 | 0.03 | 0.62 | 0.30 |
| 33 | 1.13 | 0.05 | 0.64 | 0.35 |
| 36 | 1.24 | 0.09 | 0.70 | 0.51 |
| 37 | 1.33 | 0.13 | 0.71 | 0.57 |
| 38 | 1.58 | 0.20 | 0.73 | 0.63 |
| 39 | 1.70 | 0.23 | 0.75 | 0.69 |
| 41 | 1.81 | 0.26 | 0.79 | 0.82 |
| 42 | 2.15 | 0.33 | 0.81 | 0.89 |
| 43 | 2.26 | 0.35 | 0.83 | 0.96 |
| 44 | 2.71 | 0.43 | 0.85 | 1.04 |
| 45 | 3.05 | 0.48 | 0.87 | 1.13 |
| 47 | 3.17 | 0.50 | 0.91 | 1.34 |
| 48 | 4.52 | 0.66 | 0.93 | 1.47 |
| 49 | 4.75 | 0.68 | 0.95 | 1.63 |
| 50 | 12.67 | 1.10 | 0.97 | 1.86 |
| 51 | 15.61 | 1.19 | 0.99 | 2.25 |

5. The best fit for the function

$$z[P(X\ ^{2}\ X_m)] = a + b\cdot X_m \qquad (4\text{-}17)$$

was calculated by least squares regression. (See Figure 4.6)

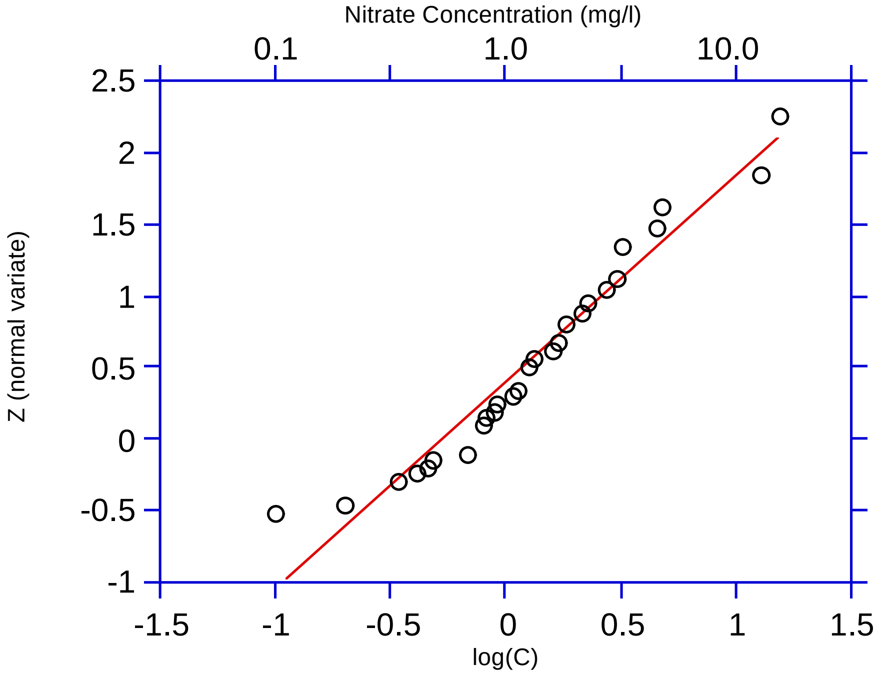


Figure 4.6 Fitting a Probability Distribution by Regression for Quadrangle 5740

6. The lognormal parameters were calculated from a and b as

$$\mu_X = -a/b \qquad (4\text{-}18)$$

$$\sigma_X = 1/b. \qquad (4\text{-}19)$$

Where $\mu_X$ is the mean and $\sigma_X$ is the standard deviation of the log-transformed concentrations.

7. An exceedence probability of a threshold concentration C is calculated by finding the corresponding normal variate

$$Z = \frac{\log_{10}(C) - \mu_X}{\sigma_X} . \qquad (4\text{-}20)$$

The exceedence probability is equal to one minus the cumulative normal probability of the variate Z.

