CRWR Online Report 95-4

A SPATIAL AND STATISTICAL ASSESSMENT OF THE VULNERABILITY OF TEXAS GROUNDWATER TO NITRATE CONTAMINATION

by

Thomas Anders Evans, B.S.M.E., M.S.E.

and

David R. Maidment, Ph.D.

Principal Investigator

November 1995

CENTER FOR RESEARCH IN WATER RESOURCES

Bureau of Engineering Research .The University of Texas at Austin

J.J.Pickle Research Campus . Austin, TX 78712-4497

This document is available online via World Wide Web at

http://civil.ce.utexas.edu/centers/crwr/reports/online.html

ACKNOWLEDGMENTS

The work presented in this dissertation was supported by the United States Environmental Protection Agency, through agreement number CR818116, administered by the Environmental Monitoring Systems Laboratory. Grateful acknowledgment and credit for the data analyzed here are due to the staffs of the Texas Water Development Board, the Water Utilities Division of the Texas Natural Resource Conservation Commission, the US Geological Survey, and the Soil Conservation Service of the US Department of Agriculture. Robert Blodgett, Roger Quincy, and William Battaglin deserve individual recognition, as do a number of others, who, alas, will not receive it here.

This work could not have been completed without the support of my dissertation committee, Philip Bennett, Randall Charbeneau, Raymond Loehr, Daene McKinney, and especially Professor David Maidment, who paid the bills.

Finally, the author wishes to thank the Burgin family for life support, and Brenda Nelson and Sheri Whitley for actions above and beyond the call of their job descriptions.

ABSTACT

Nitrate concentrations in approximately 46,000 water samples from Texas wells are analyzed using spatial and statistical representations on a grid of 7.5' quadrangles. In each quadrangle containing at least 12 measurements, the probabilities of exceeding four threshold concentrations probabilities of exceeding four concentration thresholds of nitrate (0.1, 1.0, 5.0, and 10.0 mg/l nitrate as nitrogen) are estimated as the ratio of observed exceedences to the total number of measurements. An alternative probability analysis using the lognormal distribution yields exceedence probabilities that show some systematic difference from those computed directly from the data.

Five representative aquifers were chosen for additional analysis. Nitrate exceedence probabilities are relatively uniform within aquifers, but differ significantly from one aquifer to another. The exceedence probability for the 1 mg/l threshold was selected as best representing vulnerability to nitrate contamination. The five aquifers, ranked from lowest to highest vulnerability by this criterion are: Carrizo-Wilcox, Edwards (Balcones Fault Zone), Hueco-Mesilla Bolson, Ogallala, and Seymour. Evidence suggests that nitrate levels are increasing across the state, and in the Ogallala in particular, but such trends are not consistent across aquifers.

Linear regression was used to assess the relationship between nitrate exceedence probabilities potential indicator parameters. The dominant parameter is the aquifer from which the sample is drawn. Setting this aside, the only consistently significant indicator is average annual rainfall: groundwater is more likely to be contaminated in regions where rainfall is low than in regions

where rainfall is hi gh. No significant relation between the spatial patterns of nitrate contamination and the sale of nitrogen fertilizers was found.

Table of Contents

List of Figures....9

List of Tables....13

Chapter 1: Introduction....16

1.1 Motivation....18

1.2 Objectives....25

1.3 Scope of Study....26

1.4 Project Summary....26

1.5 Contributions of Study....27

1.6 Outline of Dissertation....28

Chapter 2: Background and Literature Review....30

2.1 Uses for Groundwater Vulnerability Assessment....30

2.2 Groundwater Vulnerability Assessment methods....32

2.3 Statistical Groundwater Vulnerability Assessment....37

2.4 Choice of Method....41

2.5 Nitrate in Groundwater....44

2.6 Outline of Proposed Vulnerability Analysis Method....49

2.6.1 Comparison of Method with Previous Studies....50

2.6.2 Application in Present Work....53

Chapter 3: Data Sources and Description....55

3.1 Nitrate Measurement Data....56

3.1.1 Nitrate Reporting Limits....58

3.1.2 Sampling Period....60

3.1.3 Measurement Record Accuracy....61

3.2 Well Data....63

3.2.1 TWDB Well Numbers....63

3.2.2 Location Accuracy 67
3.2.3 Selected Aquifers 68
3.2.4 Well Description Accuracy 72
3.3 Soil Data 74
3.3.1 STATSGO Map and Data Structure 74
3.3.2 Using STATSGO Data 80
3.3.3 Range and Distribution of Soil Parameter Values 88
3.4 Precipitation Data 93
3.4.1 Preparation of Annual Average Precipitation Map 93
3.4.2 Range and Distribution of Precipitation Data 96
3.5 Fertilizer Sales Data 97
3.6 Water Utilities Division Nitrate Monitoring Data 100
3.7 Herbicide and Nitrate Data from Midwestern U.S. 101

Chapter 4: Methods 103
4.1 Nitrate as a Surrogate for Vulnerability 104
4.2 Identification of Analysis Regions 107
4.3 Grouping Data for Analysis 113
4.3.1 Database Management Systems 113
4.3.2 Geographic Information Systems 119
4.4 Statistical Model of Vulnerability 124
4.4.1 Discrete Probability Estimates 127
4.4.2 Lognormal Probability Estimates 137
4.4.3 Discussion 143
4.5 Mapping of Indicator Parameters 147
4.5.1 Polygons and Their Attributes 148
4.5.2 Overlaying Polygons 150
4.5.3 Raster Cells and Attributes 150
4.5.4 Rasterized Polygons: A Compromise 153
4.6 Regression on Indicators 154

4.7 Confirming Assumptions 156

Chapter 5: Procedures 158

5.1 Important Features of INFO and TABLES 158

5.2 Data Entry 164

5.2.1 TWDB Well and Nitrate Data 164

5.2.2 Soil Data 169

5.2.3 Precipitation Data 169

5.2.4 Fertilizer Sales Data 172

5.2.5 Water Utilities Division Data 172

5.2.6 Midwest Herbicide and Nitrate Data 174

5.2.7 A Note on Map Projections 175

5.3 Calculation of Data Derived from STATSGO 175

5.4 Preparation of Quadrangle Maps 179

5.5 Calculation of Statistics 181

5.5.1 Discrete Probability Estimates 182

5.5.2 Lognormal Probability Estimates 184

5.6 Map Preparation 186

5.7 Indicator Values and Stepwise Linear Regression 187

5.8 Exceedence Frequencies from WUD Data Set 189

Chapter 6: Results 193

6.1 Statewide Results 194

6.2 Selected Aquifers 219

6.2.1 Carrizo-Wilcox Aquifer 222

6.2.2 Edwards Aquifer (Balcones Fault Zone) 226

6.2.3 Hueco-Mesilla Bolson Aquifer 230

6.2.4 Ogallala Aquifer 234

6.2.5 Seymour Aquifer 239

6.2.6 Aquifer Summary 243

6.3 Indicators and Regression 249

6.4 Comparison with Water Utilities Division Data 259
6.5 Nitrate and Herbicides in Midwest Data Set 265
6.6 Summary 268

Chapter 7: Conclusions 271
7.1 Vulnerability Assessment Method Summary 271
7.2 Results in Texas 279
7.3 Recommendations 285

References 293

Appendix A: Database Definitions 298

Appendix B: Computer Programs 313

Vita 358

List of Figures

Figure 1.1 Nitrate-Related Ranking of Texas Counties....24
Figure 2.1 Simplified Biological Nitrogen Cycle....47
Figure 3.1 Reported Detection Limits for Nitrate....59
Figure 3.2 Reported Nitrate Concentrations....60
Figure 3.3 Nitrate Measurements Reported by Year....61
Figure 3.4 TWDB Well-Numbering System....66
Figure 3.5 TWDB Location of Study Aquifers....70
Figure 3.6 Well Depths (less than 200 feet)....73
Figure 3.7 Map Unit Area Histogram....76
Figure 3.8 STATSGO Map and Data Organization....79
Figure 3.9 STATSGO Map Units....81
Figure 3.10 Map Unit TX071 with Quadrangles for Size Comparison....86
Figure 3.11 Soil Depth Histogram....89
Figure 3.12 Spatial Distribution of Soil Thickness....90
Figure 3.13 Soil Organic Matter Histogram....91
Figure 3.14 Spatial Distribution of Soil Organic Content....92
Figure 3.15 Spatial Distribution of Annual Average Precipitation....95
Figure 3.16 Thiessen Polygon Area Histogram....96
Figure 3.17 Precipitation Histogram....97
Figure 3.18 Sales of Nitrogen Fertilizers by County....99
Figure 4.1 Measurement Histograms for 2.5' and 7.5' Quadrangles....111
Figure 4.2 Well Locations in Quadrangle 5740....112

Figure 4.3 Locating Points in Polygons 122
Figure 4.4 Estimating Probability of Success from a Sample of 12 Trials 134
Figure 4.5 Effects of Different Inclusion Criteria..... 135
Figure 4.6 Fitting a Probability Distribution by Regression for Quadrangle 5740 142
Figure 4.7 Discrete Probabilities from Continuous Distributions 144
Figure 4.8 Polygon Overlay 151
Figure 4.9 Resolving Indicators and Exceedence Probabilities to a Common Grid..... 155
Figure 5.1 Example of a RELATE..... 160
Figure 5.2 Using a RELATE to Count Measurements in Wells..... 162
Figure 5.3 One INFO Item Redefined into Four Items..... 164
Figure 5.4 Simplified Quadrangle/Map Unit Intersection..... 188
Figure 6.1 Nitrate Detection Histogram..... 197
Figure 6.2 Spatial Distribution of Nitrate Detection Probabilities (Binomial Estimate)..... 198
Figure 6.3 Nitrate 1 mg/l Exceedence Histogram 200
Figure 6.4 Spatial Distribution of 1 mg/l Exceedence Probabilities (Binomial Estimate)..... 201
Figure 6.5 Nitrate 5 mg/l Exceedence Histogram 203
Figure 6.6 Spatial Distribution of 5 mg/l Exceedence Probabilities (Binomial Estimate)..... 204
Figure 6.7 Nitrate 10 mg/l Exceedence Histogram 206

Figure 6.8 Spatial Distribution of 10 mg/l Exceedence Probabilities (Binomial Estimate) 207
Figure 6.9 Spatial Distribution of Upper 95% Confidence Limit on 1 mg/l Exceedence Probabilities (Binomial Estimate) 209
Figure 6.10 Spatial Distribution of 1 mg/l Exceedence Probabilities (Lognormal Estimate) 211
Figure 6.11 Comparison of Discrete and Lognormal Exceedence Probability Estimates 212
Figure 6.12 Variation of Exceedence Probabilities with Depth 215
Figure 6.13 Measurements by Year in Base Data Set 216
Figure 6.14 Variation of Exceedence Probabilities Over Time 217
Figure 6.15 Regression of 1 mg/l Exceedence Probability Against Time 218
Figure 6.16 7.5' Quadrangles Associated with Study Aquifers 220
Figure 6.17 Spatial Distribution of Nitrate Exceedence Probabilities in the Carrizo-Wilcox Aquifer 223
Figure 6.18 Variation of Exceedence Probabilities with Depth in the Carrizo-Wilcox Aquifer 224
Figure 6.19 Variation of Exceedence Probabilities Over Time in the Carrizo-Wilcox Aquifer 225
Figure 6.20 Spatial Distribution of Nitrate Exceedence Probabilities in the Balcones Fault Zone of the Edwards Aquifer 227
Figure 6.21 Variation of Exceedence Probabilities with Depth in the Balcones Fault Zone of the Edwards Aquifer 228

Figure 6.22 Variation of Exceedence Probabilities over Time in the Balcones Fault Zone of the Edwards Aquifer 228
Figure 6.23 Spatial Distribution of Nitrate Exceedence Probabilities in the Hueco-Mesilla Bolson Aquifer 231
Figure 6.24 Variation of Exceedence Probabilities with Depth in the Hueco-Mesilla Bolson Aquifer 232
Figure 6.25 Variation of Exceedence Probabilities over Time in the Hueco-Mesilla Bolson Aquifer 233
Figure 6.26 Spatial Distribution of Nitrate Exceedence Probabilities in the Ogallala Aquifer 235
Figure 6.27 Variation of Exceedence Probabilities with Depth in the Ogallala Aquifer 236
Figure 6.28 Measurements by Year in Ogallala Aquifer 237
Figure 6.29 Variation of Exceedence Probabilities in the Ogallala Aquifer Over Time 238
Figure 6.30 Spatial Distribution of Nitrate Exceedence Probabilities in the Seymour Aquifer 240
Figure 6.31 Variation of Exceedence Probabilities with Depth in the Seymour Aquifer 241
Figure 6.32 Variation of Exceedence Probabilities over Time in the Seymour Aquifer 242
Figure 6.33 Comparison of Exceedence Probabilities Statewide and in Five Study Aquifers 244

Figure 6.34 Histograms of Nitrate Concentrations in Texas and Five Study Aquifers 246
Figure 6.35 Boxplots of Well Depths and Nitrate Concentrations in Five Study Aquifers 248
Figure 6.36 Quad-by-Quad Comparison of Estimated 0.1 mg/l Exceedence Probabilities with WUD Nitrate Measurements 260
Figure 6.37 Aggregated Comparison of Estimated 0.1 mg/l Exceedence Probabilities with WUD Nitrate Measurements 261
Figure 6.38 Quad-by-Quad Comparison of Estimated 1 mg/l Exceedence Probabilities with WUD Nitrate Measurements 262
Figure 6.39 Aggregated Comparison of Estimated 1 mg/l Exceedence Probabilities with WUD Nitrate Measurements 262
Figure 6.40 Quad-by-Quad Comparison of Estimated 5 mg/l Exceedence Probabilities with WUD Nitrate Measurements 263
Figure 6.41 Aggregated Comparison of Estimated 5 mg/l Exceedence Probabilities with WUD Nitrate Measurements 263
Figure 6.42 Quad-by-Quad Comparison of Estimated 10 mg/l Exceedence Probabilities with WUD Nitrate Measurements 264
Figure 6.43 Aggregated Comparison of Estimated 10 mg/l Exceedence Probabilities with WUD Nitrate Measurements 264
Figure 6.44 Herbicide and Nitrate Measurements Grouped by Geologic Parameters 267

List of Tables

Table 3.1 Nitrate Measurement Data 57
Table 3.2 Nitrate Measurements from Well 5740304 57
Table 3.3 Excluded Measurement Records 62
Table 3.4 Well Description Data 64
Table 3.5 Location Accuracy Codes 67
Table 3.6 Aquifer Codes 69
Table 3.7 Wells and Measurements in Selected Aquifers 69
Table 3.8 Soil Series in Map Unit TX071 "Brackett-Purves-Real" 82
Table 3.9 Layers in Purves Component of Map Unit TX071 82
Table 3.10 Derived Values for Soil Organic Content in Purves Series (Map Unit TX071) 84
Table 3.11 Soil Series Parameters for Map Unit TX071 84
Table 4.1 Predicate Symbols for Relational Calculus 115
Table 4.2 Results of Query 4-1 116
Table 4.3 Nitrate Concentrations in Quadrangle 5740 131
Table 4.4 Estimated Exceedence Probabilities for Quadrangle 5740 132
Table 4.5 Data for Lognormal Fit to Quadrangle 5740 141
Table 5.1 Codes for Study Aquifer Identification 169
Table 5.2 CMY Components of Probability Map Colors 187
Table 6.1 Nitrate Exceedences in Texas 195
Table 6.2 Probability Plot Correlation Coefficient Test Results 213
Table 6.3 Sign Test Results 214

Table 6.4 Regression Results For Threshold Exceedences through Time 218
Table 6.5 Wells and Measurements in Selected Aquifers 221
Table 6.6 Nitrate Exceedences in the Carrizo-Wilcox Aquifer 222
Table 6.7 Nitrate Exceedences in the Balcones Fault Zone of the Edwards Aquifer 226
Table 6.8 Nitrate Exceedences in the Hueco-Mesilla Bolson Aquifer 230
Table 6.9 Nitrate Exceedences in the Ogallala Aquifer 234
Table 6.10 Regression Results For Threshold Exceedences through Time in the Ogallala Aquifer 238
Table 6.11 Nitrate Exceedences in the Seymour Aquifer 239
Table 6.12 Regression Results for Quads with 12 or More Measurements 252
Table 6.13 Regression Results for Quads Associated with Study Aquifers 254
Table 6.14 Regression Results for Quads Associated with Study Aquifers, Including Dummy Variables for Aquifers 256
Table 6.15 Regression Results for 1 mg/l threshold within Study Aquifers 258
Table 6.16 Aquifer Categories Ranked by Nitrate and Herbicide Detection Rates 267
Table 7.1 Exceedence Probability Summary 284

Chapter 1: Introduction

The objective of the study presented in this dissertation is to develop a methodology for assessing the vulnerability of groundwater to contamination by agricultural chemicals. Federal water quality regulations have created a need for such assessments, but do not specify the methods to be used, or rigorously define groundwater vulnerability. The present work advocates a statistical approach to vulnerability assessment, and, in keeping with that approach, suggests that *probability of contamination*, a quantity that can be expressed numerically, be used as a surrogate for *vulnerability*, which remains a rather nebulous and unquantifiable commodity.

In this work, the words *susceptibility*, *vulnerability*, and *risk* represent related, but distinct, ideas. 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. Risk, unlike susceptibility and vulnerability can be described by a number. In other words, risk is quantifiable, while susceptibility and vulnerability are not.

In addition to risk of contamination, there are other risks associated with groundwater quality: risk of human exposure to the contaminant, risk of adverse public health effects, and so on. Although these risks are important to the

formation of public policy, they lie beyond the scope of this study, which is concerned solely with the likelihood that a contaminant is present in a groundwater supply.

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

This work proposes that an answer to this question lies in the following postulate: For any body of groundwater and any chemical constituent, there exists a probability distribution function, $P(C_t)$, describing the likelihood that a sample, chosen at random from that body, will contain a concentration of the constituent greater than a threshold concentration, C_t. While this concentration probability distribution is not identical to risk of contamination, susceptibility, or vulnerability, it is closely related to all three, and is both quantifiable and measurable, to the extent that its parameters can be estimated from measurements of concentration taken from the groundwater body.

A body of groundwater contains an infinite number of potential water samples—a *population*, in statistical argot—the concentration probability distribution $P(C_t)$, describes that population. Actual measurements of constituent concentrations in this body of groundwater make up a *sample* of that population. Properties (called *parameters*) of the concentration probability distribution can

be estimated by calculations performed on the sample. These estimates are called *statistics*. The methodology advocated here uses statistics calculated from groups of groundwater quality (i.e. constituent concentration) measurements as surrogates for risk of contamination, which cannot be measured, and for susceptibility and vulnerability, which cannot be quantified.

The particular results presented here form a spatial and statistical study of the presence of nitrate in groundwater in Texas. This work analyses nitrate measurements collected throughout the State from 1962 to 1993 and recorded in the Texas Water Development Board's Ground Water Data System (Nordstrom and Quincy, 1992), and uses statistical methods in conjunction with a geographic information system (GIS) and a relational database management system to organize the data and form conclusions.

Although the present work was directed toward the vulnerability of groundwater to agricultural chemicals, of which nitrate is a widely measured representative, the methods developed in the course of this study are not specific to agricultural chemicals or to groundwater. The same approach could easily be applied to industrial contamination of air, or any number of other forms of pollution.

1.1 MOTIVATION

This impetus for this study comes from the National Primary Drinking Water Regulations (40 CFR 141), which took effect in January 1993. These regulations implement provisions of the revised Safe Drinking Water Act by listing 60 maximum contaminant levels (MCLs) for constituents that must be

monitored by operators of public water supplies, and imposing schedules for monitoring those constituents. Earlier regulations listed only 34 MCLs, so the costs of monitoring have increased significantly, especially since most of the additional MCLs are for organic chemicals such as industrial solvents, like toluene and trichloroethylene, and pesticides, like atrazine and alachlor, which require more expensive analytical methods than inorganic or nutrient constituents. To reduce the financial burden on regulators and water systems, the regulations allow the State agencies responsible for enforcement of the Safe Drinking Water Act to waive some monitoring requirements for a number constituents, including several agricultural chemicals, in water systems that have been shown, over several monitoring cycles, to be free from contamination from those constituents.

Waivers may also be granted to systems that have been shown, through a vulnerability assessment, to be secure from contamination. The choice of vulnerability assessment method is left to the State, subject to approval by the Environmental Protection Agency (EPA), but must include either sufficient knowledge of previous use of the constituent in regions contributing to the water supply that the State can be sure that no source of the constituent is present, or evidence that the water supply is protected by soil or geological conditions, and the structure of the well.

The Water Utilities Division of the Texas Natural Resource Conservation Commission is responsible for enforcement of the National Primary Drinking Water Regulations in Texas. The Water Utilities Division is engaged in an

ongoing effort to record the locations and descriptions of public water supply wells in Texas in a geographic information system (GIS), in part to facilitate the analysis of wells and their surroundings for the purpose of granting monitoring waivers.

The original purpose of this study was to devise an automated system for vulnerability analysis using the Water Utilities Division's GIS data.

It soon became apparent that the data that was available in Statewide GIS coverages and databases was not adequate to form the basis of a vulnerability assessment system. In particular, hydrogeologic information such as aquifer composition, degree of confinement, and groundwater flow direction do not exist in GIS form for the State as a whole. In the absence of such data, the study focused on evaluating the usefulness of the data that *is* available in GIS for predicting groundwater vulnerability, and developing a method for deriving a statistically based groundwater vulnerability assessment method from existing groundwater quality measurements.

Concentration Thresholds. Laws like the Safe Drinking Water Act and related regulations set thresholds to trigger regulatory action, so the likelihood of exceeding thresholds is of more practical value as an measure of vulnerability than other statistical measures such as average concentrations. This study explores the use of exceedence probability as a measure of vulnerability.

Databases and Geographic Information Systems. Data management technologies, such as GIS, will play an increasingly large role in forming environmental policy and EPA has identified GIS as an important technology for

groundwater protection. The "Ground-Water Data Management Summary and Recommendations" chapter of the 1991 final report of EPA's Ground-Water Task Force states that

> GIS is an emerging tool for cross-media planning and integrated environmental management, and base program activities such as permitting, inspection, and enforcement. In addition, it is particularly useful in risk-based priority setting of Regional program commitments and resource requirements. GIS has been found to be increasingly useful in program planning and priority setting activities, once the investment in area-specific mapping has been accomplished. As EPA begins using GIS in its decision making, it is also important to begin promoting the use of GIS by the State's [*sic*] in their decision making process. (USEPA, 1991)

Data Stockpiles. Government agencies have collected and stored huge amounts of environmental data. GIS and database management systems offer a means for manipulating and analyzing this data *en masse*. This study attempts to address questions like "What benefits do this mass of data offer?" "What additional value does GIS give to that data?" "What are the shortcomings of publicly available data sets, and how can they be improved for future use?"