### 4.4.3 Discussion

The two probability models represent two different approaches to statistical estimation. The discrete or binomial estimation method is a non-parametric approach, in that it does not rely on an assumed probability distribution function. The lognormal estimation method, because it depends on a particular analytical function to form its predictions, is a parametric approach. Each approach has advantages and disadvantages.

**Binomial Model.** The chief advantage of the binomial approach is that it retains the same validity no matter what underlying probability distribution describes the data. In both distributions shown in Figure 4.7, the total probability mass to the right of the vertical line—the exceedence probability for the threshold represented by the line—is equal to 0.25. Since the binomial method is based only on the total probability of exceeding the threshold, the difference in the shape of the two distributions makes no difference in the estimating procedure. The lognormal model would fit the left distribution, which is lognormal, well , but not the one on the right, which is the sum of a normal and lognormal distribution.

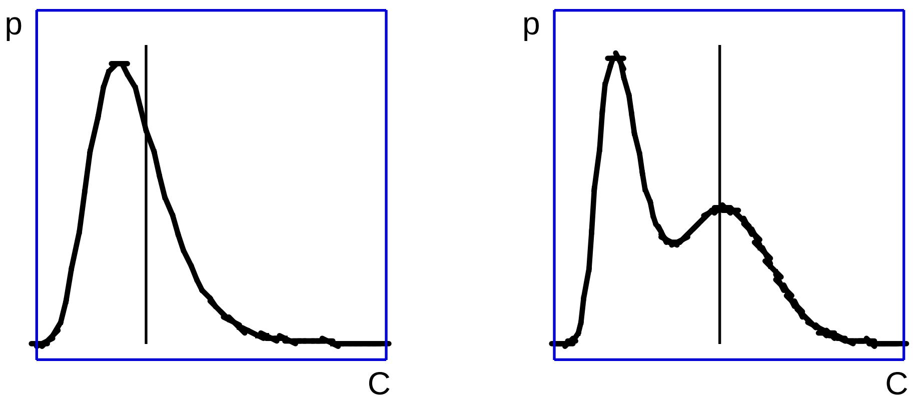


Figure 4.7 Discrete Probabilities from Continuous Distributions

The fit of the data from quadrangle 5740, shown in Figure 4.6 is typical of those examined individually in this study; the lognormal model fits well through the middle of the range of concentrations, but deviates from the data at the ends of the data. In the case of quad 5740, the model underpredicts the number of measurements with low nitrate concentrations.

The discrete model also gives meaningful confidence intervals on its estimates. More measurements produce less uncertainty in a predictable and understandable way. Although it is possible to estimate errors from the regression fitting the lognormal distribution, these describe the goodness-of-fit of the regression, and not uncertainties in the estimated probabilities. A lognormal model based on two data points will show a perfect fit, and no standard error; this has no meaning for predicting the accuracy of the model's predictions.

**Lognormal Model.** The lognormal model does offer some advantages, however. Once the parameters have been fit, it is not necessary to revisit the original data to estimate the probability of exceeding a new threshold value, as it is for the discrete model. The lognormal parameters, indicating the central tendency and spread of the data, are more informative about the range of concentrations seen in the region than the single probabilities produced by the discrete model.

The data from quad 5740 also point to a deficiency in the discrete model. The estimated exceedence probabilities for the 5 mg/l and 10 mg/l thresholds are identical, because the two measurements greater than 5 mg/l were also greater than 10 mg/l. Intuitively, we would expect a higher exceedence probability for the lower threshold. The lognormal model would fit this expectation better than the discrete model.

**Caveats.** Some limitations and warnings apply to both models. Defining exceedence probabilities on regions implies that the behavior of the whole region can be adequately characterized by that number. This would be true only if the probability of detecting an excess of the constituent were the same at every point in the region. Because the regions are inhomogeneous, this is not true. The 37 wells located in quad 5740 and included in the study draw water from the Glen Rose Limestone, from the Hosston Formation, and from the Trinity Group. The wells have depths ranging from 80 to 500 feet. Over this range of conditions, there must be significant variation in exceedence probabilities.

The exceedence probability would still characterize the region as a whole, if not every point in it, if the samples were truly randomly selected, or chosen as representative of the region.  Since the measurements are collected from existing wells, and some wells are more frequently sampled than others, even this claim is weakened.  Most of the wells in quad 5740 were sampled only once.  One was sampled twelve times.  All of the measurements from these samples were treated as equally representative of the quad.

This can be justified in part by the fact that water moves through the region, and that several measurements from a well taken at different times can represent a region around the well.  However, twelve measurements at a single location are not the same as twelve measurements at twelve locations.  No attempt was made to correct for biases introduced by the TWDB sampling schedule.

The exceedence probability estimates should not be taken as absolute predictions of exceedence rates, but should instead be viewed relative to each other.  A region with a high estimated exceedence probability is different from one with a low estimated exceedence probability, and more measurements lead to greater confidence that the difference is real.  The confidence intervals on the exceedence probability estimates cannot account for bias in the sampling scheme, but offer a set of "best case" bounds.  The true exceedence probability for the region lies between those bounds *if* the sample is representative of the region.  The data used in the study provide no basis for judging how well the regions are represented by the samples.  It is assumed that the samples are sufficiently representative that the differences from one quadrangle to another (particularly when the quads are widely separated) are more significant than the inhomogeneities within the quadrangles.

**Preferred Method.** On balance, the binomial approach to estimating exceedence probabilities seems more suited to the problem of characterizing groundwater vulnerability. The probability distribution of nitrate concentrations cannot reasonably be expected to follow the same functional form everywhere. In some cases, the lognormal distribution will fit well, in others it will fit over a limited range of concentrations. Since water quality regulations incorporate threshold concentrations in the form of maximum contaminant levels and monitoring trigger levels, it makes sense to use a method that estimates the probability of exceeding those thresholds regardless of the form of the underlying probability distribution. In the presentation of results in Chapter 6, the lognormal model is used in only one map.

## 4.5 MAPPING OF INDICATOR PARAMETERS

The soil property, precipitation, and fertilizer sales data, which are tested as indicators of vulnerability to nitrate contamination, are contained in polygon coverages in the vector GIS system of Arc/Info. The polygons—STATSGO map units, counties, and Thiessen polygons—are irregularly shaped and, with the exception of the two soil parameters derived from the STATSGO soil data set, the extents of polygons associated with one parameter do not coincide with the extents of polygons associated with any other parameter, or with the quadrangles for which exceedence probabilities have been calculated. The maps in Chapter 3 clearly illustrate this.

In order to compare the variation of the indicator parameters with the variation of the exceedence probabilities, all the indicator parameter values were re-mapped onto the quadrangles.  The discussion that follows examines the meaning of this re-mapping.

### 4.5.1  Polygons and Their Attributes

In a vector GIS, a polygon is a contiguous, bounded area on the surface of the earth.  Within a coverage or thematic layer, the boundaries of a polygon are determined by differences in the values of the attributes that express the theme.  Examples of attributes that define polygons are:  political affiliations, like counties; geological or other physical characteristics, like the soil associations in the STATSGO soil data; or arbitrary divisions along made-up boundaries, like 7.5' quadrangles.

Locating a point inside a polygon can be compared to identifying a member of a set.  If a location lies inside a given polygon, it meets the criteria that define the polygon.  Consider a Theissen network constructed around rain gauges.  For gauge number 123 there is a polygon defined as "the set of all points that are closer to gauge 123 than to any other gauge."  Attribute values may be assigned to a point (such as the location of a well) based on the attributes of the polygon in which it is located, and all points lying within the boundaries of a polygon would necessarily have the same values for the attributes assigned to them from the polygon.  If the average annual rainfall at gauge 123 is 28 inches, then the statement "The average annual rainfall at the nearest gauge is 28 inches" is true for all points in the Theissen polygon surrounding gauge 123.

A GIS polygon with an attribute value is something like a bin or a bucket with a label on it. The label applies to all the contents of the bucket, and no distinction can be made between one part of the contents and another, or one part of a polygon and another. This does not imply that such distinctions do not exist, only that they cannot be represented by the GIS without sub-dividing the polygon.

Although some statements, like the descriptions of the Theissen polygons above, are true for every point within a polygon, others apply only to the polygon as a whole. For example, polygon 123 might have an area of 25 square miles, but the statement "the area of every point in polygon 123 is 25 square miles" is meaningless.