Spatial Patterns of Nitrate in Texas Groundwater. Figure 1.1 illustrates some of the essential points of the methods developed in this study. The figure shows three maps of the 254 counties of Texas. In each map, the counties are collected in groups containing one-fifth (20% or 51) of the counties, based on the level of a nitrate-related value defined for all counties. For , for example, a well is considered "vulnerable" if the Texas Water Development Board's groundwater database shows that a nitrate concentration in excess of the MCL of 10 mg/l has been detected there. The counties are ranked by the proportion of vulnerable

wells to the total number of wells listed for that county in the database. The 50 counties with the highest proportion of vulnerable wells are shaded red. The next-highest 51 counties are shaded orange, and so on. The resulting ranking of the counties can be used as an estimate of the relative vulnerability of the groundwater supplies in those counties to contamination by nitrates.

This estimate of vulnerability can then be compared to a candidate indicator, such as nitrogen fertilizer sales, to test the value of that candidate for predicting groundwater vulnerability.

The figures rank the counties according to:

Figure 1a: The proportion of wells where nitrate concentrations above 10 mg/l (as Nitrogen) have been detected

Figure 1b: The proportion of wells where nitrate concentrations above 1 mg/l (as Nitrogen) have been detected

Figure 1c: Nitrogen fertilizer sales during the years 1986–1991

The data sources for the three maps are described in Chapter 3.

The figures show some clear patterns, some of which run counter to intuition. A striking contrast can be seen between the fertilizer sales and the appearance of nitrate in groundwater. The belt of high fertilizer sales in east Texas does not appear to create a corresponding high level of nitrate in groundwater. In fact, the region with the highest rate of nitrate concentrations in excess of 10 mg/l (the MCL for nitrate) lies northwest of Dallas, spatially separate from the regions of highest fertilizer use. Fertilizer sales figures appear to have relatively little value as an indicator of the likelihood of finding

groundwater nitrate concentrations in excess of either of the two thresholds considered in Figure 1.1. The figure does, however, show large-scale regional variation in frequency of elevated nitrate concentrations and different patterns, which suggests that data with coarse spatial resolution can have some value as indicators.

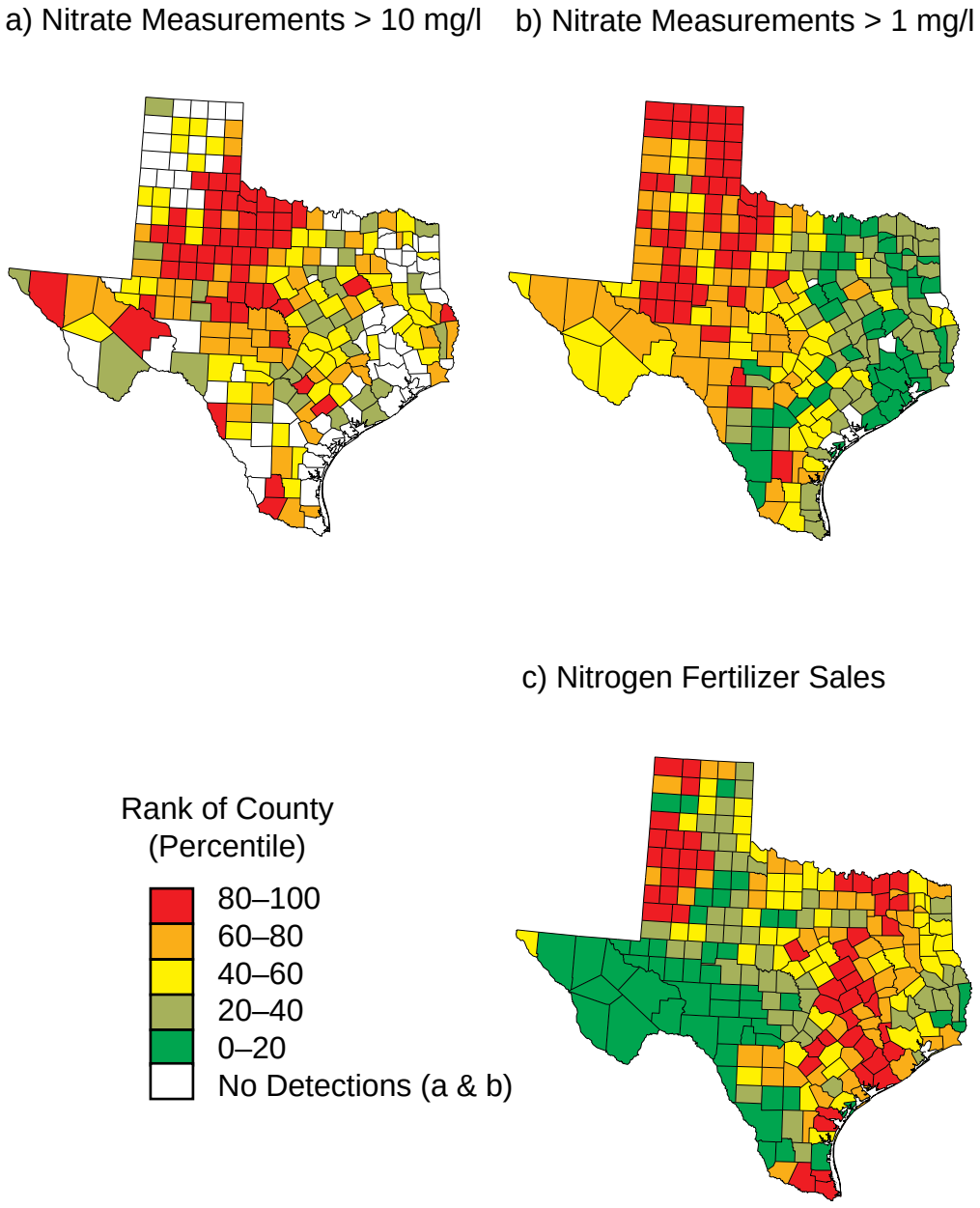


Figure 1.1 Nitrate-Related Ranking of Texas Counties

1.2 OBJECTIVES

At the time this research was proposed, the intended objective was to develop an automated system for granting vulnerability-based waivers for water quality monitoring under the National Primary Drinking Water Regulations. Because those waivers require a high degree of certainty in identifying regions that are *not* vulnerable, and because of a lack of statewide geologic data in GIS form, this goal was found to be impractical. The focus of the study shifted to improving vulnerability assessment methods using available data.

The objectives of this study are:

1) To formulate a spatially variable statistical model capable of representing in a compact form the information contained in tens of thousands of water-quality measurements spread over an area the size of Texas (691,000 km^2).
2) To apply this model in identifying spatial patterns of nitrate detection in Texas as a whole and in five selected major aquifers.
3) To estimate the relative importance of a small number of indicators—soil conditions, precipitation rates, fertilizer sales—in predicting the likelihood of contamination of groundwater by agricultural chemicals.
4) To evaluate the usefulness of a geographic information system and a database management system in carrying out an empirical study based on historic data.
5) To evaluate the usefulness of publicly available, computerized environmental data for estimating the vulnerability of groundwater to contamination.

1.3 SCOPE OF STUDY

The following limitations define the scope of this study.

1) The analysis of nitrate concentrations is restricted to data in the Texas Water Development Board's Groundwater Data System. This limits the study area to Texas and provides a single, consistent source for well descriptions and nitrate concentration data.

2) All the data used in this study comes from databases maintained by government agencies and available on a Statewide basis. This excludes, for example, data collected for studies of single aquifers or groundwater systems, unless they have been incorporated into Statewide databases. For example, maps of dominant groundwater flow direction, which exist for some aquifers, are not used because this data is unavailable over most of the State.

1.4 PROJECT SUMMARY

The study can be divided into the following three major steps.

1) Define bodies of groundwater for this study and sort measurements of water quality by their association with these bodies. Two types of definition are used. The first, based purely on location, divides Texas into seven-and-a-half minute (7.5') quadrangles, and defines a distinct body of water for each quadrangle. A measurement is associated with a given quadrangle if the well from which it was collected is located in that quadrangle. The second set of groundwater bodies is composed of five aquifers selected from the Texas Water Development Board's map of major and minor

aquifers (Ashworth and Flores, 1991). A measurement is associated with a given aquifer if the well from which it was collected draws water from a hydrogeologic unit associated with that aquifer.

2) Calculate statistical estimates of nitrate concentration probability distributions associated with the bodies of groundwater. Both discrete probabilities (estimates of the probability that various nitrate levels will be exceeded) and continuous distributions (estimates of the parameters of a probability density function) are calculated for the groundwater bodies identified in step 1.

3) Relate the statistics calculated in step 2 to indicator variables. Potential indicators of water quality considered in this study are: average annual precipitation, average soil thickness, average soil organic content, and estimated nitrate fertilizer application rate. These indicators were chosen as candidates because they were readily available, and could be plausibly linked to the degree of vulnerability of the groundwater in a region. The variation in the chosen indicators will be compared with the variation in the statistics using stepwise linear regression.

1.5 CONTRIBUTIONS OF STUDY

The study makes the following contributions to knowledge and understanding of groundwater vulnerability analysis:

1) The formulation of a spatially varying statistical mode from which exceedence probabilities (estimates of the likelihood that a constituent

will be found in concentrations exceeding a selected threshold value) can be calculated as a quantifiable measure of groundwater vulnerability.

2) The development of a quantitative, statistical method for assessing the relative value of indicators of groundwater vulnerability, and a demonstration of this method with a small number of potential indicators.

3) Application of the above to a large body of data drawn from a diversity of hydrologic and geologic settings.

4) Insight into the variation of groundwater vulnerability in Texas, and the factors that influence that vulnerability.

1.6 OUTLINE OF DISSERTATION

This dissertation consists of seven chapters.

Chapter One, this chapter, sets out the motivation, goals, scope and plan of the research project.

Chapter Two, Literature Review, summarizes the existing state of knowledge about the problems of groundwater vulnerability analysis, with particular emphasis on statistical and empirical approaches.

Chapter Three, Data Sources and Description, describes the data that are analyzed in the research, where they came from and how they were manipulated to support the needs of the research.

Chapter Four, Methods, describes the mathematical models and methods that were employed in the research. The emphasis in this chapter falls on the theoretical and mathematical aspects of the research.

Chapter Five, Procedures, focuses on the details of carrying out the analyses. This "how to" chapter describes the computer programs that actually carried out the mathematics described in Chapter Four.

Chapter Six, Results, presents maps, tables, and summary statistics that describe the distribution of nitrate in Texas, its relation to indicator variables, and the relationship between nitrate distribution and the occurrence of pesticides in groundwater in the midwestern United States.

Chapter Seven, Conclusions, finishes the dissertation by offering a summary of the completed project and the meaning of the results presented in Chapter Six.

Chapter 2: Background and Literature Review

The purpose of this chapter is to set the present study in the context of other studies of groundwater vulnerability. Since this study employs a statistical approach to vulnerability assessment, the literature review emphasizes those studies that have applied statistical methods to this problem. In addition, the use of nitrate as an indicator of vulnerability to contamination by agricultural chemicals is discussed.

This chapter addresses the following questions:

- What uses are there for groundwater vulnerability analysis?
- What methods are used for groundwater vulnerability analysis?
- Why use a statistical approach?
- How have statistical methods been applied to groundwater vulnerability analysis?
- What does the occurrence of nitrate indicate about agricultural contaminants?
- How does the method used in the present study differ from previous statistical approaches?

2.1 Uses for Groundwater Vulnerability Assessment

A groundwater vulnerability analysis identifies regions where groundwater is likely to become contaminated as a result of human activities. The objective of vulnerability analyses is to direct regulatory, monitoring, educational, and policy development efforts to those areas where they are most needed for the protection of groundwater quality. Fundamentally, this is an

economic goal, rather than a scientific one. Vulnerability analysis should provide an answer to the question "Where should groundwater protection efforts be directed to return the most environmental and public health benefits for the least cost?"

In its 1991 final report, EPA's Ground-Water Task Force states as part of its "Ground Water Protection Principals" that "Efforts to protect ground water must also consider the use, value, and *vulnerability* of the resource, as well as social and economic values." (USEPA, 1991, emphasis added). The report goes on to list consideration of groundwater resource vulnerability as part of a "mature" method for setting priorities for groundwater protection. As an example of State efforts EPA regional offices should use as indicators while evaluating progress in the implementation of State Ground Water Protection Plans, the report cites development of

> a comprehensive State vulnerability assessment effort that can assist in developing State Pesticide Management Plans; targeting mitigation measures under State Nonpoint Source Management Plans; and prioritizing ground-water areas for geographically-targeted education; permitting; enforcement and clean up efforts across all ground-water related programs.

Two specific examples of EPA's intended use of groundwater vulnerability analysis are the existing regulations defining National Primary Drinking Water Standards, and the proposed differential protection strategy for imposing more restrictions on pesticide use where groundwater is vulnerable.

The first example was discussed in Chapter 1. The second example, EPA's proposed differential protection strategy for pesticides, is summarized as follows

> Under the new strategy of differential protection, if EPA determines that a pesticide poses a significant human health or environmental risk (because it may leach to groundwater) and the risk cannot be dealt with by labeling or national restricted use provisions, a state management plan (SMP) will be required for the sale and use of the pesticide in a state. The plan must describe how the risks will be addressed. As part of these plans, states will target specific areas, distinguishing those locales that warrant enhanced protection from those that merit less attention because of the lower value of the groundwater and/or their lower vulnerability to groundwater contamination. (GAO, 1992)

The National Research Council (NRC, 1993) has identified four general categories for the use of groundwater vulnerability analysis. These are: policy analysis and development, program management, to inform land use decisions, and to improve general education and awareness of a regions hydrologic resources. Judging by EPA's regulatory actions and stated groundwater protection strategy, by the publication of the NRC report, and by the results of a General Accounting Office survey (GAO 1992) stating that 42 of 45 responding states had conducted some form of groundwater vulnerability analysis, it is reasonable to conclude that groundwater vulnerability analyses are going to play some role in public policy on groundwater quality, and that methods for improving them should be studied.

2.2 GROUNDWATER VULNERABILITY ASSESSMENT METHODS

Comprehensive reviews of groundwater vulnerability assessment methods are presented in reports by the General Accounting Office (GAO, 1992) and the National Research Council (NRC, 1993). Both reports divide groundwater vulnerability assessment methods into three categories: (1) overlay and index

methods, (2) methods employing process-based simulation models, and (3) statistical models. The same categories will be applied here.

Overlay and Index Methods. Overlay and index methods (the GAO report calls these "parameter weighting" methods), combine maps of parameters considered to be influential in contaminant transport. Each parameter has a range of possible values, indicating the degree to which that parameter protects or leaves vulnerable the groundwater in a region. Depth to the groundwater, for example, appears in many such systems, with shallow water considered more vulnerable than deep.

The simplest overlay systems identify areas where parameters indicating vulnerability coincide, e.g. shallow groundwater and sandy soils. More sophisticated systems assign numerical scores based on several parameters. The most popular of theses methods, DRASTIC (Aller, et al. 1987) uses a scoring system based on seven hydrogeologic characteristics of a region.

The acronym DRASTIC stands for the parameters included in the method: Depth to groundwater, Recharge rate, Aquifer media, Soil media, Impact of vadose zone media, and hydraulic Conductivity of the aquifer. DRASTIC is applied by identifying mappable units, called hydrogeologic settings, in which all seven parameters have nearly constant values. Each parameter in a hydrogeologic setting is assigned a numerical rating from 0–10 (0 meaning low risk; 10 meaning high risk) which is multiplied by a weighting factor varying from 1–5. Two sets of weights, one for general vulnerability, another for vulnerability to pesticides can be used. A score for the setting is

calculated as the sum of the seven products. DRASTIC scores are roughly analogous to the likelihood that contaminants released in a region will reach ground water, higher scores implying higher likelihood of contamination. DRASTIC is used to produce maps of large regions showing their relative vulnerability. Its authors recommend that it be applied on no region smaller than 100 acres.

Several other overlay and index systems for groundwater vulnerability assessment exist; the NRC report lists seven, including DRASTIC. Typically, such systems include variables related to ground water recharge rate, depth to the water table, and soil and aquifer properties. The relative importance of the variables and the methods for combining them vary from one method to another, but all share some common traits. In general, overlay and index methods rely on simple mathematical representations of expert opinion, and not on process representation or empirical data.

Mathematical Models. Process-based mathematical models such as PRZM, GLEAMS, and LEACHM can predict the fate and transport of contaminants from known sources with remarkable accuracy in a localized area by applying fundamental physical principals to predict the flow of water in porous media and the behavior of chemical constituents carried by that water. In the hands of knowledgeable analysts with the appropriate site-specific information, such models allow threats to the safety of ground water supplies to be recognized and can play an important role in planning remediation efforts. Unlike other

groundwater quality prediction methods, mathematical models predict variations of water quality both in space and in time.

Although process models offer the most sophisticated, and potentially most accurate predictions of water quality, they are not widely used for regional groundwater vulnerability analysis. Reporting on the vulnerability assessment methods used by state agencies, the GAO found that none used mathematical process models (GAO, 1992).

The Federal Republic of Germany, however, has sponsored a modeling project to identify the regions most susceptible nitrate contamination of groundwater (Wendland et al. 1993). The data and model are based on a grid of the nation consisting of nearly 40,000 3 x 3 km cells. The data include five hydrologic themes, seven soil themes, three hydrogeologic themes, six themes describing regional groundwater flow, and five themes contributing to the nitrogen cycle. From this data, the model produces a map of "Denitrification Conditions" and three maps of potential nitrate concentrations under different flow assumptions. The quantity of data required for this study, both in terms of characteristics mapped and detail of mapping, requires greater resources than any State in the U.S. has presently devoted to groundwater vulnerability analysis.

Statistical Methods. Empirical or statistical methods are the least common vulnerability assessment methods in the literature. Although statistical studies are used as tests for other methods, and geostatistical methods such as kriging are frequently used to describe the distribution of water quality parameters, very few vulnerability assessment methods are directly based on statistical methods. The

GAO report identifies one statistically based method, and the NRC report adds one more. These will be discussed in the following section. In addition, the GAO reports that although twelve states use empirical methods for assessing the vulnerability of groundwater to pesticide contamination, their methods are not published, and have not been verified.

Checklists. A fourth category, not included in the GAO or NRC reports, encompasses the methods used by Texas and several other states for their Primary Drinking Water Standards enforcement. These methods provide a checklist or decision tree, based on well construction, geologic and soil factors, and the presence of chemical sources in the vicinity of the well. The vulnerability assessment method used by the Texas Natural Resource Conservation Commission (Blodgett 1993) is a representative example.

The assessment consists of the following steps:

1. Determining the location of the water supply well.
2. Acquisition of well construction and material setting descriptions, and driller's logs for the well.
3. Verification of proper well construction, and identification of a *vulnerability point*—typically the bottom of a cemented well casing, the top of a gravel pack, or the top of the well's shallowest open interval. A well lacking cemented casing, or otherwise improperly constructed is considered susceptible to contamination.
4. Examination of driller's logs to determine geologic susceptibility. The thickness of aquitards (materials with low vertical conductivity) above

the vulnerability point is tabulated. If the vulnerability point lies below a single aquitard layer thirty feet thick (forty feet if the aquitard is exposed at the surface) or below multiple aquitard layers with a total thickness of 100 feet or more, the well is considered protected (not susceptible). A different method is used for wells in fractured rock or carbonate aquifers.

5. Delineation of a zone of contribution for susceptible wells. The limits of the zone for a forty-year time-of-travel are calculated with a semi-analytical computer model, WHPA, also used for wellhead protection programs in Texas and other states.
6. Review of contaminant use in the zone of contribution. A variety of databases with spatial coordinates are used for this purpose.
7. Waiver determination. Using the results of the preceding steps, a list of contaminants to be tested for is generated. Three- to nine-year waivers are given for contaminants not requiring monitoring.

The above procedure, and a similar vulnerability assessment method for Wisconsin (Wisconsin Bureau of Water Supply 1992), rely on a process similar to the overlay and index methods described earlier. Like those methods, the checklist applies expert knowledge and opinion systematically to the problem of vulnerability assessment, but does not employ a specific process model (except in an ancillary role) or an empirical/statistical basis for its recommendations.

2.3 STATISTICAL GROUNDWATER VULNERABILITY ASSESSMENT

Between them, the GAO and NRC reports on vulnerability assessment methods found only two published methods for statistical groundwater

vulnerability analysis. Although a number of studies have applied statistical methods to verifying other methods, or have sought to prove or disprove a correlation between single environmental parameters (land use/land cover, for example) and groundwater quality, attempts to produce a predictive method for groundwater quality from empirical data are uncommon. A literature search revealed only six studies (including the two listed in the GAO and NRC reports) that attempt to identify and rate the importance of multiple indicators of groundwater vulnerability or groundwater quality. None of these studies used geostatistical methods.

Teso et al. (1988) used discriminant analysis—a statistical method for assigning objects to categories based on their location in a multi-dimensional data space—to identify sections (one mile squares) in Fresno County, California as susceptible (or not) to contamination by 1,2-dibromochloropropane (DBCP). They compiled both groundwater DBCP measurements and soil taxonomic groups for 835 sections. Based on the DBCP measurements they sorted the section into categories of "contaminated," meaning that DBCP had been detected in a well located in that section or "not contaminated," meaning that no wells in the section had detectable levels of DBCP. 511 of the 835 sections were classified as contaminated. In addition, the presence or absence of soils belonging to 228 taxonomic groups was encoded in a 228-dimensional binary vector for each section. A 1 in the nth dimension of a section's soil vector indicates the presence of soil type n; a 0 in the same place indicates its absence. The 835 sections were used to calibrate a discriminant function that identifies

any point in the 228-dimensional soil data space as "contaminated" or "not contaminated." A similar analysis with a smaller number of higher-order soil classifications (the 228 taxonomic groups were reduced to only six soil series) yielded a discriminant function based on the presence or absence of only six soil series in a section. This reduced discriminant function yielded a 0.776 success rate for classification of sections in Fresno County. When tested on an independent data set from nearby Merced County, the same function yielded a success rate of 0.573.

Chen and Druliner (1986) applied multiple linear regression to measurements of nitrate and herbicide concentrations in 82 wells tapping the High Plains Aquifer in Nebraska. They used the regression method to identify those environmental factors most strongly related to the concentration of nitrate and triazine herbicides (a class of herbicides that includes atrazine, cyanazine, and others). They found that three variables (well depth, irrigation-well density, and nitrogen-fertilizer use) explain 51% of the variation in nitrogen concentrations, and that two variables (specific discharge and well depth) explain 60% of the variation in triazine herbicide concentrations. Using nitrate concentration in combination with specific discharge explains 84% of the variation in triazine herbicide concentrations.