Still other statements, while applying in a rigorous sense only to the polygon as a whole, still have some meaning for points within the polygon. This is true for average or total values calculated over a polygon. The statements "this point lies in a polygon where atrazine is applied at an average rate of 0.5 kg per square kilometer" and "atrazine is applied at an average rate of 0.5 kg per square kilometer at this point" are not equivalent. A great deal of GIS-based data is collected and reported as averages or totals over polygons. In such cases, it is necessary to approximate values at points from averages or totals over polygons, because no other data is available. This is true of most of the polygon-based data used in this study, including the exceedence probabilities calculated from the TWDB data.

### 4.5.2 Overlaying Polygons

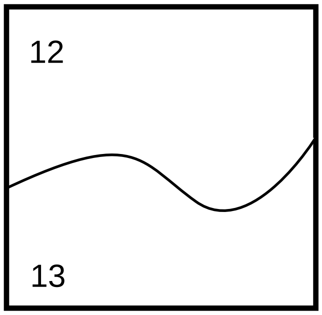
In order to study the predictive power of more than one parameter on the behavior of groundwater quality, it is necessary to combine data from several thematic layers. This process, called overlaying, is similar to constructing the intersection of sets. Overlaying two polygon coverages—for example soil polygons and Theissen polygons, as shown in Figure 4.8—preserves the boundaries of both sets of original polygons and creates a more complex set of polygons.

Combining thematic layers through polygon overlay preserves all of the information present in the original coverages, but frequently results in small, oddly shaped polygons. It would be possible to overlay all the polygons associated with the indicator parameters, group wells and nitrate measurements according to location in the resulting polygons, and calculate statistics on those groups, as was done in the 7.5' quadrangles. The irregularity and highly variable size of the resulting polygons, however, makes comparisons between them difficult. An alternative method partitions the location space into uniform pieces and interpolates attribute data onto the resulting partitions.

### 4.5.3 Raster Cells and Attributes

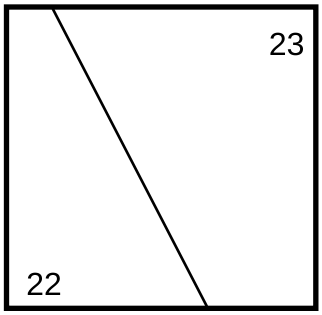
Like a polygon, a raster or grid cell is a contiguous bounded area with associated data. Unlike a polygon, its boundaries are determined by a regular pattern, like a checkerboard, not by changes in the data values associated with it. Rasters are frequently used to express continuously varying quantities. Rasters approximate continuous variation as a series of discrete steps.

Soil



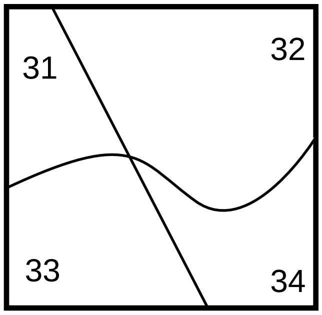
| Soil-id | Thick |
|---|---|
| 12 | 45 |
| 13 | 12 |

Precipitation



| Prec-id | AAPrec |
|---|---|
| 22 | 28 |
| 23 | 30 |

Overlay



| Over-id | Soil-id | Prec-id | Thick | AAPrec |
|---|---|---|---|---|
| 31 | 12 | 22 | 45 | 28 |
| 32 | 12 | 23 | 45 | 30 |
| 33 | 13 | 22 | 12 | 28 |
| 34 | 13 | 23 | 12 | 30 |

**Figure 4.8 Polygon Overlay**

The single value associated with a raster cell can be an average value or a dominant (maximum, maximum area, maximum weight) value over the cell's area  For continuously varying data this is plainly an approximation, but a tolerable one if the area of an individual cell is small enough that the variations within an individual cell are small compared to the range of variation over the area mapped by the whole grid.

The great advantage of rasters over polygons is that when thematic layers are combined, the spatial structure remains unchanged, because the grid of cell boundaries is the same in each layer.   No irregular fragments are formed when raster layers are combined.

If the surface data and the exceedence probability estimates are all represented on a common grid, then linking probability values to indicator values becomes a matter of extracting several attribute values for a single grid cell, which is a trivial GIS operation.  The limitations of raster GIS, however, make resolving the probabilities and the surface data to a common grid difficult.

The most serious limitation of the raster system is its limited representation of topology.  All data in a raster GIS consists of cells.  A point can be represented approximately by a single cell, a line by a series of adjacent cells, and a polygon by a cluster of cells, but spatial concepts like the location of a point in a polygon cannot be represented in a raster GIS.  Since wells and nitrate measurements were grouped by location within 7.5' quadrangles for this study, this limitation needs to be overcome before all the data can be represented in a common grid.

### 4.5.4 Rasterized Polygons: A Compromise

In order to preserve the point-and-polygon topology necessary to group the nitrate data for statistical analysis, and to allow surface data (rainfall amounts, soil parameters, and fertilizer application) to be compared consistently with the variation of the exceedence probability estimates, a compromise was developed.

The polygon coverage used to group the wells and nitrate readings was overlaid on each of the indicator parameter coverages, resulting in a highly fragmented polygon coverage. Each fragment, however, was associated with the original polygons that formed the overlay through the coverage's attribute table. It was possible to calculate an area-weighted average for the parameter values for each quadrangle by grouping the fragments according to the quadrangle IDs in their attribute tables. The averages could then be linked to the quadrangle coverage, along with the exceedence probability estimates. The steps required to carry out this averaging and linking are described in section 5.7. Figure 4.9 illustrates the process of resolving the exceedence probabilities and the indicator parameters to a common grid. By using the quadrangle numbers as a key to link the tables containing the parameter averages and the exceedence probabilities, it is possible to form a single table containing exceedence probabilities and indicator parameter values for each quadrangle.

The contents of this table can then be linked to the quadrangle coverage and used to map the variation of the exceedence probabilities or the values of the indicator parameters over the quadrangles. The values of the exceedence probabilities and the indicator parameters can also be written to an external file, and used as input to a regression analysis.