Statistical Studies of Groundwater Quality Indicators. In addition to the studies identified by the GAO and NRC reports, other research has used statistical methods to identify relationships between small numbers of indicators

and measured water quality parameters, although not directed toward producing a vulnerability assessment method.

Burkart and Kolpin (1993a) examined the influence of a variety of hydrogeologic and land-use factors on the concentrations of nitrate and atrazine in shallow aquifers over an area encompassing portions of twelve States in the midwestern U.S. They sought to identify correlations between individual factors, such as aquifer type or depth to groundwater, and the concentrations of the constituents. Using non-parametric methods, including the Mann-Whitney rank sum test and contingency tables, they found significant differences in nitrate and herbicide concentrations when wells are grouped by aquifer class (bedrock or unconsolidated) and by depth of unconsolidated material over the aquifer.

Nightingale and Bianchi (1980) used linear correlation coefficients and multiple linear regression to examine the relationship between soil and aquifer permeabilities and measurements of conductivity, anion, and cation concentrations. Like the work of Teso et al., this study was based on historical measurements grouped by the sections from which they were taken. They found that salinity was correlated to soil and aquifer permeability, but that nitrate levels correlated only with the estimated specific yield of the aquifer system.

Helgesen et al. (1992), seeking a connection between land use and water quality, delineated discrete regions of uniform land use over a portion of the High Plains aquifer in southern Kansas. They selected one well at random from each region and tested a water sample for a variety of agricultural and petroleum related chemicals. Non-parametric hypothesis tests showed significantly higher

mineral concentrations under irrigated croplands and petroleum-producing areas than under undeveloped range land.

Baker et al. (1994) used an approach similar to that of Burkart and Kolpin (1993), but applied it to a larger body of samples, collected through a voluntary well testing program. Samples of water from rural wells submitted by more that 43,000 participants in twelve states were analyzed for nitrate and herbicide concentrations. Non-parametric statistical methods were applied to compare the analysis results with descriptions of the wells and their surroundings submitted by the participants with the water samples. They found that the age of the well, its depth, and its proximity to feedlots or barnyards significantly influence the likelihood of finding elevated nitrate concentrations in the samples. Likelihhoods increased dramatically when two "risk factors" were combined. They also found that factors influencing nitrate exerted similar influences on herbicide concentrations.

2.4 CHOICE OF METHOD

A statistical approach was selected for this study for two reasons. The first is dissatisfaction with index/overlay methods and process-based models. The second is the appropriateness of this approach to GIS-based analysis.

Although they represent informed opinion, and apply consistent standards to all regions, overlay and index methods lack a sound methodological foundation, being based neither on direct observation nor first principles. "These methods are driven largely by data availability and expert judgment, with less emphasis on processes controlling ground water contamination. One can argue

whether the factors included in the methods are the relevant ones for vulnerability assessment and whether the factor ratings are appropriate" (NRC, 1993). These doubts are supported by studies carried out to test DRASTIC. The GAO report observes that "…tests of DRASTIC generally indicated a poor relationship between model predictions (that is, relative groundwater vulnerability), and monitoring results (that is, where pesticides are found)" (GAO 1992).

Overlay and index methods are also difficult to interpret quantitatively and provide no estimates of uncertainty. Is a region with a DRASTIC score of 200 twice as vulnerable to contamination as one with a score of 100? Does a DRASTIC score of 150 mean "between 140 and 160" or "between 100 and 200?" DRASTIC's authors do not provide answers to these questions and caution against any absolute interpretation of the index. This places serious limitations on the value of DRASTIC as a guide to forming policy. Since DRASTIC is the most thoroughly studied of the index/overlay systems, others should be viewed with less confidence.

Mathematical models of groundwater processes have the great advantage of being based on sound principles, rather than opinion, but this does little to enhance their value for policy guidance at a state or regional level. The models require more expertise and (as illustrated by the German example) more detailed data than state agencies can provide on a regional scale. The NRC report offers the following view of process models.

> It must be recognized that sophisticated models may not necessarily provide more reliable outputs, especially for regional-

> scale, and even for field-scale applications. Since data for many of the required input parameters for sophisticated models are not always available, their values have to be estimated by indirect means using surrogate parameters or extrapolated from data collected at other locations. Errors and uncertainties associated with such estimates or extrapolations can be large and may negate the advantages gained from a more rigorous process description in the simulation model. (NRC 1993)

Given the state of available data, such models are not well suited to the task of regional assessment of groundwater vulnerability.

Statistical approaches offer the possibility of a method that is as easy to apply as an index/overlay method, but with a more defensible foundation. The weighted-sum approach of DRASTIC looks like the product of a multiple linear regression, and the NRC report observes that "Vulnerability assessment methods that use overlay/indexing techniques are an eyeballed form of multivariate discriminant analyses that lack probability estimates" (NRC 1993). Since overlay methods look like the results of statistical analysis, why not develop one that *is* what it looks like? Although it is risky to apply empirical methods outside the range of conditions over which it was calibrated, such methods are at least based on real measurements, not just a set of opinions.

Data Requirements. Statistical methods require data, the more data and the higher the quality, the better. Collection of groundwater quality data is expensive and time-consuming, driving up the cost of statistical investigations. Burkart and Kolpin orchestrated the collection of samples from 303 wells throughout the midwest during the spring and summer of 1991. This was a substantial undertaking with very careful quality control, and it produced roughly 600 measurements of herbicide, nitrate, and ammonium concentration. Given

the size of the region under study, this is a small number of measurements on which to base broad conclusions of cause and effect. Anyone attempting a regional-scale study of water quality faces a very substantial problem in gathering sufficient data.

At the time this study was begun, the existing body of pesticide data in Texas was not sufficient to form the basis of a statistical study. EPA's *Pesticides in Groundwater Database* (EPA, 1992), which compiles monitoring study results over the period 1971–1991, contains only 511 pesticide measurements in Texas. The Texas Department of Agriculture (Aurelius, 1989) carried out a pilot study in 1897 and 1988 to estimate the extent to which rural domestic wells are contaminated by pesticides from nonpoint agricultural sources. 175 wells were tested for nine pesticides, arsenic, and nitrate. The study was confines to high-risk areas and cannot be considered as representative of the State as a whole.

Since pesticide measurements in groundwater were not adequate to support the development of a statistical method for groundwater vulnerability analysis, another constituent—nitrate, which has been extensively measured in groundwater—was chosen.

2.5 NITRATE IN GROUNDWATER

This section presents a brief review of nitrate in groundwater, relevant to the present study, rather than a comprehensive review of the extensive literature on nitrate in groundwater. In particular, the nitrate cycle is discussed, and important concentration values are identified.

High concentrations of nitrate (NO_3^-)in drinking water may cause the disease methemoglobinemea in small children (Hem 1989). Because of this and other diseases linked to nitrate (and possibly because it is inexpensive to measure), its concentration in public water supplies is monitored and regulated by federal law. The National Primary Drinking Water Standards (40 CFR 141) set the maximum contaminant level (MCL) for nitrate at 10 mg/l (measured as nitrogen). Groundwater systems must monitor for compliance with the MCL annually. If nitrate in excess of 5 mg/l is detected, the system must increase its monitoring to quarterly for at least one year.

Nitrate occurs naturally from mineral sources and animal wastes, and anthropogenically as a byproduct of agriculture and from human wastes. Nitrate is the most highly oxidized form of nitrogen in the nitrogen cycle, which includes activities in the atmosphere, hydrosphere, and biosphere. Figure 2.1 shows the following major transformations from the nitrogen cycle (Madison and Brunett, 1985)

Assimilation of inorganic forms of nitrogen (ammonia and nitrate) by plants and microorganisms.

Heterotrophic conversion of organic nitrogen from one organism to another.

Ammonification of organic nitrogen to produce ammonia during the decomposition of organic matter.

Nitrification of ammonia to nitrate and nitrite by the chemical process of oxidation.

Denitrification (bacterial reduction) of nitrate to nitrous oxide (N_2O) and molecular nitrogen (N_2) under anoxic conditions.

Fixation of nitrogen (reduction of nitrogen gas to ammonia and organic nitrogen) by microorganisms.

Madison and Brunett (1985) list the following as major anthropogenic sources of nitrate: "fertilizers, septic tank drainage, feedlots, dairy and poultry farming, land disposal of municipal and industrial wastes, dry cultivation of mineralized soils, and the leaching of soil as the result of the application of irrigation water." Natural sources include: "soil nitrogen, nitrogen-rich geologic deposits, and atmospheric deposition."

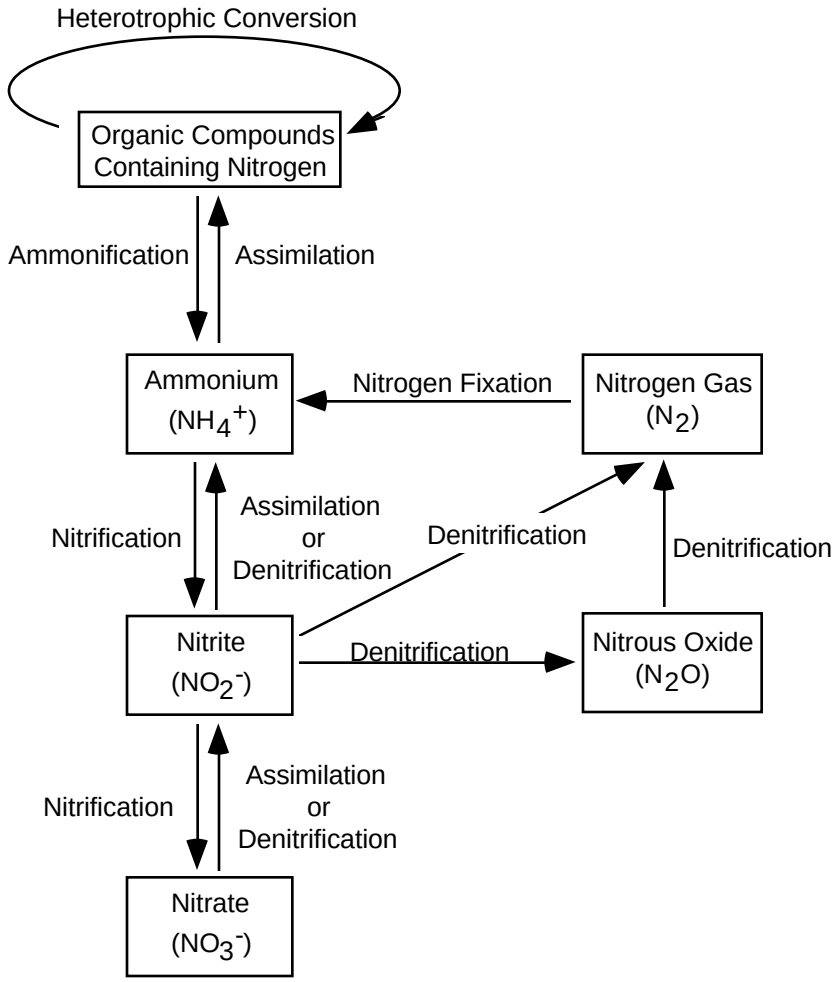


Figure 2.1 Simplified Biological Nitrogen Cycle
[after Madison and Brunett (1985)]

According to Hem (1989), nitrogen occurs in water as nitrate or nitrite anions, as ammonium cations, and in a variety of organic compounds. Nitrite and the organic species are unstable in aerated water. Ammonium cations are strongly adsorbed on mineral surfaces, but the anionic species are readily transported in water and are stable over a wide range of conditions.

Given the wide range of nitrate sources associated with agriculture, its chemical stability in water, and its high mobility—to say nothing of the frequency with which it has been measured in water—nitrate is a natural choice as an indicator for vulnerability of groundwater to contamination to nonpoint agricultural sources. This use has been suggested by Cohen et al. (1984), and has been tested by a number of investigators. Domagalski and Dubrovsky (1992) found no significant difference in nitrate concentrations between wells with and without triazine herbicide residues in the San Joaquin valley of California. An examination of the report by Burkart and Kolpin (1993a) shows that the geological factors associated with high frequencies of herbicide contamination are also associated with high frequencies of excess nitrate detection. Baker et al. (1994) found a similar correspondence between nitrate and pesticide vulnerability in samples collected from rural wells in 17 States.

Nitrate concentrations are usually reported in units of miligrams per liter (mg/l) with the mass representing either the total mass of nitrate ion in the water (nitrate-NO_3), or as the mass of only the nitrogen (nitrate-N). The molecular weight of nitrate is 62; the molecular weight of nitrogen is 14, so the ratio of a concentration measured as nitrate-NO_3 to an equivalent concentration measured as nitrate-N is 4.43. The MCL of 10 mg/l nitrate-N is equivalent to 44.3 mg/l nitrate-NO_3.

In their nationwide study of nitrate in the groundwater of the U.S., Madison and Brunett assigned the following interpretations to ranges of nitrate concentrations (in nitrate-N)

- Less than 0.2 mg/l—Assumed to represent natural background concentrations.

- 0.21 to 3.0 mg/l—Transitional; concentrations that may or may not represent human influence.

- 3.1 to 10 mg/l—May indicate elevated concentrations resulting from human activities

- More than 10 mg/l—Exceeds maximum concentration for National Interim Primary Drinking-Water Regulations.

Their selection of 3.0 mg/l as a threshold to indicate human influence has been followed by many investigators, including Burkart and Kolpin, and Baker et al. The use of individual concentration levels in this study is discussed further in Section 4.1.

2.6 OUTLINE OF PROPOSED VULNERABILITY ANALYSIS METHOD

The general form of the approach to statistical groundwater vulnerability analysis advanced in this work can be summarized in six steps. These are:

1. Select a constituent or set of constituents, whose presence will indicate the degree of vulnerability of a groundwater source.
2. Identify a set of distinct mappable regions of the surface or subsurface.
3. Assemble a body of measurements of the constituent identified in step 1 that can be linked with the regions identified in step 2.
4. Calculate descriptive statistics for the body of measurements linked with each region.
5. Map the variation of the descriptive statistics from region to region.
6. Relate the variation of the descriptive statistics to the variation of indicator parameters by forming a mathematical expression that mimics

the relationship between the descriptive statistics and indicator values mapped over the same set of regions.

The results of these steps include maps and numerical values associated with the regions, indicating their vulnerability to contamination as represented by the descriptive statistics, and a mathematical model that permits those results to be extended to areas where water quality data have not been collected, but values of the indicator parameters are known.

2.6.1 Comparison of Method with Previous Studies

The six steps are proposed as a synthesis of the approaches taken in the statistical studies cited in Section 2.2. The work of Teso et al. (1988), and Nightingale and Bianchi (1980) follows steps 1 through 4 by dividing the study area into regions by square-mile section, forming groups of water quality measurements from historic data based on the location of sampling sites in the sections, and forming summary statistics for each section—binary classifications in based on the presence or absence of DBCP in any well in the section in Teso et al, arithmetic averages of nitrate concentrations and electrical conductivity for all measurements from the section in Nightingale and Bianchi. Similarly, Burkart and Kolpin (1993b) grouped the measurements collected in their reconnaissance of agricultural contaminants in the mid-continental U.S. by their location in major land resource areas (MLRAs) and calculated a third type of summary statistic—the frequency with which threshold concentrations of nitrate and herbicides were exceeded in measurements collected in the MLRAs.

Burkart and Kolpin, Baker et al. and Teso et al. mapped their results (step 4), but not Nightingale and Bianchi did not.

Comparison of summary statistics to indicator parameters and formation of a mathematical model (step 5) is carried out in all of the cited studies except for Burkart and Kolpin (1993b). Chen and Druliner (1986), and Helgesen et al. (1992) compared indicator parameters directly to concentrations reported in individual water samples rather than statistics calculated on groups of measurements, although Helgesen et al. intend each well to represent a region. Burkart and Kolpin (1993a) re-group their measurements for each indicator, rather than forming one set of groups and comparing their statistics to indicator variations over the same groups. Teso et. al and Nightingale and Bianchi base their results on region-based statistics and indicator values from the same regions.

The cited studies approach data compilation in one of two ways. These can be identified as the *well-oriented* approach and the *region-oriented* approach.

The *well-oriented* approach, taken by Burkart and Kolpin, by Chen and Druliner, and by Baker et al. is to select a relatively small number of wells to represent a each region or setting. Measured variations in constituent concentration from well to well are compared to variations in the characteristics of the wells and their surroundings. Barringer et al. (1990) point out that results from such studies can be biased due to spatial autocorrelation if the wells are too close together. A well-oriented study requires careful planning or data screening to assure that the selected wells are typical of the regions where they are located.

The *region-oriented* approach is to define a set of regions, calculate two sets of statistics on the regions—one of water quality and one of potential indicators—and study the relationships between the two sets of statistics. This is the method that Teso et al. and Nightingale and Bianchi applied in their studies California. In both studies the regions were surveying sections. In Teso et al., the water quality statistics were the binary classification of the sections by having or not having DBCP detections, the indicator statistics were the soil taxonomy vectors, and the relationship between the two was analyzed with discriminant analysis. In Nightingale and Bianchi, the water quality statistics were arithmetic averages of conductivity or cation and anion concentrations, the indicator statistics were averages of aquifer and soil permeability, and the relationships were examined with linear correlation coefficients for paired variables and multiple linear regression for multiple variables. Helgeson et al. identified regions by land use, and characterized each by a single randomly selected water sample. In another report on the results of their groundwater reconnaissance of the midwest, Burkart and Kolpin (1993b) used a GIS to identify regions—STATSGO soil polygons (see Chapter 3) or Major Land Resource Areas—as more or less vulnerable to contamination, based on the frequency that atrazine was detected in wells in those regions.

Region-oriented studies avoid some of the problems of well-oriented studies, but are subject to some limitations. Bias due to autocorrelation is reduced by aggregating samples, giving each region equal weight in evaluating the relationship between indicators and water quality. The potential for an

atypical well to incorrectly characterize a region is reduced (if sufficient data is available) by the contributions of several wells to the description of water quality in the region. The regional orientation, however, precludes any study of the effects of well-specific characteristics such as pumping rates or construction characteristics. On balance, the regional approach was judged more suitable for the data available, and the objectives of the study.

2.6.2 Application in Present Work

In this study, the five steps were implemented as follows.

1. Use *nitrate* to represent the vulnerability of groundwater.
2. Divide Texas into a grid of *7.5' quadrangles*, based on the well-numbering system used by the Texas Water Development Board (TWDB) in its Ground-Water Data System (Nordstrom and Quincy, 1992). The well-numbering system and the quadrangles are described in Section 4.2.
3. Form groups of groundwater nitrate measurements recorded in the TWDB Ground-Water Data System based on the location by quadrangle of the wells from which the water samples were collected.
4. Calculate statistical estimates of *exceedence probabilities*, the likelihood that nitrate concentrations measured in water samples collected in the quadrangles will exceed selected threshold values.
5. Prepare maps of the quadrangles showing the variation of the exceedence probabilities for the selected thresholds.
6. Prepare maps of four indicator parameters—average annual precipitation, average soil thickness, average soil organic matter content, and average

annual nitrogen fertilizer sales—and use *stepwise multiple linear regression* to construct a simple linear model of exceedence probabilities based on these indicators.

The italicized words in the list above indicate specific choices made in the course of this investigation that make it distinct from the general model described at the beginning of this section. All of these choices will be discussed in later sections of this chapter.

In addition to the 7.5' quadrangles, five aquifers—the Carrizo-Wilcox, the Edwards (Balcones Fault Zone), the Hueco-Mesilla Bolson, the Ogallala, and the Seymour—were used as an alternate set of regions to divide a subset of the TWDB data into groups for an analysis similar to that performed on the quadrangles. The variation of exceedence probabilities for this subset was compared from aquifer to aquifer as well as by the four parameters listed in step 5 above.

The choice of nitrate for study, the methods used to form the data into groups for analysis, the methods used to calculate the exceedence probabilities, and the use of stepwise multiple linear regression are described in Chapter 4. The data used in the analyses are described in Chapter 3.

Chapter 3: Data Sources and Description

The conclusions that this study presents are based on statistics calculated from 46,507 nitrate measurements taken from 29,485 wells throughout Texas. Following the methods outlined in Section 2.6, and described in detail in Chapter4, the spatial variation of the statistics is mapped to identify regions of high or low vulnerability to nitrate contamination. The spatial variation in the statistics is then compared to the spatial variation of potential water quality indicators, including soil parameters, average annual precipitation, and fertilizer sales, in order to assess the value of these data as indicators of water quality.

Because the structure and limitations of these data strongly influence the choice of the methods used, this chapter, which describes the data itself, is a necessary prelude to Chapters 4 and 5, which describe the methodology and procedures followed in the study. This chapter contains seven sections, one for each data set used in the study. These data sets can be divided into three groups:

1) Primary data, consisting of groundwater nitrate concentration measurements and descriptions of the wells where the groundwater was collected for testing. The nitrate data are described in Section 3.1 and the well data are described in Section 3.2.
2) Data to be considered as potential indicators of water quality. These include soil thickness and organic content described in Section 3.3; annual average precipitation, described in Section 3.4; and average annual nitrogen fertilizer sales, described in Section 3.5.

3) Independent measurements of nitrate and herbicides, used to test assumptions made in the study. These include measurements of nitrate in public water sources collected by the Water Utilities Division of the Texas Natural Resource Conservation Commission, described in Section 3.6, and the first year's results of the U.S. Geological Survey's reconnaissance of nitrate and herbicides in groundwater in the Midwest, described in Section 3.7.

3.1 NITRATE MEASUREMENT DATA

The nitrate measurements used in this study come from the Texas Water Development Board's (TWDB) Groundwater Data System (Nordstrom and Quincy, 1992). This statewide database contains physical descriptions of wells and their surroundings in Texas, and levels of chemical constituents measured by a variety of public agencies. The TWDB maintains the database to characterize the quantity and quality of groundwater available throughout the state, in support of the preparation of the Texas Water Plan (TWDB, 1994).

For every nitrate measurement listed in the Groundwater Data System as of October 1993—a total of 62,692 database records—the data fields listed in Table 3.1 were retrieved for use in this study. Of these data fields, the well ID, date, and nitrate level have values in all records. Many records have no values for the collecting agency or reliability remarks. The values in the flag field are discussed in section 3.1.1.

Table 3.1 Nitrate Measurement Data

Name of Data Field	Description
Well ID	Identification number of well where collected (see section 3.2.1)
Date	Date collected
Agency	Collecting agency (e.g. USGS, TWDB, etc.)
Reliability Remarks	Numeric code indicating handling and analysis reliability
Nitrate Level	Concentration (mg/l NO_3) of nitrate.
Nitrate Flag	Code ("<" or ">") indicating level is reporting limit rather than measured concentration

Table 3.2 Nitrate Measurements from Well 5740304

Year	Month	Reported Nitrate (mg/l NO_3)	Adjusted Nitrate (mg/l N)
1966	4	< 0.4	[2] 0.10
1966	12	< 0.4	[2] 0.10
1967	6	14.0	3.17
1968	6	12.0	2.71
1968	7	13.5	3.05
1971	6	8.0	1.81
1972	5	8.0	1.81
1974	3	5.9	1.33
1976	8	4.7	1.06
1980	3	3.9	0.88
1986	6	2.13	0.48
1991	8	0.44	[2] 0.10

Nitrate concentrations in the TWDB database are listed as mg/l nitrate (nitrate-NO_3). However, unless otherwise noted the values used in this study's statistical analyses and reported here are in equivalent values of nitrate as nitrogen (nitrate-N), the units used in EPA regulations. 1 mg/l nitrate-N equals 4.42 mg/l nitrate-NO_3. Each nitrate-NO_3 value in the data set was converted to an equivalent nitrate-N value. To maintain a uniform reporting limit for all records used in the study, all values at or below a value of 0.1 mg/l nitrate-N will be treated as ² 0.1 mg/l. A nitrate concentration greater than 0.1 mg/l will be considered a "detection" and concentrations less than or equal to this value will be considered to be "below detection limit." As an illustration of this conversion and adjustment, Table 3.2 shows the nitrate measurements listed in the TWDB database for well 5740304 and the adjusted values used for analysis in this study. Of the twelve measurements shown, nine are considered detections of nitrate and three fall below the detection limit.

3.1.1 Nitrate Reporting Limits

The flag field in a nitrate measurement record may be blank or may contain a "<" or ">" character. A blank should indicate that the value listed for nitrate concentration in the nitrate level field is the actual value measured in the water; a "<" or ">" indicates that the value is a detection or reporting limit, rather than an actual value. The ">" character appeared 5 times in the retrieved data. The "<" character appears in 4047 (6.5%) of the records. A value of 0.40 mg/l nitrate-NO_3 (approximately equal to 0.1 mg/l nitrate-N) appears most frequently as a reporting limit, as the histogram in Figure 3.1 illustrates. (Not

shown in the figure are 403 records with detection limits greater than 1 mg/l NO_3.)

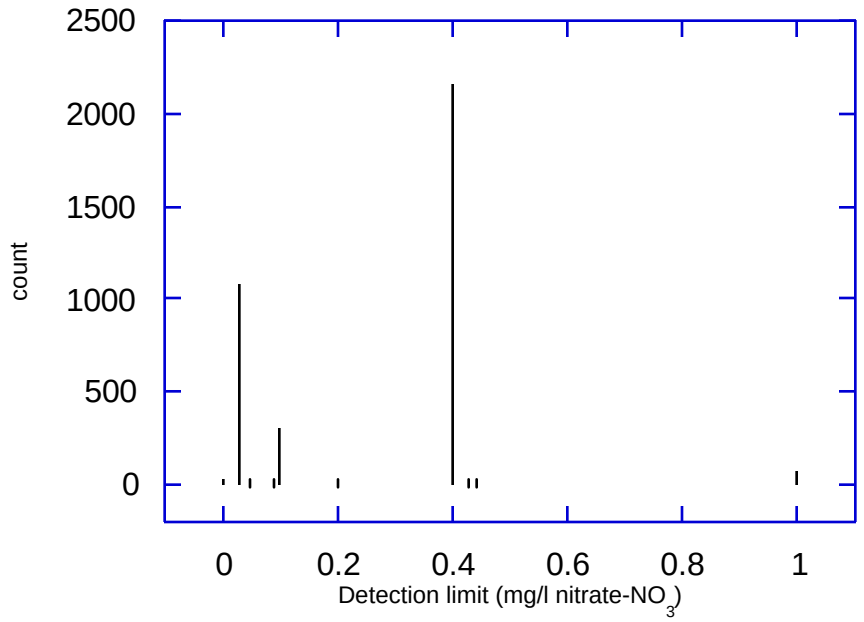


Figure 3.1 Reported Detection Limits for Nitrate

Although a blank in the flag field should indicate that the nitrate level in the record is a true measured concentration, the number of occurrences of some values suggests otherwise. Figure 3.2 shows a histogram of nitrate levels below 1 mg/l nitrate-NO_3 in records with blank flag fields. The value 0.4 appears 9,793 times in the 58,640 records with blank flag fields. It seems very unlikely that 17% of the water measurements reported in this database should have exactly this value. Since 0.4 is also the most common reporting limit value, a much more plausible explanation of this high incidence would be that the nitrate concentration in many of these cases was below 0.4 mg/l nitrate-NO_3, and that the "<" flag was omitted from the record. Because of the ambiguous meaning of

"0.4 mg/l nitrate-NO_3," this study will treat all occurrences of this value as meaning "less than or equal to 0.4 mg/l nitrate-NO_3."

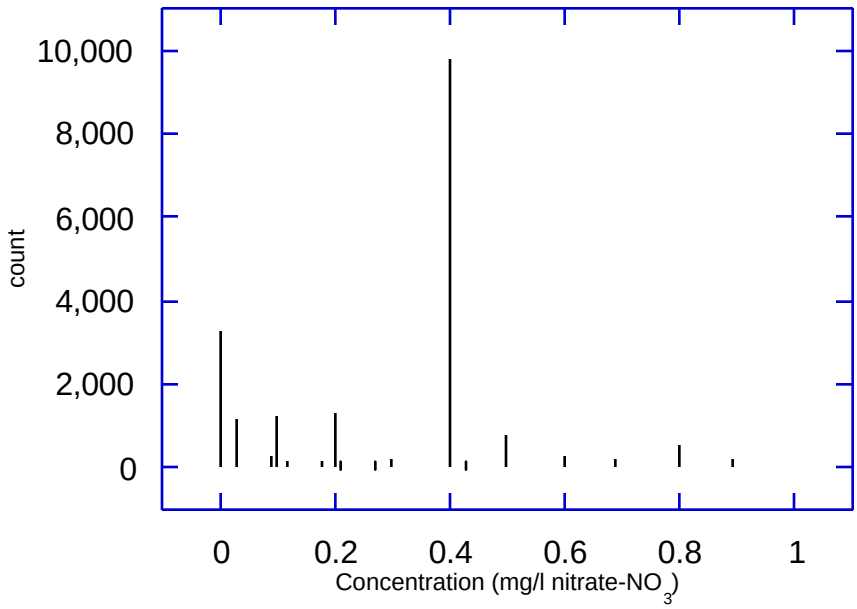


Figure 3.2 Reported Nitrate Concentrations

3.1.2 Sampling Period

The records retrieved from the TWDB database indicated sampling dates from 1896 to 1993. The histogram in Figure 3.3 shows the number of measurements taken in each year. As will be shown in the discussion of the results of this study in Chapter 6, there has been a slight increase over time in the amount of nitrate found in Texas groundwater. In order to reduce the effects of this increase on the data, the study was confined to measurements taken during the years 1962 to 1993. This period was chosen in part because of the sharp increase in the number of nitrate samples collected per year from 1962 onward.

Omitting nitrate measurements prior to this date retained a substantial majority of the database in the study while removing the measurements least likely to be representative of the present condition of Texas groundwater.

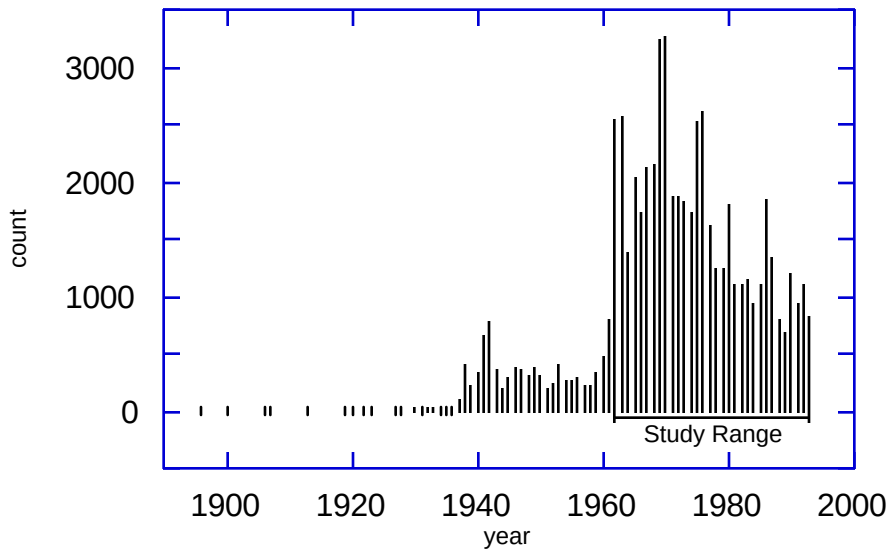


Figure 3.3 Nitrate Measurements Reported by Year

3.1.3 Measurement Record Accuracy

Because the nitrate measurements recorded in the TWDB database come from a variety of sources, they do not conform to a uniform set of quality control standards. In fact, there is evidence in the data to suggest that many values may be questionable. As the preceding section describes, it appears that a "<" flag was omitted from many records in the database. In addition, 140 records indicate nitrate concentrations over 500 mg/l NO_3, a suspiciously high level. (Concentrations of 500 mg/l have been found in waters in the unsaturated zone

below irrigated crops, and levels over 1000 mg/l have been found in pools in the parts of Carlsbad Caverns where bats roost (Hem, 1989). It seems unlikely that concentrations this high are representative of natural groundwater.) 51,329 of the 62,692 nitrate records retrieved from the TWDB database had blank reliability remark fields; while this provides no grounds for excluding the records, it is not a ringing endorsement either.

In spite of these reservations, this study has taken an "innocent until proven guilty" approach to the measurement records. The data were included in the study "as is" unless substantial evidence indicated that they should be excluded. As shown in Table 3.3, records were excluded if reliability remarks indicated questionable collection or handling, if no record could be found of the well from which the water was collected, if the well had bad location data (see following section), if the reported value was "less than" a threshold greater than

Table 3.3 Excluded Measurement Records

Reason	**Criteria**	**# Records Excluded**
Reliability	Remarks = 01, 02, or 03	7,020
Well Data	No well record	11
Well Location	Well mis-located	418
Lower threshold	flag = "<" and nitrate > 0.45 mg/l NO_3	407
Upper threshold	flag = ">"	5
Collection Date	Year < 1962	9,087
Total Excluded		16,185

0.1 mg/l nitrate-N (0.45 mg/l NO_3), if the reported value was "greater than" any threshold, or if the measurement was taken before 1962 (see preceding section). These exclusions left 46,507 nitrate measurement records in the study. This set of nitrate measurement records will be called the "base data set" in the remainder of this document.

3.2 WELL DATA

The data providing physical descriptions of the wells included in the study comes from the same TWDB database as the nitrate measurement data. For each well for which a nitrate measurement was recorded—a total of 38,740 database records—the data fields listed in Table 3.4 were retrieved.

3.2.1 TWDB Well Numbers

TWDB has adopted a system of identification numbers for wells in Texas, based on the location of the wells expressed in latitude and longitude. The following description and Figure 3.4 explain the numbering system.

> [The numbering system] is based on division of the state into a grid of 1-degree quadrangles formed by degrees of latitude and longitude and the repeated division of these quadrangles into smaller ones as shown…
>
> Each 1-degree quadrangle is divided into sixty-four 7-1/2-minute quadrangles, each of which is further divided into nine 2-1/2-minute quadrangles. Each 1-degree quadrangle in the state has been assigned an identification number. The 7-1/2-minute quadrangles are numbered consecutively from left to right, beginning in the upper-left-hand corner of the 1-degree quadrangle, and the 2-1/2-minute quadrangles within each 7-1/2-minute quadrangle are similarly numbered. The first 2 digits of a well number identify the 1-degree quadrangle; the third and fourth digits, the 7-1/2-minute quadrangle; the fifth digit identifies the 2-1/2-minute quadrangle; and the last two digits identify the well

within the 2-1/2-minute quadrangle. (Nordstrom and Quincy, 1992)

Table 3.4 Well Description Data

Name of Data Field	Description
Well ID	Identification number of well (see section 3.2.1)
Aquifer Code	Alphanumeric code for aquifer or geologic unit associated with well
County	Numeric code for county where well is located (FIPS code)
Latitude	Latitude of wellhead location (DMS)
Longitude	Longitude of wellhead location (DMS)
Location Method	Numeric code indicating accuracy of latitude and longitude
Depth	Depth of completed well from land surface (feet)
Depth Method	Alphabetic code indicating source of depth measurement
Altitude	Elevation of land surface at wellhead (feet above mean sea level)
Altitude Method	Alphabetic code indicating source of altitude measurement
Primary Use	Alphabetic code indicating primary purpose served by well

The TWDB well-numbering system will be used throughout this report not only for wells and well locations, but also for numbering 1_, 7.5', and 2.5'

quadrangles used to divide the state for analysis. Well number 5740304 is located in 1_ quad 57, 7.5' quad 5740, and 2.5' quad 57403.

Locating Well 5740304

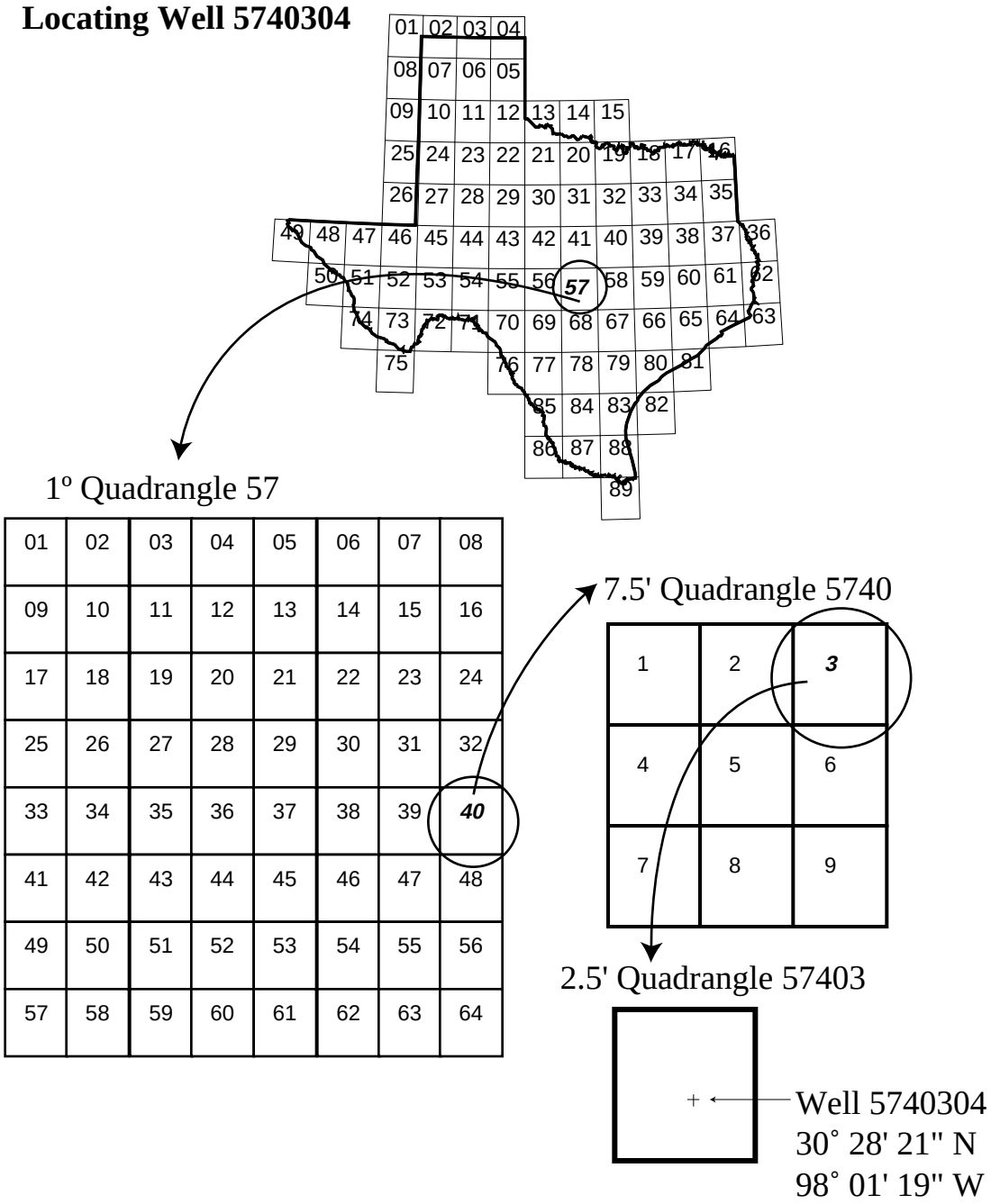


Figure 3.4 TWDB Well-Numbering System

3.2.2 Location Accuracy

The latitude and longitude of a well listed in the database do not perfectly represent the true location of that well. Different location methods have different degrees of precision and accuracy. The TWDB Ground-Water Data System assigns a numerical code to each well location, indicating the reliability of the given coordinates. The meanings of these codes are summarized in Table 3.5, which also lists the number of wells and associated measurements falling into each accuracy group.

Table 3.5 Location Accuracy Codes

Code	Accuracy	# wells	# measurements
1	± 1"	12,180	22,049
2	± 5"	2,832	4,801
3	± 10"	3,814	4,936
4	± 1'	12	17
5	*	5,628	7,412
none	unknown	4,779	7,260

*—latitude and longitude are given for center of 2.5' quadrangle

A location method code of 5 indicates that the given latitude and longitude are for the center of the 2.5' quadrangle, rather than the well itself. The TWDB states that this is a temporary measure, necessary to include wells listed in an older database that did not require latitude and longitude for well records. Nearly 20% of the wells included in the study (and 16% of the nitrate measurements) can be located only by 2.5' quadrangle.

3.2.3 Selected Aquifers

Wells and nitrate measurements were grouped for statistical and spatial analysis primarily by their location in the 7.5' quadrangles numbered according to the system described in Section 3.2.1. A subset of the wells and measurements selected for further examination were grouped by association with five aquifers, the Carrizo-Wilcox, the Balcones Fault Zone of the Edwards, the Hueco-Mesilla Bolson, the Ogallala, and the Seymour. The TWDB designates these as Major Aquifers, meaning that they supply "large quantities of water in large areas of the State" (Ashworth and Flores, 1991).

The field "Aquifer Code" in the Texas Groundwater Data System "is adopted from U.S. Geological Survey's WATSTORE Data File. The code consists of three digits designating the geologic Era, System, and Series followed by a four or five [character alphabetic] code designating the aquifer(s) or stratigraphic unit(s)" (Nordstrom and Quincy, 1992).

For example, the code "124WLCX" refers to the Wilcox Group, which belongs to the Cenozoic Era, the Tertiary System, and the Paleocene Series. The code has been modified to describe wells in ambiguous settings, or which draw water from more than one formation or aquifer (Nordstrom, 1994). For example, the code "110AVQW" refers to a combination of alluvium, Queen City Sands, and the Wilcox Group.

Based on the aquifer delineation criteria described by Ashworth and Flores (1991), and geologic descriptions from the Geologic Atlas of Texas (BEG, various years), wells were assigned to aquifer groups according to the TWDB aquifer codes listed in Table 3.6. Note that a well was assigned to an aquifer

group only if the TWDB code associated it with a single formation or aquifer. A well with the code "110AVQW" was not assigned to the Carrizo-Wilcox, because it is associated with alluvium and the Queen City Sands as well as the Wilcox Group. The number of wells and measurements associated with these aquifers are summarized in Table 3.7.

Table 3.6 Aquifer Codes

Aquifer	**TWDB Codes**
Carrizo-Wilcox	124CRRZ 124WLCX 124CZWX 124CZWXA
Edwards (Balcones Fault Zone)	218EBFZA
Hueco-Mesilla Bolson	112HCBL 112MSBL
Ogallala	121OGLL
Seymour	112SYMR

Table 3.7 Wells and Measurements in Selected Aquifers

Aquifer	**Wells**	**Measurements**
Carrizo-Wilcox	2292	4597
Edwards (BFZ)	412	1691
Hueco-Mesilla Bolson	404	1908
Ogallala	3483	4430
Seymour	1993	2526

The five aquifers are shown in Figure 3.5. The map was created by combining the outlines of the aquifers from five GIS coverages prepared by TWDB, and represents that agency's estimate of the extent of the aquifers on surface and the limits of the unexposed (downdip) regions that provide usable water. Brief descriptions of the aquifers follow.

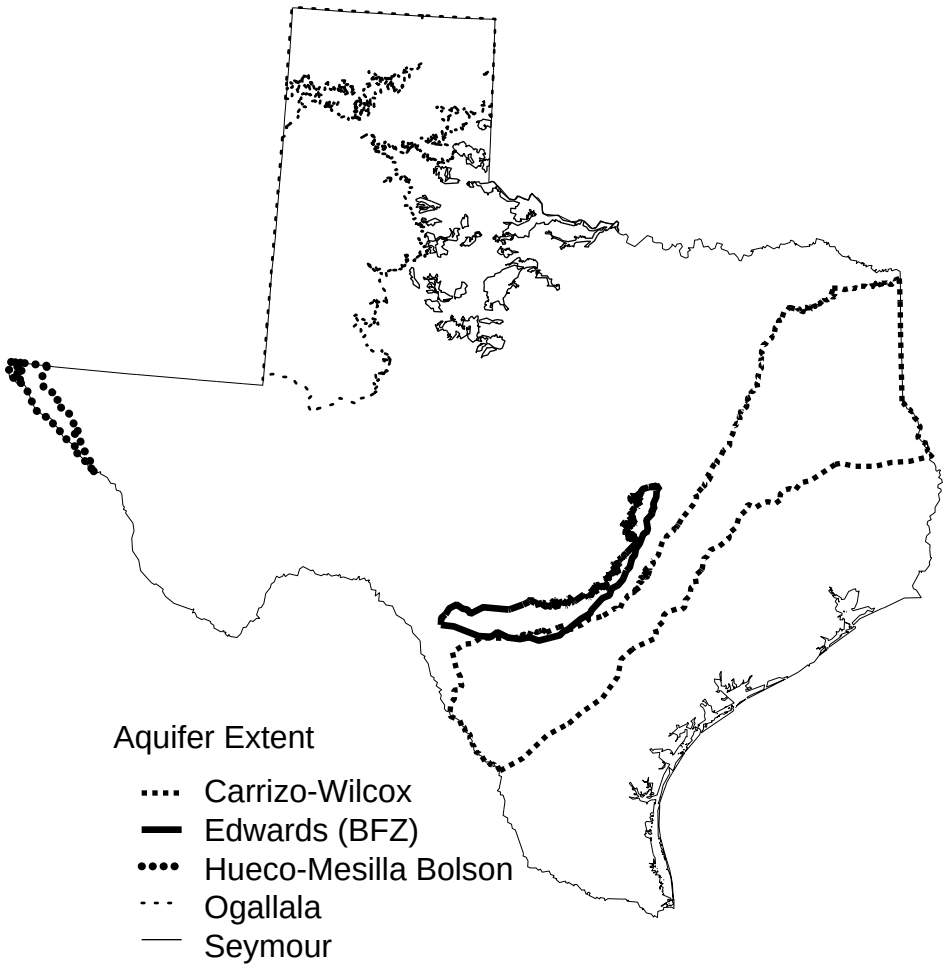


Figure 3.5 Boundaries of Study Aquifers as Identified by TWDB

Carrizo-Wilcox Aquifer. "The Carrizo-Wilcox aquifer includes the Carrizo Formation and the entire Wilcox Group. It extends across the State from Mexico to Louisiana" (Ashworth and Flores, 1991). The Carrizo Formation consists primarily of quartz sand, feldspar, and sandstone (BEG, 1974a and 1968). The Wilcox Group consists primarily of quartz sand, mudstone, clay, and silt (BEG, 1974 and 1968). The TWDB aquifer codes selected for this aquifer group are "124CRRZ" for Carrizo Sand, "124WLCX" for Wilcox Group, "124CZWX" for Carrizo Sand and Wilcox Group—Undifferentiated, and "124CZWXA" for Carrizo Wilcox Aquifer. (Norstrom and Quincy, 1992).