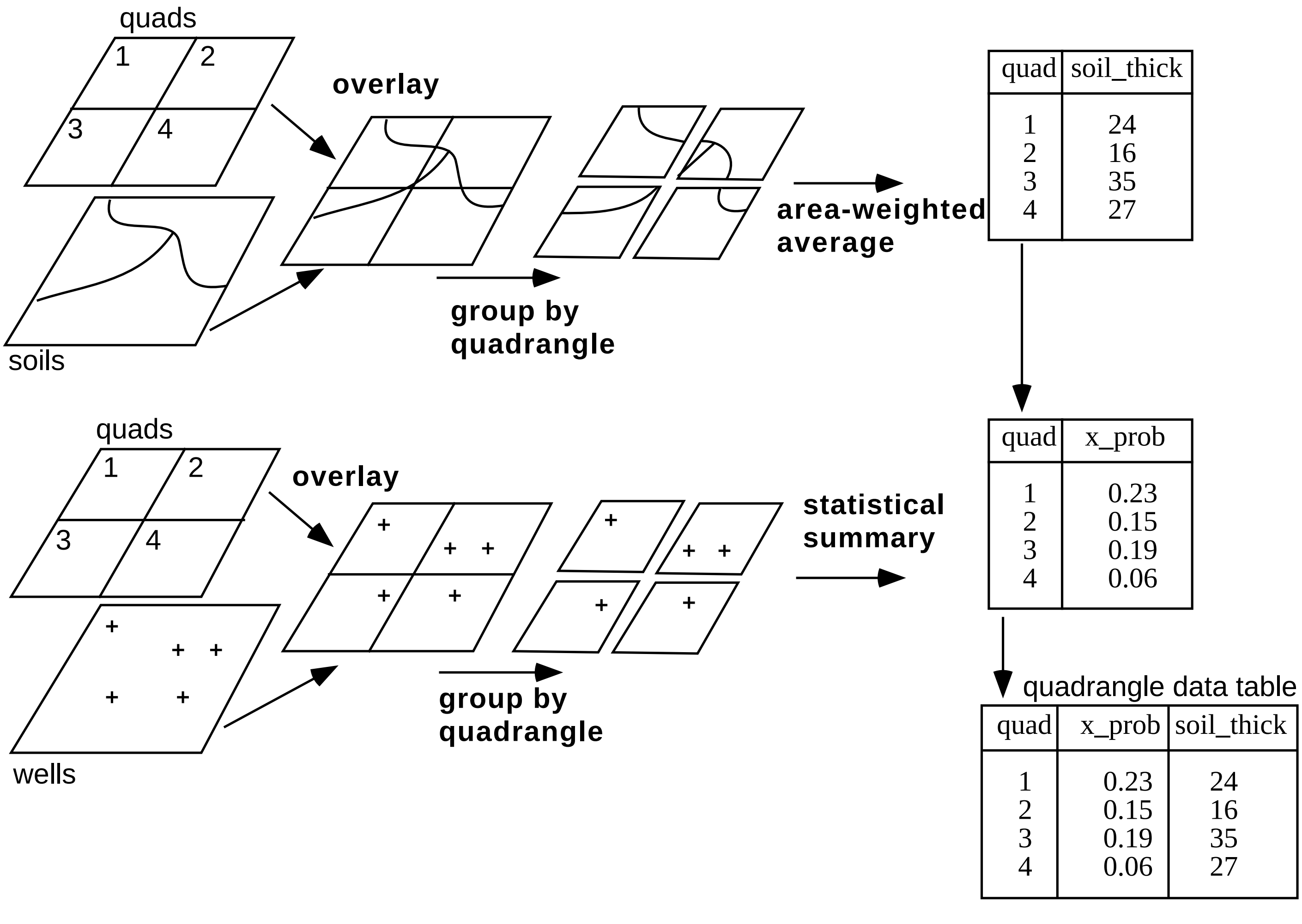
## 4.6 REGRESSION ON INDICATORS

Once the indicators and the exceedence probabilities have been linked to a common grid, values for all these data can be tabulated and used independently of their spatial relationships. The values of exceedence probability, average precipitation, soil thickness, etc. can be treated as a dependent variable (exceedence probability) and a series of independent variables (precipitation, etc.) in a multiple linear regression to produce a model of the form

$$P = ß_0 + ß_1X_1 + ß_2X_2 + ß_3X_3 + ...$$

where each $ß_n$ is a found by fitting the values of P and the various $X_n$s.

The regression method used in this study to quantify the relationship between the indicators and the exceedence probabilities is stepwise multiple regression. In this procedure, variables are added to or deleted from the model one at a time according to the significance of their coefficients, as measured by the partial and sequential F statistics (Draper and Smith, 1981). In this work, an F statistic of 4, indicating a 95% probability that the coefficient differs from zero, was used as the inclusion criterion.



| quad | soil_thick |
|---|---|
| 1 | 24 |
| 2 | 16 |
| 3 | 35 |
| 4 | 27 |

| quad | x_prob |
|---|---|
| 1 | 0.23 |
| 2 | 0.15 |
| 3 | 0.19 |
| 4 | 0.06 |

quadrangle data table

| quad | x_prob | soil_thick |
|---|---|---|
| 1 | 0.23 | 24 |
| 2 | 0.15 | 16 |
| 3 | 0.19 | 35 |
| 4 | 0.06 | 27 |

**Figure 4.9 Resolving Indicators and Exceedence Probabilities to a Common Grid**

## 4.7 Confirming Assumptions

To generalize the present study, two assumptions must be confirmed. The first is that the historic database used to form the exceedence probability estimates is sufficiently typical of groundwater in Texas that those estimates can predict where nitrate contamination is likely to occur. The second is that vulnerability to nitrate contamination is related to contamination by other constituents, specifically agricultural chemicals.

To test these assumptions, two additional data sets were included in the study.

Nitrate measurements collected by the Water Utilities Division of the Texas Natural Resource Conservation Commission from public water supplies over a period of just under two years (in 1993 and 1994) are compared to results of the analysis of the TWDB database for the years 1962–1993. This comparison tests whether water in public supplies differs significantly from the general sampling conducted by TWDB, and whether changes in the occurrence of nitrate in groundwater over time make the more recent WUD data different from the thirty years of TWDB data.

A completely independent data set, collected by the U.S. Geological Survey in the midwestern U.S. (samples from North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri, Michigan, Wisconsin, Illinois, Indiana, and Ohio), is used to test the assumption that the same conditions leading to high vulnerability to nitrate also lead to vulnerability to other contaminants. The methods used to analyze theses data sets are described in the discussion of procedures and results in Chapters 5 and 6.