Edwards Aquifer (Balcones Fault Zone). "The Edwards (BFZ) aquifer consists of all the units formations and other members below the Del Rio Formation and above either the Glen Rose Limestone or, when it is present, the Walnut Formation." (Ashworth and Flores, 1991). The Balcones Fault Zone of the Edwards Aquifer is made up of a variety of limestone formations with some included dolomite and shale (BEG, 1974a and 1974b). The TWDB aquifer code selected for this aquifer group is "218EBFZA" for Edwards and Associated Limestones—Balcones Fault Zone.

Hueco-Mesilla Bolson Aquifer. "The Hueco-Mesilla Bolson aquifer consists of Cenozoic alluvial and bolson deposits that occur within the valleys that flank the Franklin Mountains; and extend north and west into New Mexico, and south into Mexico… Although hydrologically connected, the aquifer does not include the overlying Rio Grande alluvium." (Ashworth and Flores, 1991). The Hueco and Mesilla deposits include alluvium and "fluviatile deposits of clay, silt, sand and gypsum in bolsons" (BEG, 1993). The TWDB aquifer codes selected for this

aquifer group are "112HCBL" for Hueco Bolson Deposits and "112MSBL" for Mesilla Bolson Aquifer.

Ogallala Aquifer. "The Ogallala aquifer consists primarily of the Ogallala Formation and extends north, west, and east into adjacent states. The boundary of the formation is mapped along the eastern High Plains escarpment and along the Canadian River Valley, where the formation outcrop is in contact with underlying formations of Cretaceous, Triassic, or Permian age. The southern extent is placed at the estimated formation pinchout" (Ashworth and Flores, 1991). The Ogallala Formation consists of "fluviatile sand, silt, clay, and gravel capped by caliche" (BEG, 1967). The TWDB aquifer code selected for this aquifer group is "121OGLL" for Ogallala Formation.

Seymour Aquifer. "The Seymour aquifer occurs in isolated, eroded alluvial remnants in north-central Texas. The areas delineated are based on surface extent, well development and usage. Consequently many smaller remnants that provide little water or are not developed, are not mapped" (Ashworth and Flores, 1991). The Seymour Formation consists of "Thick deposits… mostly sand, silty orange-brown to red, thick-bedded, massive, locally with large-scale cross-beds and gravel" (BEG, 1987). The TWDB aquifer code selected for this aquifer group is "112SYMR" for Seymour Formation.

3.2.4 Well Description Accuracy

In addition to the location of the well, the accuracy of a well's depth and aquifer code are of particular interest to this study. The histogram of well depths less than 200 feet shown in Figure 3.6 illustrates the overabundance of reported well depths equal to zero or integer multiples of 10 feet. Well depths are often

reported by drillers or well owners, who may not always respond to data requests with scientific precision. Although the TWDB Ground-Water Data System Data Dictionary does not say so, the large number of zero depths suggests that zero may mean "no data" in many cases. The assignment of aquifer codes usually comes from a geologist's interpretation of driller's logs, or from data provided by an agency other than the TWDB, such as the U. S. Geological Survey or various state water districts, that provide well data to the TWDB. This process is not under a uniform quality-control program, and is certainly subject to some errors. However the number of erroneous classifications should be expected to be small in comparison to the database as a whole (Nordstrom, 1994).

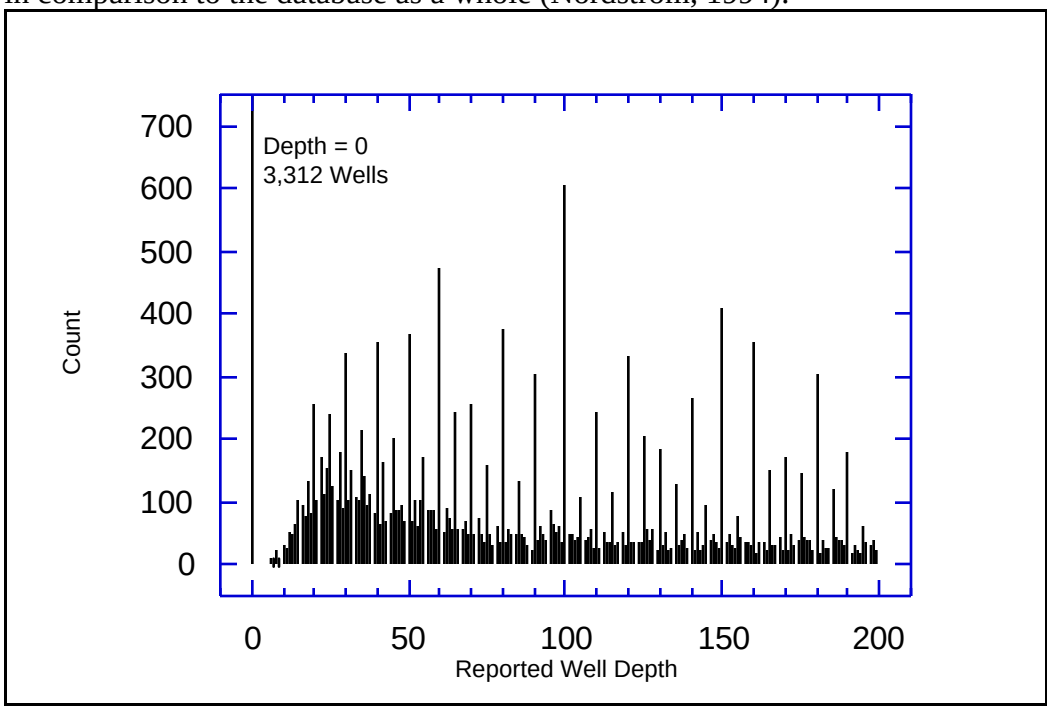


Figure 3.6 Well Depths (less than 200 feet)

The well description data included in the study, like the nitrate measurement data, were accepted "as is" without many exclusions. This does not mean that the data is considered error-free, but reflects the belief that the

quantity of data is large enough that individual errors will not significantly effect the study's conclusions.

Well description records were excluded from the study if the well's latitude and longitude lay outside the quadrangle indicated by its ID number (290 records), or if no nitrate measurements from that well were left in the nitrate measurement table after the deletion of unsuitable records (9,485 records, including the mis-located wells). These deletions left 29,255 well description records in the study.

3.3 SOIL DATA

The soil data used in this study comes from the U. S. Department of Agriculture's State Soil Geographic Database (STATSGO) (USDA, 1993). This rather complex data set has two major components: maps—represented in a GIS—and several related database tables. This study draws data from the STATSGO map of Texas and three related database tables, the map unit, component, and layer tables. Both the map and the tables are stored and manipulated in Arc/Info. This section describes the organization of STATSGO data and the way that values for two soil parameters, soil thickness and average soil organic matter content, were extracted from the database for use in this study.

3.3.1 STATSGO Map and Data Structure

STATSGO maps are compiled from many sources, including soil survey maps, county and state general soil maps, state major land resource area (MLRA) maps, and LANDSAT images. The soil groups shown in these sources are transferred to USGS 1:250,000-scale base maps and digitized. The basic spatial

unit of organization for STATSGO is the *map unit*, a combination of associated phases of soil series with a minimum size of approximately 6.25 km^2. A map unit is identified by a code (Map Unit ID or MUID) consisting of the two-character abbreviation of the state's name and a three-digit number (for example, TX071). Map units also have names reflecting the soil groups they contain (for example, TX071 is named "Brackett-Purves-Real"). The map units are not all contiguous; the map of Texas contains 4031 polygons classified into 632 map units, so on the average a Texas map unit is made up of six discontiguous polygons. Of the 632 map units in the STATSGO database for Texas, one (TX631) has no associated polygons, and one (TXW, the water group) has no associated soil parameter values. The remaining map units range in area from 10 km^2 to 21,500 km^2, with an average area of 1,082 km^2 and a median area of 570 km^2. The histogram in Figure 3.7 shows that a substantial majority of the map units cover areas of less than 1,000 km^2.

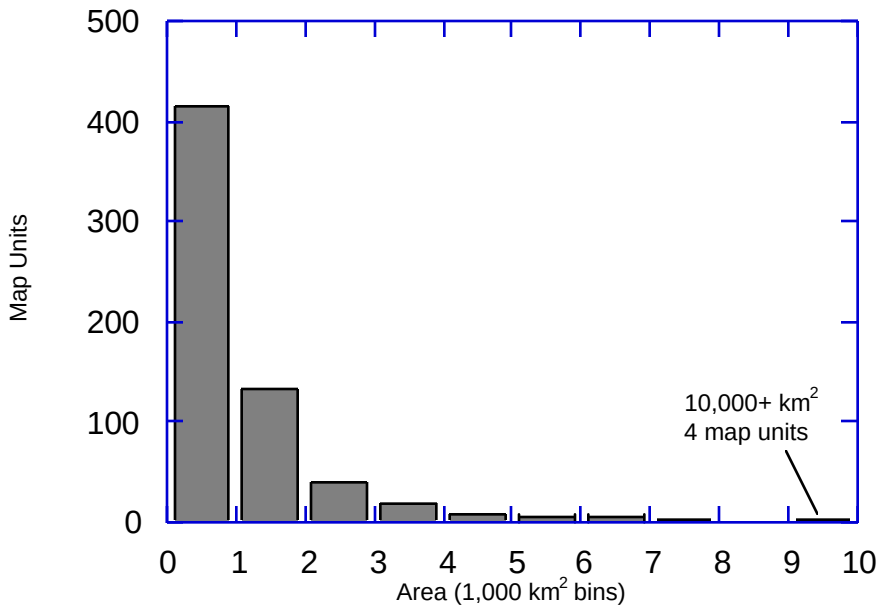


Figure 3.7 Map Unit Area Histogram

The relationship between the polygons, map units, and related tables is illustrated in Figure 3.8 and described in the following paragraphs. (The map units and data shown in Figure 3.8 are made up for purposes of illustration.)

The map units are made up of *components*, also called "soil sequences," or "soil series." Although the STATSGO map does not show components, they—like the map units—are horizontal divisions of the earth's surface, and the area of a map unit is the sum of the areas of the components it contains. Each map unit may contain from 1 to 21 components. In Texas, map units contain an average of 9 components. A component is uniquely identified by a map unit ID and a sequence number. STATSGO assigns 60 properties to the components, and stores

their values in the linked tables, including the component table. In the component table, the area of a component is expressed as a percentage of the map unit area.

The components, in turn, are made up of *layers*, which are vertical divisions of the soil. A component is a sequence of from 1 to 6 soil layers. In Texas, components contain an average of 3 layers. A layer is uniquely identified in the table by the map unit ID, the sequence number, and a layer number. STATSGO assigns 28 properties to each layer, and stores their values in linked tables, including the layer table.

The soil thickness, organic content, and bulk density values used in this study are stored in the layer table. All of these quantities are expressed as ranges, with maximum and minimum values listed in the table. For example, the

minimum depth of the top layer in a component is zero, and the maximum depth of the bottom layer in a component is equal to the thickness of the component.

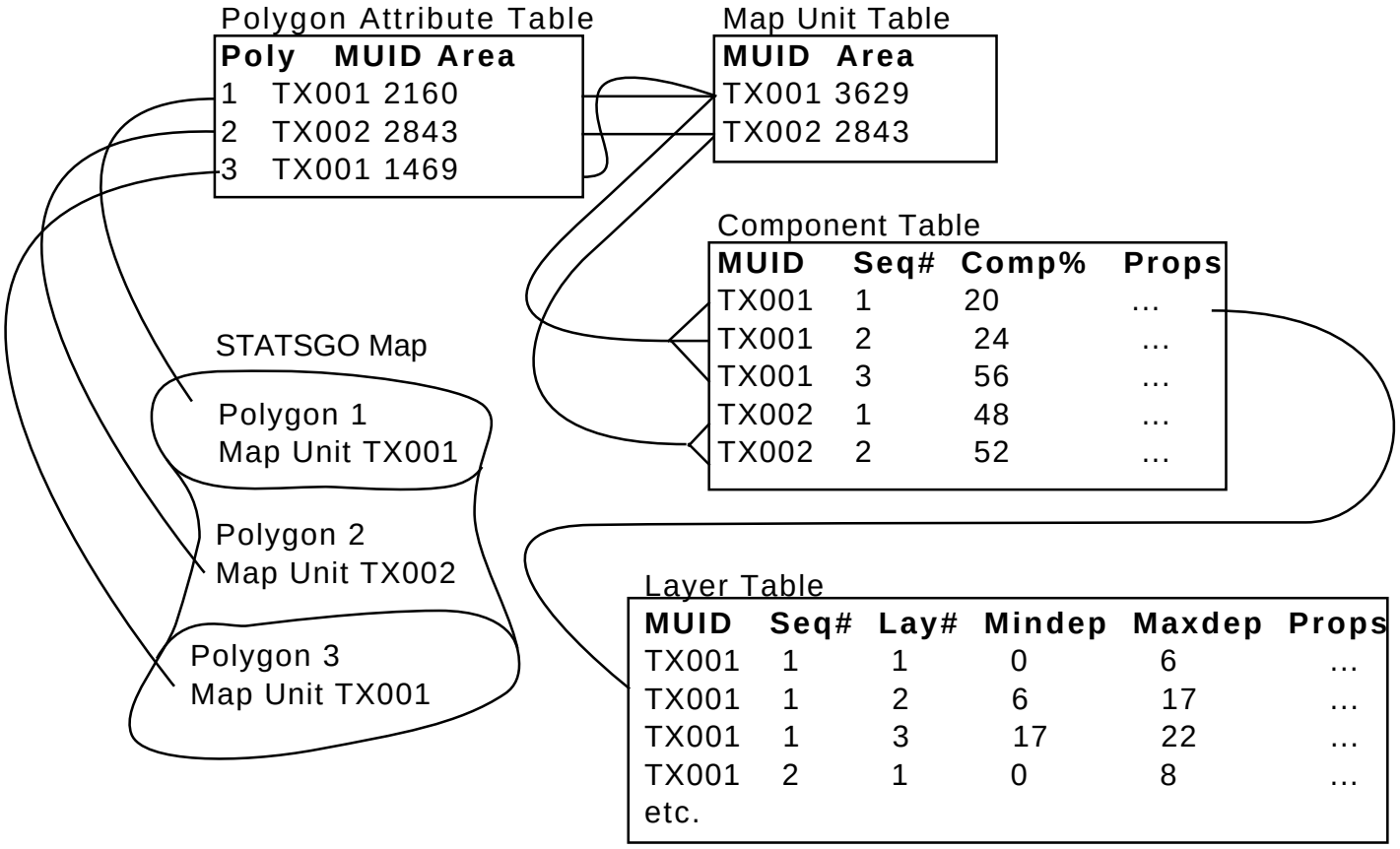


Figure 3.8 STATSGO Map and Data Organization

3.3.2 Using STATSGO Data

Figure 3.9 shows excerpts from the STATSGO map of Texas, giving some idea of the spatial structure of the map units. The area falling in the 1_ quadrangle between 30_ and 31_ N latitude and 98_ and 99_ W longitude (1_ quadrangle number 57 in the TWDB well-numbering system) is divided into roughly 140 polygons, which belong to 18 map units. The selected 7.5' quadrangle (number 5740) contains parts of two map units, which have identification codes "TXW" and "TX071." TXW is the code for all bodies of water in the state (in this case, part of Lake Travis), and TX071 is the "Brackett-Purves-Real," map unit . The soil series (also called "components") that make up TX071 are listed in Table 3.8.

Table 3.8, extracted from the component table, shows, for example, that the Purves soil series makes up 13% of map unit TX071. Table 3.9, extracted from the layer table, shows values for minimum and maximum layer depths in inches and minimum and maximum organic material content . The Purves series consists of three layers, which are 12, 2, and 6 inches thick, respectively. The total depth of the Purves series is thus 20 inches.

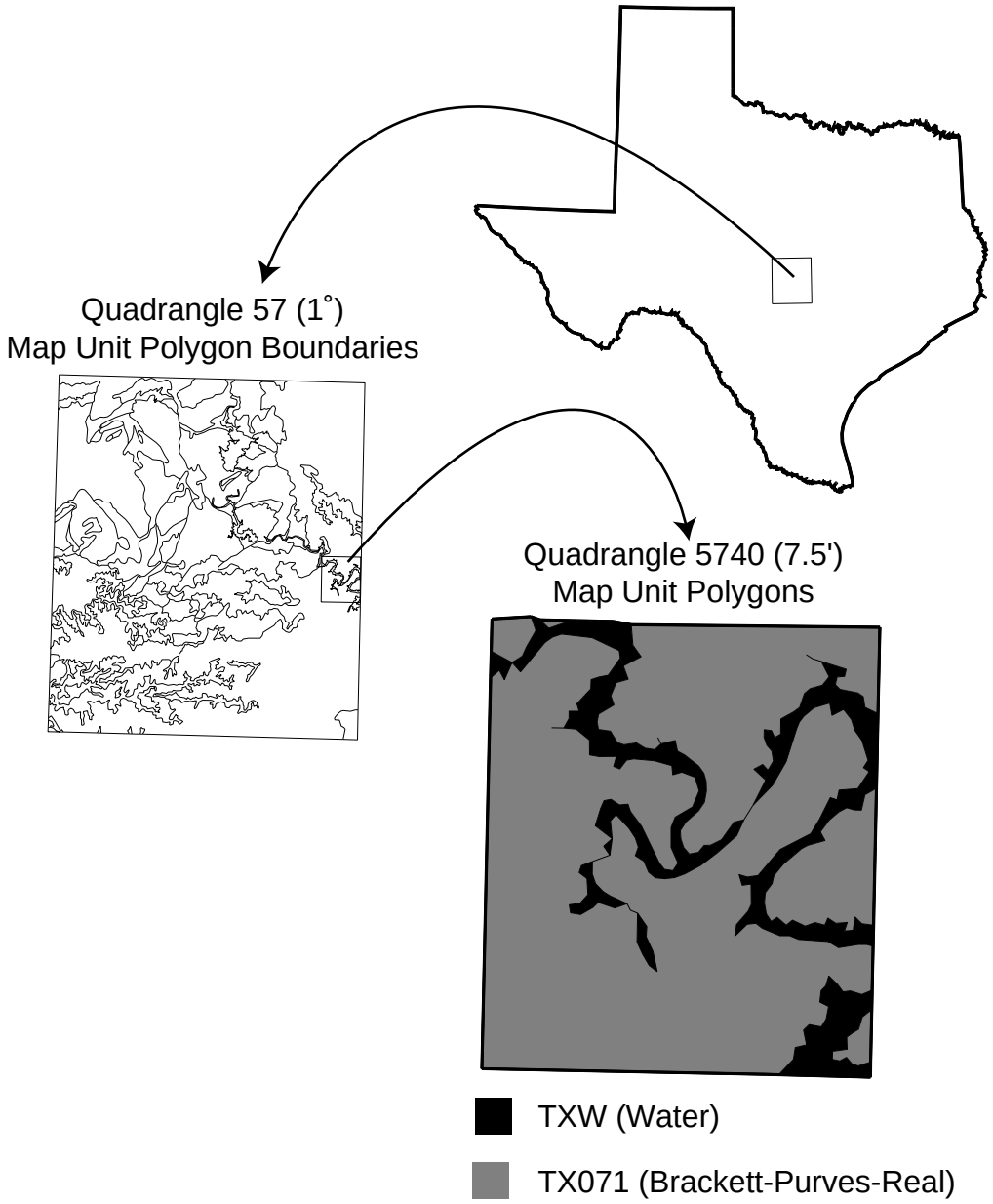


Figure 3.9 STATSGO Map Units

Table 3.8 Soil Series in Map Unit TX071 "Brackett-Purves-Real"

Seq. #	**Seq. Name**	**Comp %**
1	BRACKETT	28
2	BRACKETT	12
3	PURVES	13
4	REAL	7
5	REAL	3
6	ROCK OUTCROP	3
7	ROCK OUTCROP	3
8	COMFORT	6
9	BOLAR	4
10	DOSS	4
11	KRUM	4
12	ALEDO	5
13	OAKALLA	2
14	GRUENE	1
15	ECKRANT	2
16	BOLAR	1
17	SUNEV	1
18	TARPLEY	1
--	TOTAL	100

Table 3.9 Layers in Purves Component of Map Unit TX071

sequence number	**layer number**	**min. depth** (inches)	**max. depth** (inches)	**min. organic matter** (%)	**max. organic matter** (%)	**min. bulk density** (g/cm^3)	**max. bulk density** (g/cm^3)
3	1	0	12	1	4	1.25	1.45
3	2	12	14	1	2	1.25	1.45
3	3	14	20	0	0	0	0

Calculating the average organic material content for the layer requires more computation than the layer thickness. Organic matter is expressed as a percentage of soil mass, and must be multiplied by the bulk density of the soil to produce an organic mass density. For each layer, the average organic content and bulk density can be estimated as the midpoint between the minimum and maximum values (2.5%, 1.5%, and 0% organic matter, and 1.35, 1.35, and 0 g/cm^3, respectively). Multiplying these values by the layer thicknesses and summing over the layers produces an estimate of the organic material per unit area in the component.

$$M \approx \sum_{i=1}^{n} b_i \frac{(o_{min.} + o_{max.})_i}{2} \frac{(\rho_{min.} + \rho_{max.})_i}{2} \qquad (3\text{-}1)$$

where M is the density of organic matter (g/cm^2) for the component, b_i is the thickness (cm) of the layer, o is the weight percentage (by weight) of organic matter in the layer, ρ is the bulk density (g/cm^3) of the layer, and n is the number of layers in the component. A factor of 10 is used to convert g/cm^2 to kg/m^2. Table 3.10 shows how the organic content in the Purves series was calculated to be 11.32 kg/m^2. Note that the organic matter content for the component is expressed as a density by area, rather than volume because the organic content has been integrated over the depth of the soil.

Table 3.10 Derived Values for Soil Organic Content in Purves Series (Map Unit TX071)

sequence number	**layer number**	**thickness** (cm)	**mid bulk density** (g/cm^3)	**mid organic matter** (%)	**organic content** (kg/m^2)
3	1	30.5	1.35	2.5	10.29
3	2	5.1	1.35	1.5	1.03
3	3	15.2	0	0	0.00
3	all	50.8	--	--	11.32

Table 3.11 Soil Series Parameters for Map Unit TX071

Seq. #	**Seq. Name**	**Comp %**	**Thickness** (inches)	**Avg. om** (kg/m^2)
1	BRACKETT	28	60	23.26
2	BRACKETT	12	60	23.26
3	PURVES	13	20	11.32
4	REAL	7	36	9.07
5	REAL	3	36	9.07
6	ROCK OUTCROP	3	80	0
7	ROCK OUTCROP	3	80	0
8	COMFORT	6	20	4.44
9	BOLAR	4	44	18.86
10	DOSS	4	48	13.03
11	KRUM	4	72	28.61
12	ALEDO	5	20	5.83
13	OAKALLA	2	60	11.18
14	GRUENE	1	80	0
15	ECKRANT	2	30	17.86
16	BOLAR	1	44	18.86
17	SUNEV	1	72	22.69
18	TARPLEY	1	22	11.02
unit	TX071	100	48	15.77

Table 3.11 shows the calculated soil thicknesses and organic matter for the components of TX071. The map unit values shown on the last line of Table 3.11 are area-weighted averages, calculated by summing the products of the parameter values and the component percentages. Although values can be calculated for the soil parameters at both component and map unit levels, only the map unit averages can be located on the STATSGO map. For example, the Purves series makes up 13% of map unit TX071, but STATSGO provides no information about which 13% that is. For this reason, the STATSGO data cannot properly be applied to any areas but the STATSGO map units.

STATSGO's relatively poor spatial resolution presents a difficult problem for users of the data. In this study, the well and water quality data are organized on spatial units of 2.5' quadrangles, which are much smaller than STATSGO map units. Figure 3.10 shows the relative sizes of map unit TX071, a 1_ quadrangle, a 7.5' quadrangle, and a 2.5' quadrangle. Map unit TX071 covers about 6,700 square kilometers; in the same part of the state, a 1_ quadrangle covers about 10,000 square kilometers, a 7.5' quadrangle covers about 166 square kilometers, and a 2.5' quadrangle covers about 18.5 square kilometers. A 2.5' quadrangle is roughly the same size as the Oakalla component of map unit TX071.

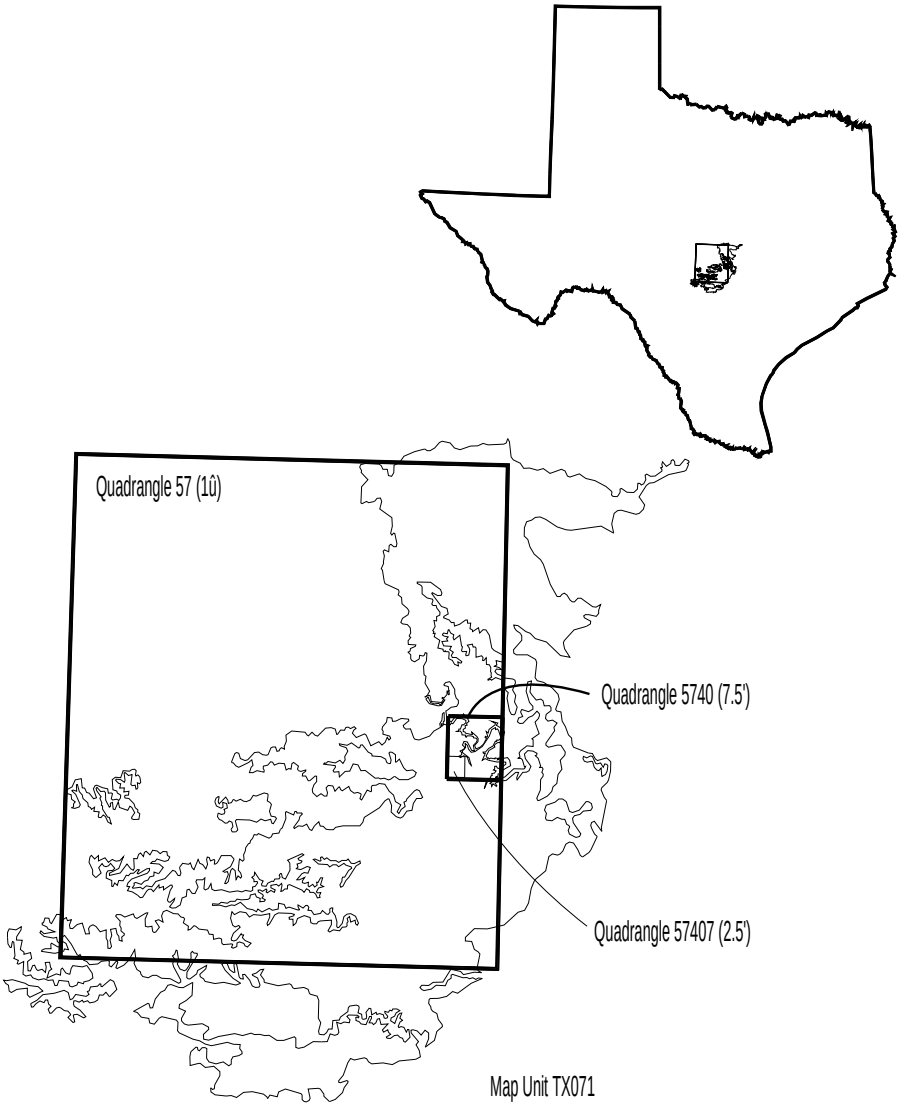


Figure 3.10 Map Unit TX071 with
Quadrangles for Size Comparison

Applying map unit values to areas other than the map units themselves—such as 2.5' quadrangles—requires the user to assume a spatial distribution of the soil series within the map units. The simplest assumption, and the best available without requiring supplementary data, is that the area-weighted averages of soil parameter values are uniformly distributed properties of the map units. This assumption contradicts fact, and the STATSGO user's guide specifically warns against it.

In spite of this warning, this study employs just this assumption. This use of the data can be justified on a variety of grounds. First, this study seeks to describe the variation of water quality through Texas using a database organized in 2.5' quadrangles. The STATSGO map units are organized in different divisions of the land surface and the two systems are irreconcilable; one must be compromised. Since the well data are primary, compromise of the STATSGO data must be tolerated. Secondly, the map units, by their nature, are groups of associated soils, so the variation in soil properties between map units ought to be greater than the variation within map units. Thirdly, since this is a statewide study, it is reasonable to assume that the errors introduced by mishandling the STATSGO data small enough that they will not significantly influence the conclusions drawn over so large a study area.

Using this compromise, soil parameters will be estimated by the following procedure. Any region (e.g., a 7.5' quadrangle) lying entirely within the boundaries of a STATSGO map unit will be assigned the average parameters for that map unit. Any region that crosses STATSGO map unit boundaries will be assigned soil parameter values equal to the area-weighted average of the

values associated with the non-water map units that lie within the region. For example, since quadrangle 5740 is composed entirely of water (TXW) and portions of map unit TX071, it would be assigned values equal to the averages for TX071.

3.3.3 Range and Distribution of Soil Parameter Values

The average soil thickness in the non-water map units ranges from a minimum value of 22.4 inches to a maximum of 88 inches. The area-weighted average of the soil thickness is 65.2 inches, and the median values is 69.9 inches. The histogram-like chart in Figure 3.11 shows the map unit area associated with ranges of soil thickness in 5-inch bins. The distribution of soil thickness over the surface of Texas is illustrated in Figure 3.12.

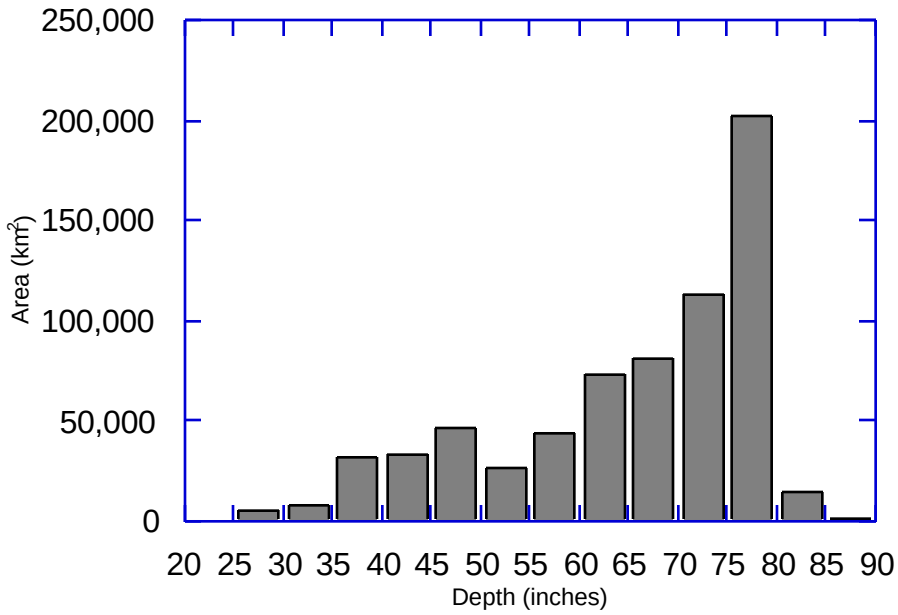


Figure 3.11 Soil Depth Histogram

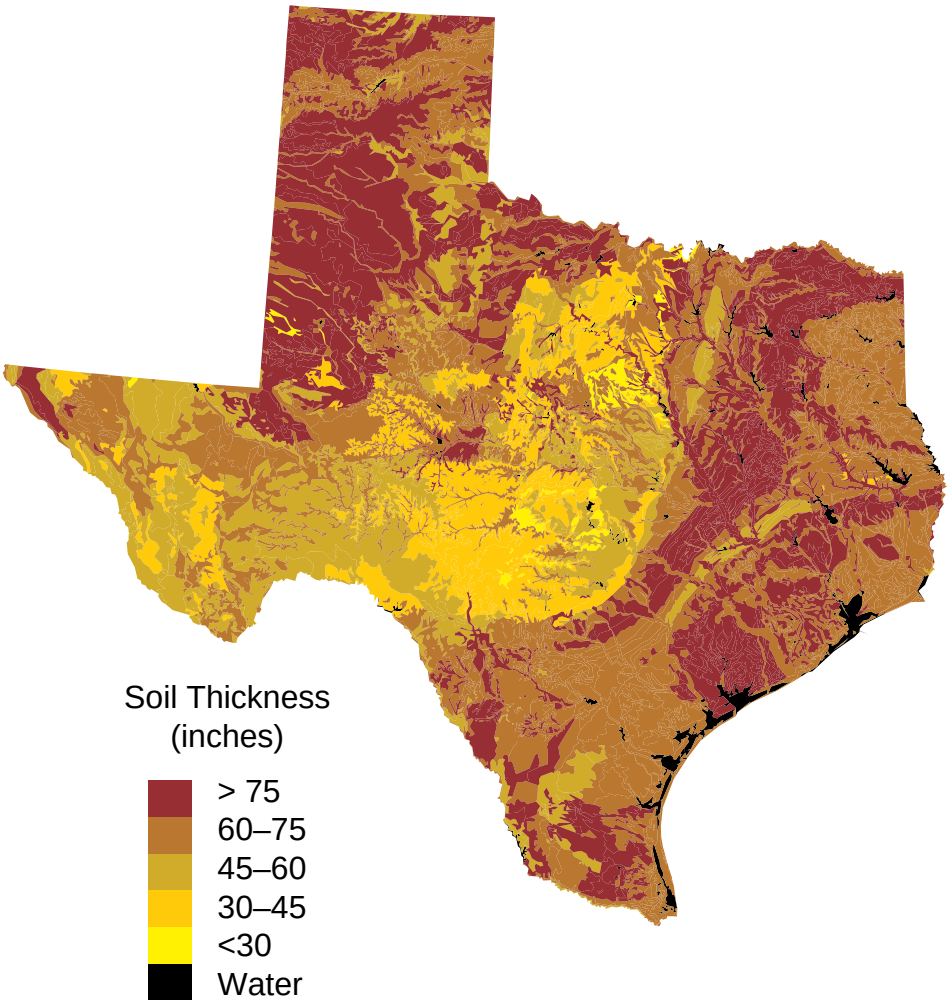


Figure 3.12 Spatial Distribution of Soil Thickness

The average soil organic content in the non-water map units ranges from a minimum value of 0.76 kg/m^2 to a maximum of 74.9 kg/m^2. The area-weighted average of the soil organic content is 16.2 kg/m^2, and the median values is 15.1 kg/m^2. The histogram-like chart in Figure 3.12 shows the map unit area associated with ranges of soil organic content in 5-kg/m^2 bins. The distribution of soil organic content over the surface of Texas is illustrated in Figure 3.13.

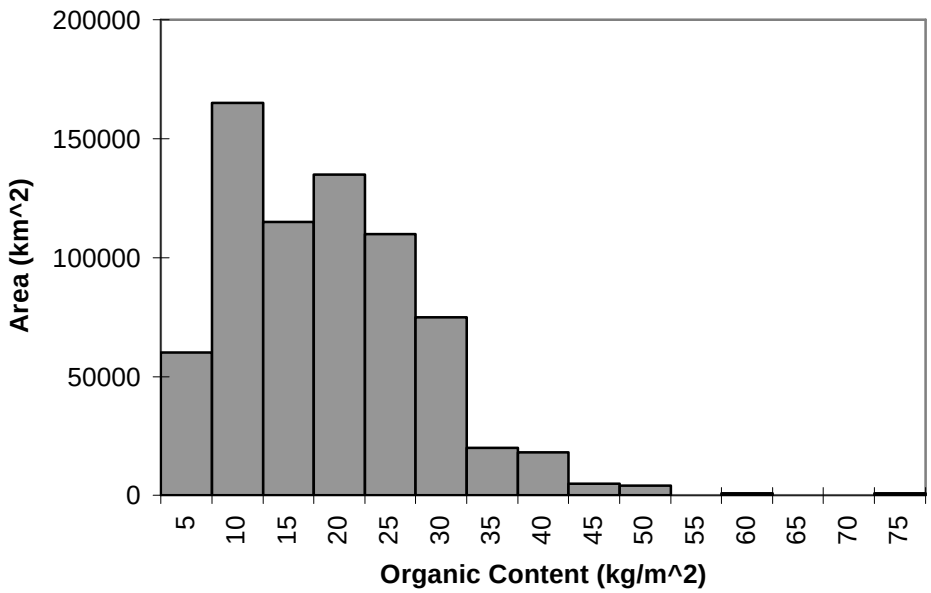


Figure 3.13 Soil Organic Matter Histogram

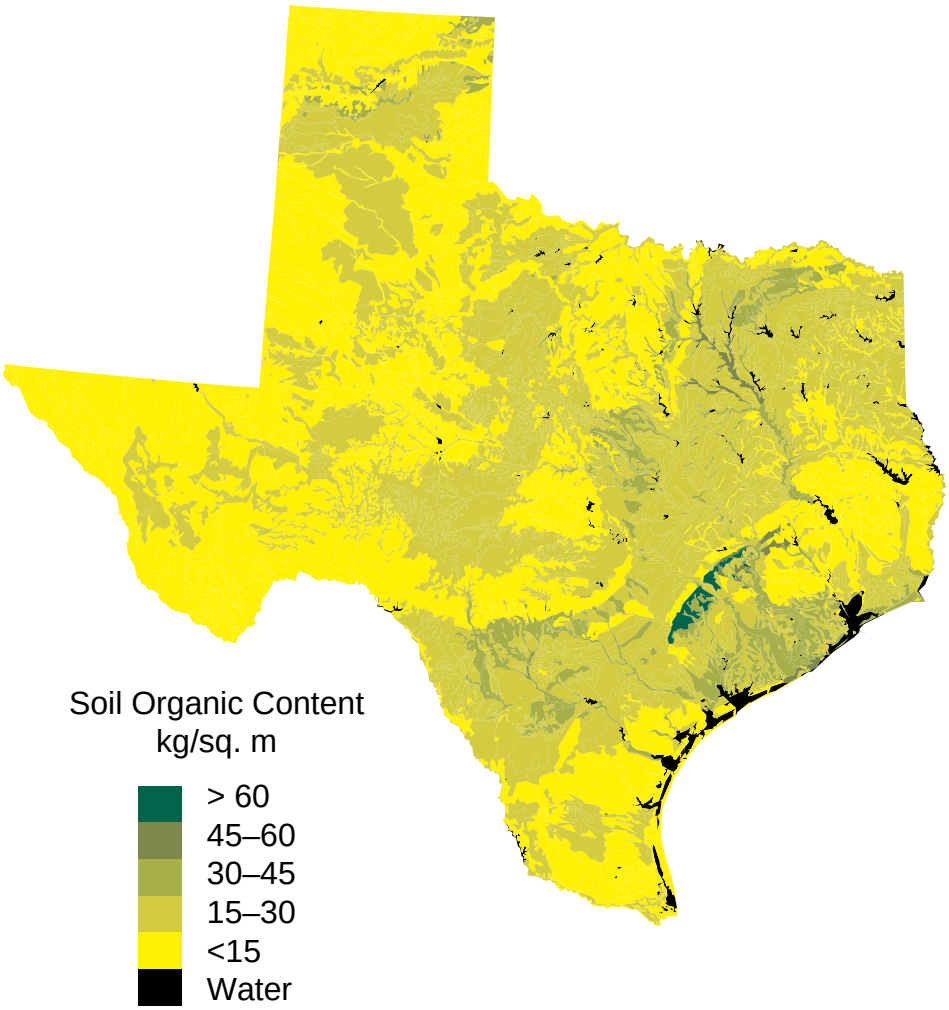


Figure 3.14 Spatial Distribution of Soil Organic Content

3.4 PRECIPITATION DATA

The precipitation data used in this study were copied from, or derived from data included in Hydrosphere Inc.'s *Climatedata* CD-ROMs (Hydrosphere Data Products, Inc., 1994). This data set consists of GIS coverages showing point locations of the observation stations, and database tables listing the daily observations of climatic data for the period of record of the TD-3200 Summary of the Day Cooperative Observer Network database of the National Climatic Data Center (NCDC).

3.4.1 Preparation of Annual Average Precipitation Map

The annual average precipitation map used in this study is intended to reflect the variation of expected rainfall across Texas. The objective in preparing the map was not to produce the best possible prediction of average annual precipitation at each station, which would require that the entire period of record be used for each station, but rather to produce the best estimate of the relative magnitudes of precipitation at different stations, which requires that the same period be reported for all stations.

This goal sets up an interesting set of conflicting requirements. For any map, including more points improves the spatial resolution, and for any time series, extending the period of record increases confidence in the calculated average values. Requiring that the period of record be the same for all stations means that stations operating for only a part of the period cannot be included in the map, so a longer period of record leads to fewer points, and vice versa.

After a trial-and-error exploration of the data, the following criteria were used to select the data for the map used in this study:

1. The period of record for the map extends from 1951 to 1980.
2. A station is deleted from the map if a sequence of than two years is missing from the station's records. (NCDC considers a year "missing" if it contains a missing month. A month is "missing" if more than nine days of data are absent.)

The selected period of record includes periods of both very low precipitation (the early-to-mid 1950s) and very high precipitation (the early 1970s), and can be considered a representative period for precipitation in Texas. Requiring a longer period of record (1951–1990) or tolerating only single-year gaps resulted in roughly 25% reductions in the number of stations included in the map.

The procedure used to generate the precipitation map is described in Section 5.2.3. The resulting map appears in Figure 3.15.

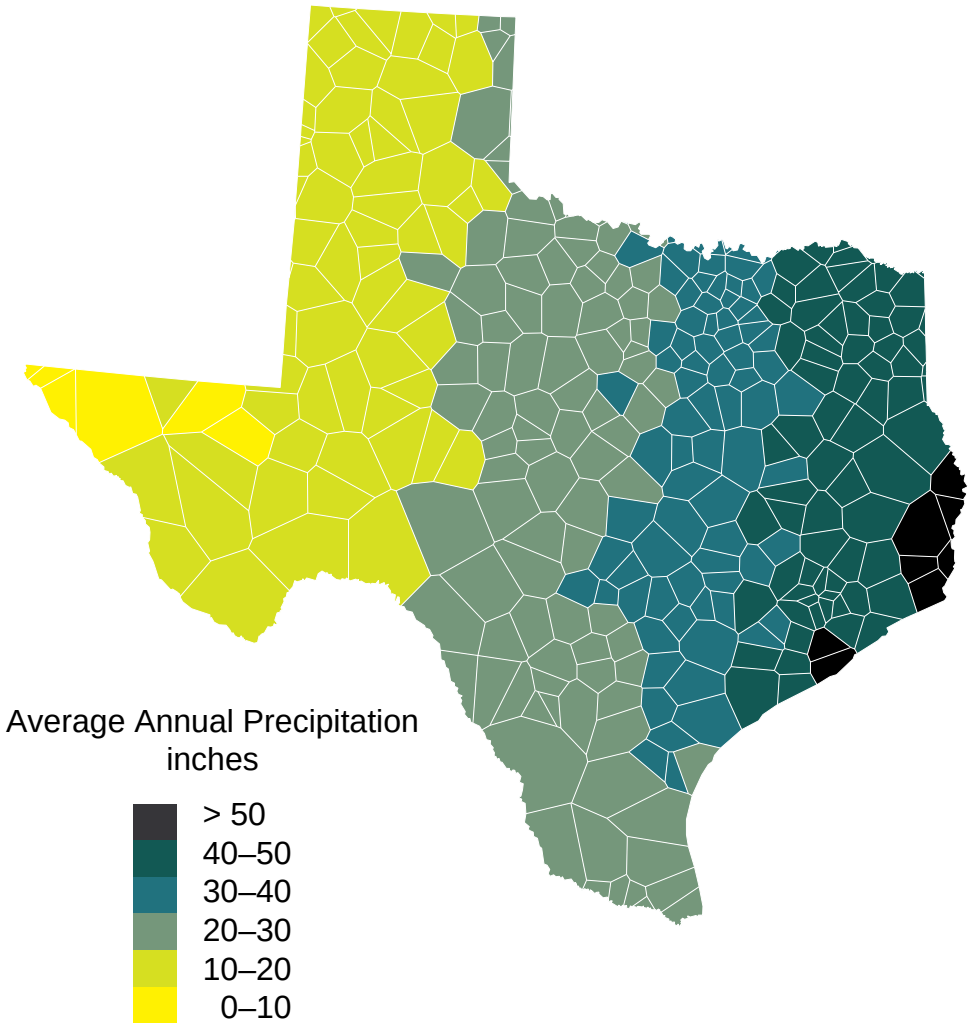


Figure 3.15 Spatial Distribution of Annual Average Precipitation (Thiessen Polygons)

3.4.2 Range and Distribution of Precipitation Data

The Thiessen polygons range in size from a minimum area of about 10 km^2 to a maximum area of about 10,600 km^2, with an average area of 2,130 km^2 and a median area of 1,690 km^2. The size of the polygons is inversely related to the density of gauges and hence to population. Polygons are small around cities and large in the unpopulated areas of west Texas. Figure 3.16 shows the frequency distribution of Thiessen polygon sizes for the study's precipitation gauging network.

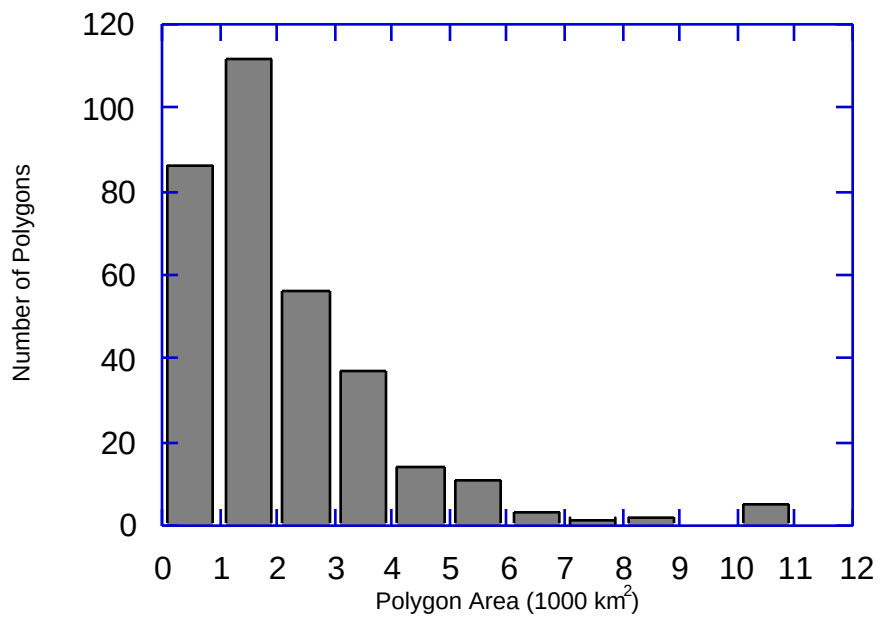


Figure 3.16 Thiessen Polygon Area Histogram

By the reckoning described in section 3.4.1, average annual precipitation ranges from a low of 7.8 inches in El Paso to a high of 59.1 inches in Orange. The area-weighted average precipitation for Texas as a whole is 26.8 inches and

the area-based median is 24.5 inches (meaning that half the area of the state averages more than 24.5 inches of precipitation per year and the other half averages less). The histogram-like diagram in Figure 3.17 shows how the Thiessen polygon area associated with the various levels of precipitation.

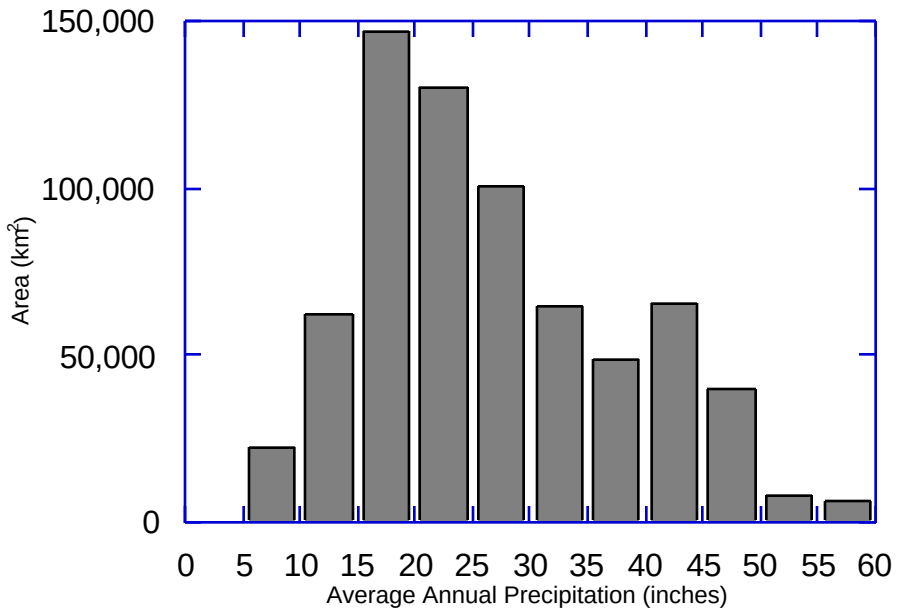


Figure 3.17 Precipitation Histogram

3.5 FERTILIZER SALES DATA

The nitrate fertilizer application data has the poorest spatial resolution of all the data used in this study. Figure 3.18 was generated from annual total fertilizer sales collected nation-wide on a county level by the EPA's office of Policy Planning and Evaluation. Battaglin and Goolsby (1995) related sales figures for the years 1986–1991 to county maps of the United States as part of a project to illustrate nationwide trends in agricultural chemical use with GIS (Mr.

Battaglin made the fertilizer data used in this study available to the author prior to the publication of the cited report). In addition to listing the total number of tons of fertilizer sold in each county, Battaglin and Goolsby divided the tons of fertilizer sold by the total area of the county to compensate somewhat for the range of variation in size of counties. The result is a number that they call "use" in tons per square mile. For the map in Figure 3.17, six years of use were averaged for each county. These averages range from a low of 0, meaning no recorded nitrate sales in the county for the six years, and a high of 18.9 tons per year of recorded nitrate fertilizer sales per square mile of county.

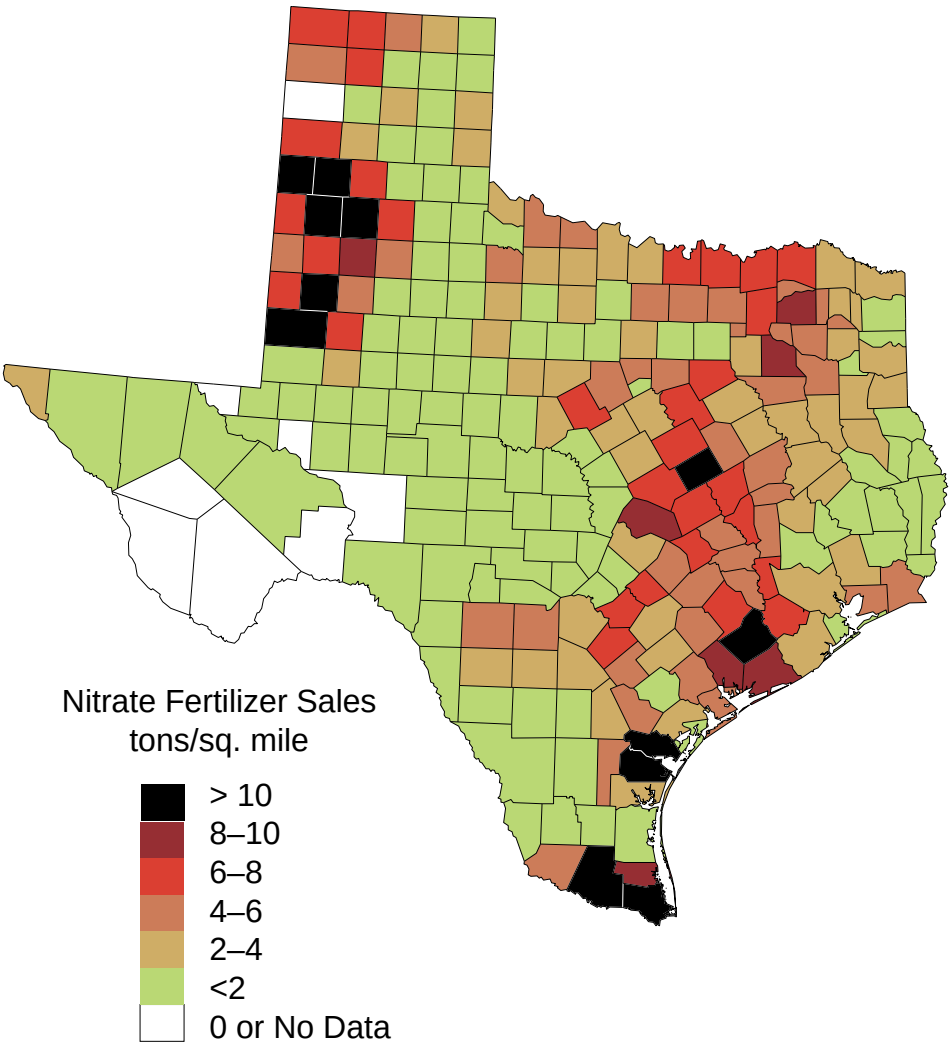


Figure 3.18 Sales of Nitrogen Fertilizers by County (Annual Average 1986–1991)

3.6 WATER UTILITIES DIVISION NITRATE MONITORING DATA

Nitrate measurements collected by the Water Utilities Division (WUD) of the Texas Natural Resource Conservation Commission as part of their Primary Drinking Water Standards enforcement efforts, were used as an independent data set to test nitrate vulnerability predictions based on the TWDB data.

The nitrate measurements reported by the WUD are collected at points of entry to public water distribution systems, i.e., after water from multiple sources has been mixed and treated. A water system may have several points of entry and several wells or surface intakes supplying those points of entry. Water samples from points of entry do not represent individual wells unless the point of entry is tied to only one well.

The data provided by the WUD include nitrate concentrations measured at points of entry, identifications of those points of entry and the wells and surface intakes supplying them, and the locations of the wells. These were represented in two database tables and a GIS coverage. The nitrate measurement table includes the system and point of entry identification for each measurement, along with the date of sample collection and analysis results. The point of entry table contains one record for each well, listing the well ID, system ID, and point of entry ID. (WUD well numbers are not the same as TWDB well numbers. They are based on county and water supply identification, rather than geographic coordinates.) By linking nitrate concentration to points of entry, points of entry to wells, and wells to locations, it is possible to tie nitrate concentrations to quadrangles for comparison to the quad exceedence probabilities calculated from

the TWDB data. The process and results of this comparison are described in Sections 5.8 and 6.4.

3.7 HERBICIDE AND NITRATE DATA FROM MIDWESTERN U.S.

Because of the lack of a sufficient quantity of measurements of herbicides and other man-made agricultural chemicals in Texas groundwater, it is not possible to determine whether vulnerability to nitrate is correlated to vulnerability to other agricultural chemicals in Texas. However, in order to generalize the results of a study of vulnerability to nitrate contamination to other agricultural chemicals, it is necessary to assume some relationship between nitrate and those other chemicals. The data presented by Kolpin et al (1993) is used to test the rather mild assumption that geologic conditions favorable to a high rate of detection of elevated nitrate levels will also be favorable to a high rate of herbicide detections.

The data were collected in 1991 from 300 wells in the Midwestern U.S. The nitrate and herbicide data were collected as part of an effort to characterize the spatial and seasonal distribution of agricultural chemicals in groundwater, and to provide data for an exploratory statistical analysis of the influence of anthropogenic, and geologic and other natural factors on the occurrence of herbicides (Kolpin and Burkart, 1991).

A full account of the reconnaissance can be found in the cited references. The data used here included the reported concentrations of nitrate and nine herbicides or herbicide metabolites (alachlor, atrazine, cyanazine, deethyl-atrazine, deisopropyl-atrazine, metolachlor, metribuzin, prometon, and simazine), and two geologic descriptors of well surroundings (depth to top of

aquifer, and aquifer type—bedrock or unconsolidated). The use of the data is explained in Section 6.5.

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

S*usceptibility, 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.

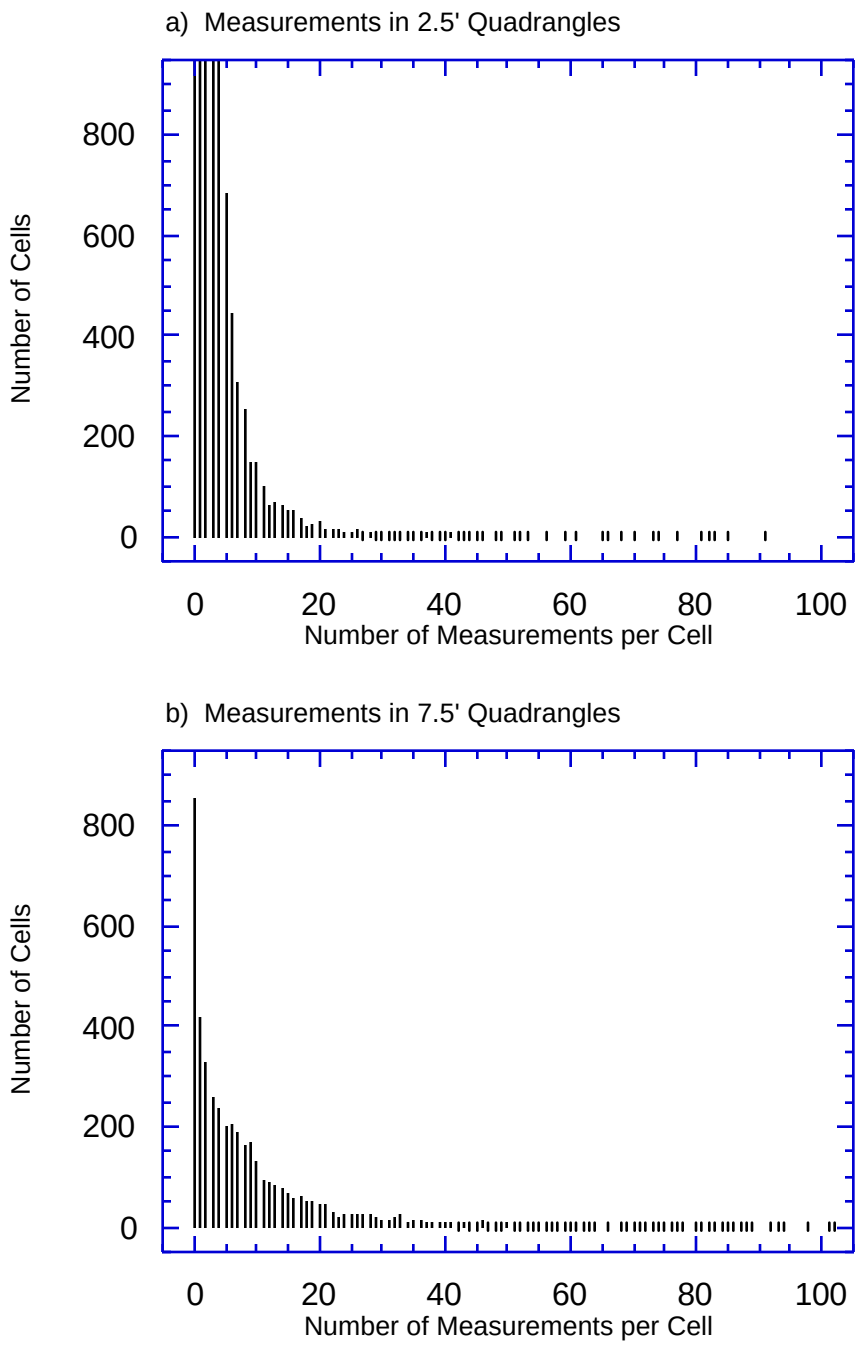


Figure 4.1 Measurement Histograms for 2.5' and 7.5' Quadrangles

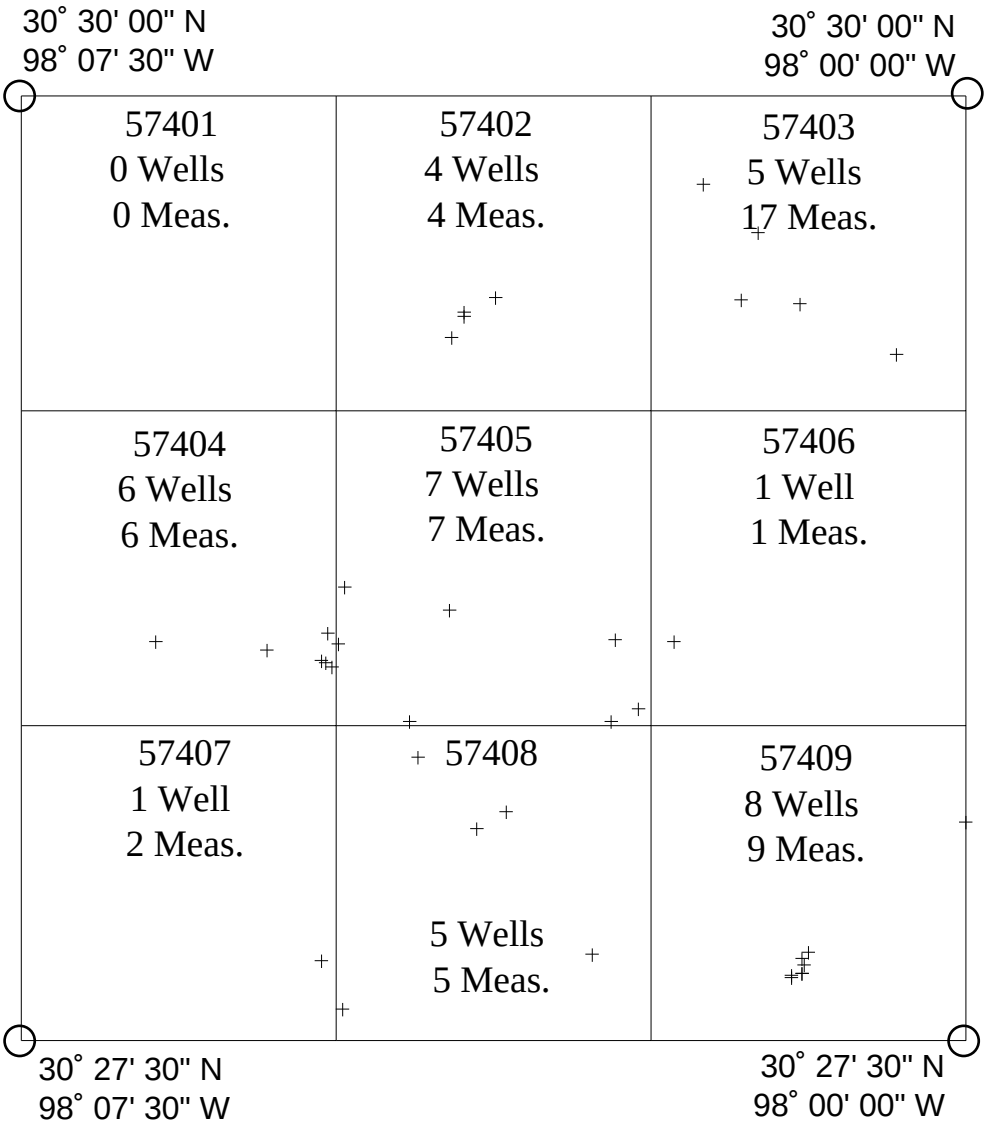


Figure 4.2 Well Locations in Quadrangle 5740

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

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

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.

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

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

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

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

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

$$\textit{well-scheme} = (\text{well-ID, depth}),$$

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

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

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

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

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

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

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

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.

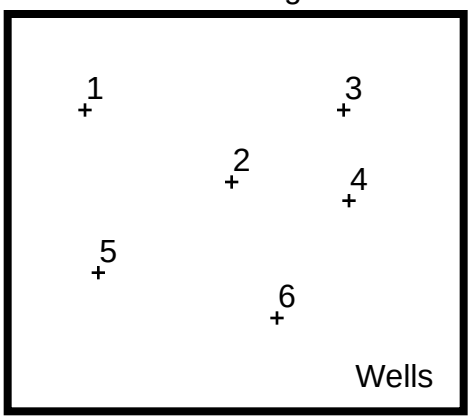
Features in a coverage can be thought of as elements of a set, like the records in a database table. Subsets of objects can be formed on the basis of location, attribute values, or a combination, and set operations such as union or intersection can be performed on these subsets.

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

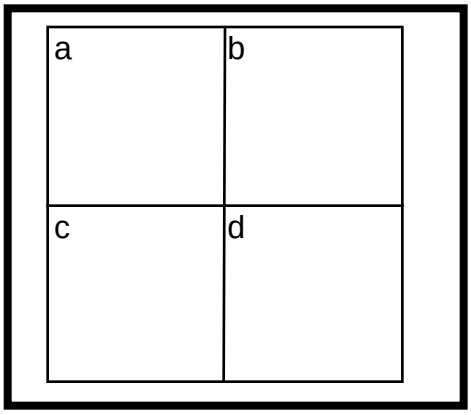
attribute table is omitted from the figure. The two coverages are combined in an *overlay* operation, and the result is shown at the bottom of the figure. Because the topology of the point coverage is unchanged, the result of the overlay is the addition of a new attribute in the point attribute table identifying the quadrangle in which the wells are located. Wells can now be grouped by quadrangle using ordinary database queries.

Point Coverage

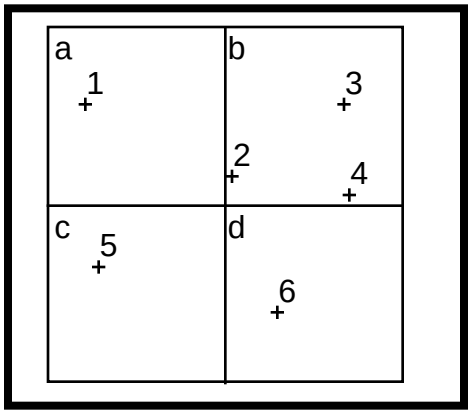


Well-ID	Depth
1	128
2	87
3	214
4	45
5	398
6	414

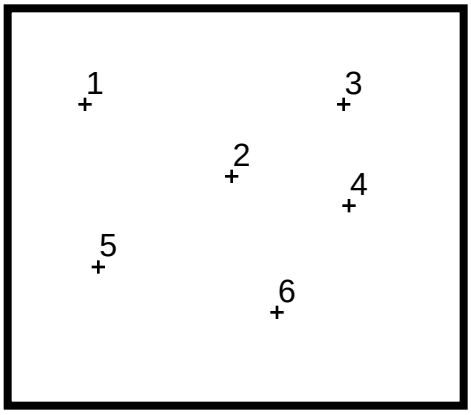
Polygon Coverage



Overlay



Result



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

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

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.

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

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

$$\hat{P} = \frac{s}{n} \quad . \tag{4-9}$$

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, $1-\alpha$, 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 (mg/l nit.-N)	$\hat{p}_e$	P_l 90% two-sided	P_h 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

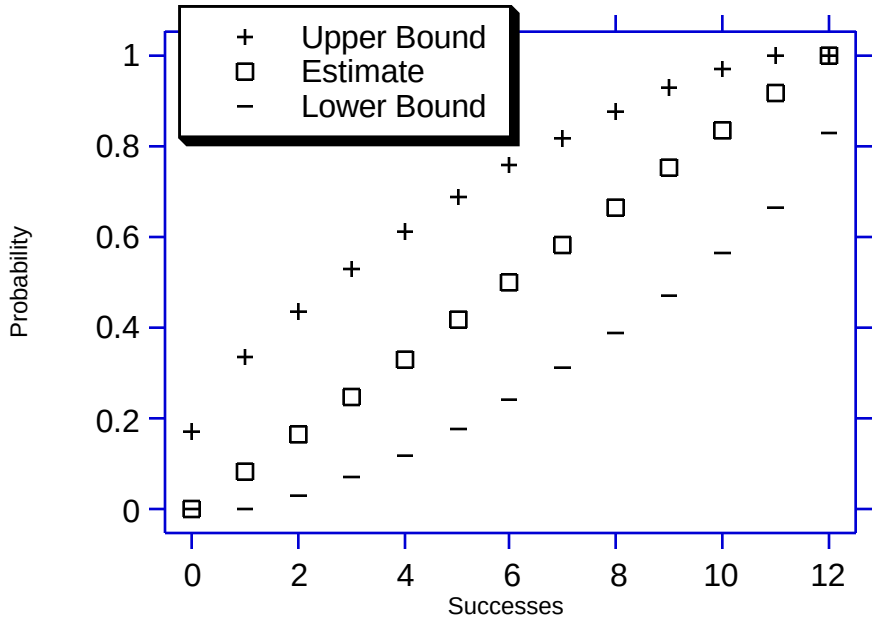


Figure 4.4 Estimating Probability of Success from a Sample of Twelve Trials

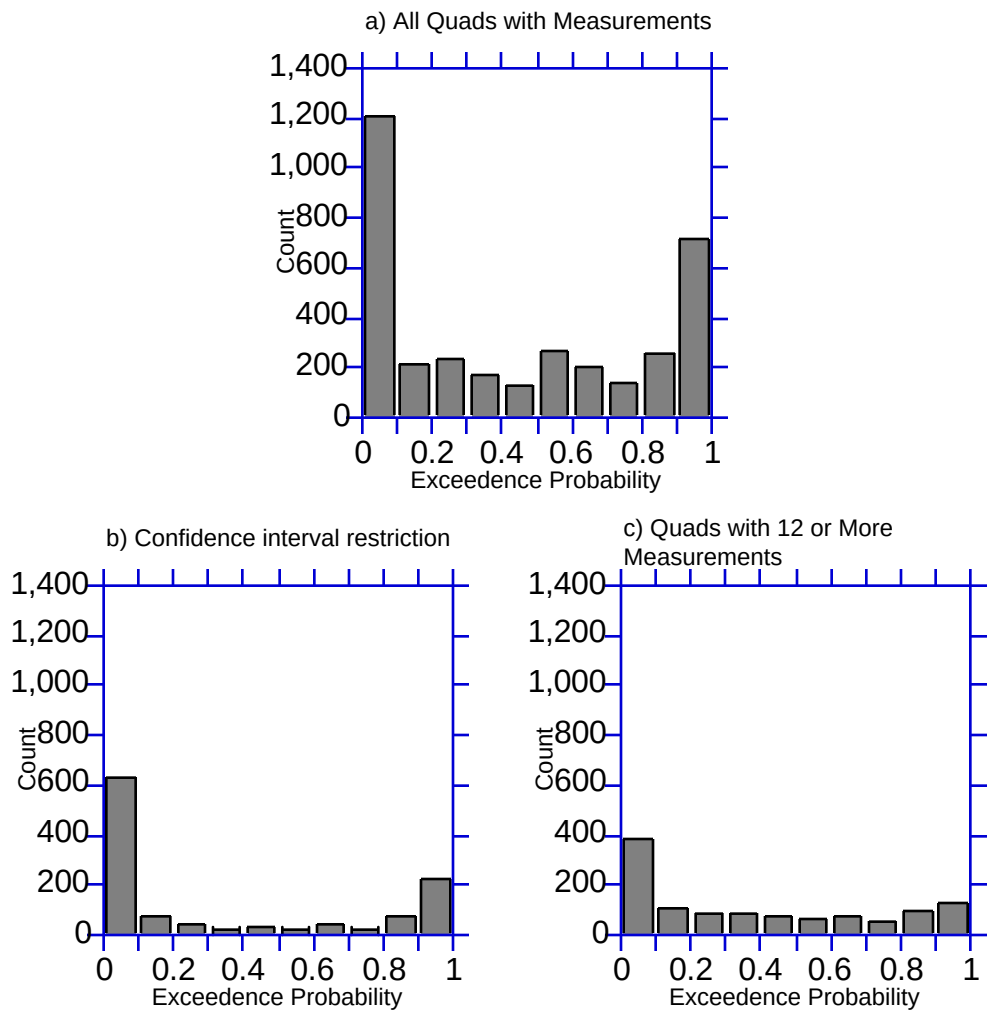


Figure 4.5 Effects of Different Inclusion Criteria

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 $0 < p < 0.5$)

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

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

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

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

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

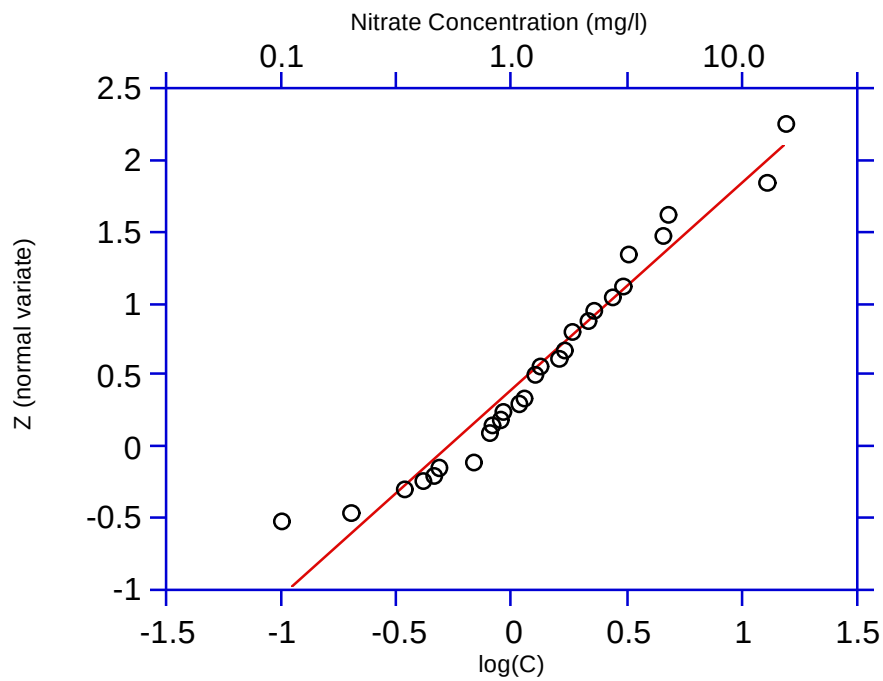


Figure 4.6 Fitting a Probability Distribution by Regression for Quadrangle 5740

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 μ_X is the mean and σ_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.

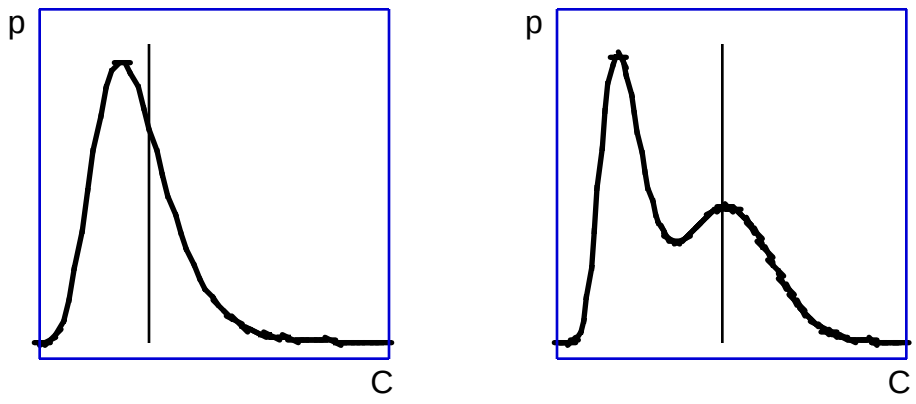


Figure 4.7 Discrete Probabilities from Continuous Distributions

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

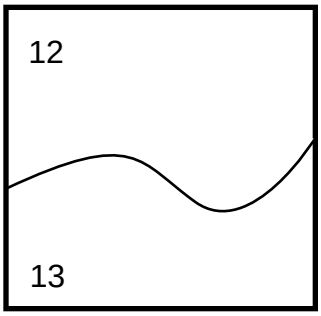
In order to study the predictive power of more than one parameter on the behavior of groundwater quality, it is necessary to combine data from several thematic layers. This process, called overlaying, is similar to constructing the intersection of sets. Overlaying two polygon coverages—for example soil 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

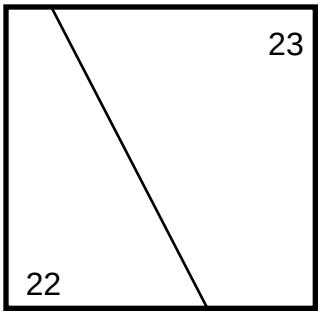
Like a polygon, a raster or grid cell is a contiguous bounded area with associated data. Unlike a polygon, its boundaries are determined by a regular pattern, like a checkerboard, not by changes in the data values associated with it. Rasters are frequently used to express continuously varying quantities.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

Soil



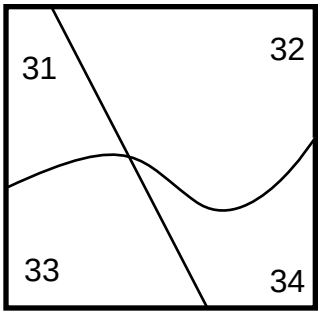
Soil-id	Thick
12	45
13	12

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_ns.

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.

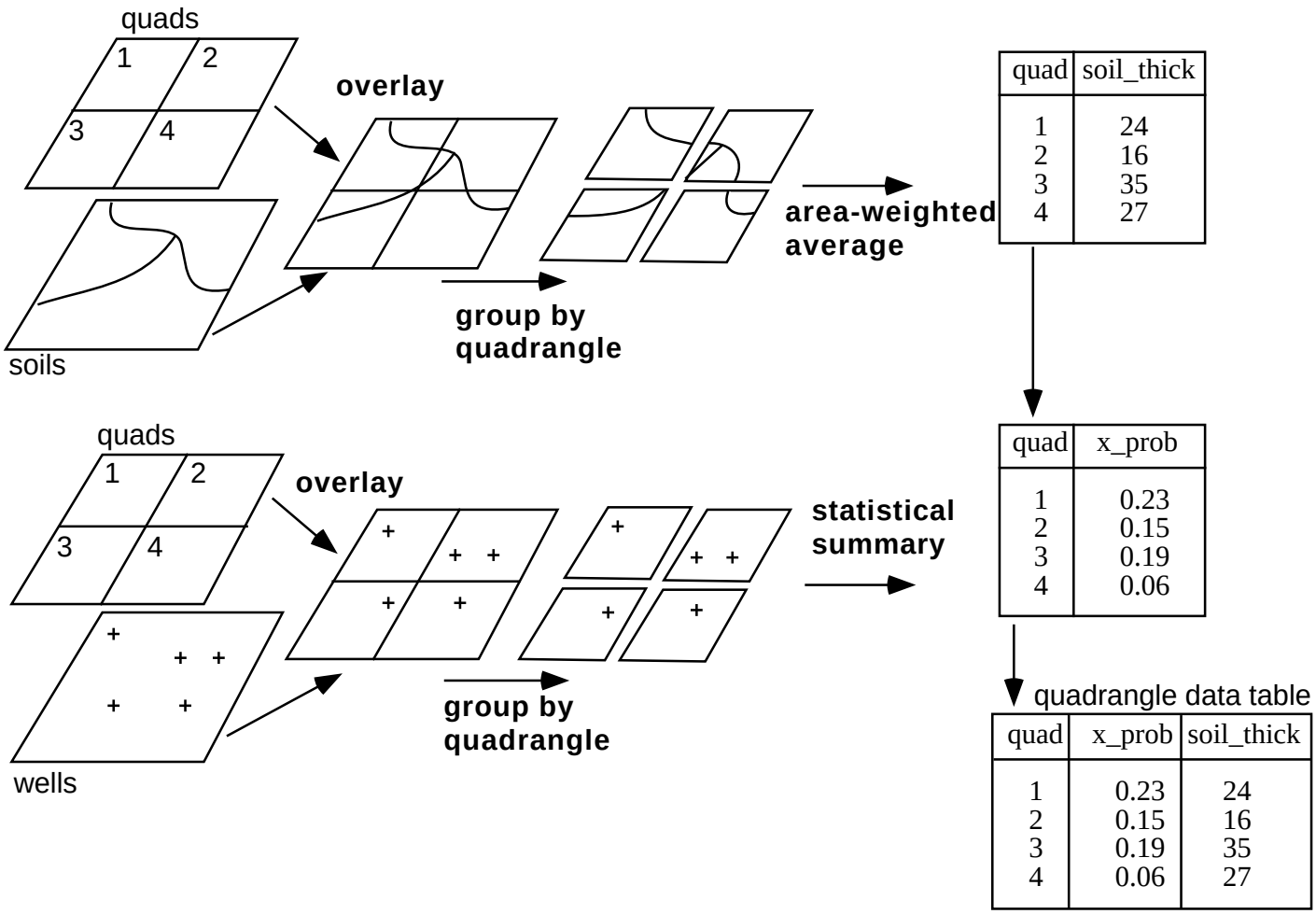


Figure 4.9 Resolving Indicators and Exceedence Probabilities to a Common Grid

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.

Chapter 5: Procedures

This chapter describes design and implementation of the computer programs used to organize and interpret the data. Definitions of the INFO database tables used in this study are listed in Appendix A. The source code for programs is presented in Appendix B. In several cases, an involved series of Arc/Info commands was entered from the keyboard to carry out a procedure. Some of the Arc Macro Language (aml) programs included in Appendix B are, in effect, transcripts of keyboard procedures with comments inserted for clarity. A reader reasonably familiar with Arc/Info should be able to reconstruct the computer analyses carried out in this study completely from the material in this chapter and the appendices. All operations were carried out on a Sun Sparc 2 workstation using Arc/Info version 6.1.1, except as noted.

5.1 Important Features of INFO and TABLES

The Arc/Info GIS incorporates the INFO database management system for management of its tabular data. Since the organization and manipulation of tabular data, both related and unrelated to spatial objects, is crucial to the methods used in this study, a brief discussion of some important features and programming tricks is required before moving on the specific procedures used. A limited set of INFO commands is included in the TABLES subsystem of Arc/Info. The database procedures in this study were carried out using TABLES commands.

In concept, an INFO table is a set of *records*, each with the same set of *items*. A record represents some object—a well, a nitrate measurement, a STATSGO map unit—and the items define a set of properties of the object. This structure is identical to the set of tuples that make up a relation in the database model discussed in Section 4.3.1. For example, a record in the table of nitrate measurements, called "meas" contains the items WELL-ID, YEAR, MONTH, DAY, and NITRATE, among others. A record in the table of well data, called "wells.dat," contains the items WELL-ID, DEPTH, LATITUDE, and LONGITUDE, among others. (This example is simplified for clarity. The definitions of the tables actually used in the study are listed in Appendix A.) Each record (or tuple) in "meas" corresponds to a nitrate measurement collected from one of those wells on a particular day, and each record in "well" corresponds to a well somewhere in Texas.

The database query expressed in relational calculus in eq. 4-1,

$$\{r \mid r \in meas \wedge r[\text{well-ID}] = 5740304\}$$

would be carried out in INFO with the commands

```
select meas
reselect well-id = 5740304
```

The first command makes "meas" the active, or queriable, table, and the second restricts the selected set of records to those meeting the stated criteria. By doing so, these commands implement the two parts of the predicate of the relational calculus query.

Because the tables "meas" and "well" contain a common item, WELL-ID, there is a logical connection between them. This connection was exploited in query 4-2, which selected tuples on the from the relation "meas" on the basis of values in the relation "well". Ordinarily, however, INFO can access the contents of only a single table (the "selected" table). To circumvent this limitation, the user must use a special mechanism called "relate," which allows a table to be expanded temporarily with items from a second table. Figure 5.1 shows a relate in concept. The tables shown in Figure 5.1 are made-up examples, not data from the study.

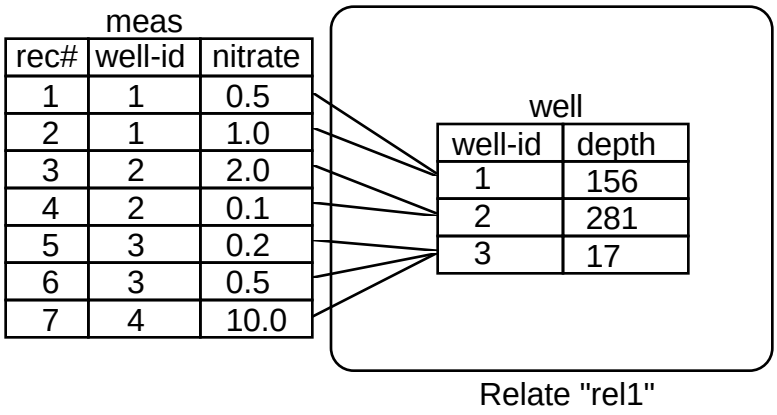


Figure 5.1 Example of a RELATE

In the figure, the relate "rel1" has been defined on well. The relate grants access to the contents of wells.dat to any other table that contains an item identical to the item "wellno". In the figure, the relate has been attached to nit.dat. This means that the contents of wells.dat can be read while meas is selected. In TABLES, the items accessed through a relate are referred to by the

relate name and the item name connected by two slash (/) characters. "rel1//depth" refers to item "depth" in wells.dat, accessed through the relate "rel1". The equivaent of the relational caluculus expression

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

would be the INFO commands

```
select meas
reselect rel1//depth < 100
```

which would select records 5 and 6 from the tables shown in Figure 5.1.

The relate mechanism gives rise to a useful programming device used several times in this study. Relates can grant access for both reading and writing to the related table, and INFO permits simple mathematical operations to be performed on the contents of tables. These operations can be combined to calculate counts, sums, averages and weighted averages. A series of examples will illustrate this device.

The TABLES command `calculate` calculates new values for items. To create an item for nitrate-NO_3 equivalents of the nitrate-N concentrations shown in Figure 5.1, a new item called conc-NO3 is added to the table "meas" and the commands

```
select meas
calculate conc-NO3 = conc * 4.429
```

produce the values shown in the table "meas" in Figure 5.2.

To count the number of measurements in each well, a new item called m_cnt is added to the table "well" and given an initial value of zero for all records. The commands

```
select meas
calculate rel1//m_cnt = rel1//m_cnt + 1
```

produce the values shown in Figure 5.2 in the table "well."

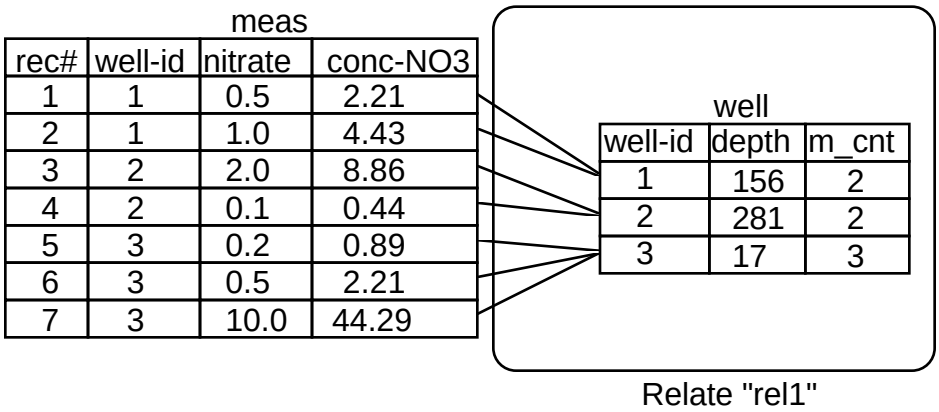


Figure 5.2 Using a RELATE to Count Measurements in Wells

This device works by exploiting the fact that TABLES performs the calculations sequentially, record by record. The item in the related record is updated once for each corresponding record in the selected table. If INFO and TABLES performed calculations in parallel, this device would not work. The same operations could be performed in external programs, or through the use of the full INFO database programming language, but this device simplifies the required programming considerably.

If the "reselect" command is used to restrict the set of selected records to those meeting a set of conditions, the relate/calculate device can be used to count records meeting the condition. For example if an item called "gt1_cnt" is added to the well.dat table (with initial values set to zero), the following TABLES commands, executed while the nit.dat table is selected, will assign the number of measurements in each well that exceed 1.0 to the new item.

```
reselect conc > 1.0
calculate well//gt1_cnt = well//gt1_cnt + 1
```

More complex uses of the combination of relates with the calculate command will appear in the sections that follow.

Another database operation that requires some explanation is the *redefine* feature, which is best defined by example. Within records of the tables "twdb_wells.dat" and "twdb_wells.nit," which contain the well descriptions and nitrate measurements extracted from the TWDB Groundwater Data System, the well number occupies the first seven bytes. The first two digits of the number, as explained in Chapter 3, identify the one-degree quadrangle in which the well is located. The first two bytes of each well record are *redefined* as "QUAD_1D", allowing the user to refer to only the part of the well number that identifies the one-degree quadrangle, without requiring parts of the well number to be entered multiple times. Similarly, the 7.5' and 2.5' quadrangles are identified from the first four and five bytes of the well number, as shown in Figure 5.3. By using these redefined items, and the relate/calculate trick described above, it is possible to count the wells in each quadrangle, the number of measurements in

each quadrangle, and the number of measurements exceeding various thresholds in each quadrangle.

5740502	well number
57	one-degree quad number
5740	7.5-minute quad number
57405	2.5-minute quad number

Figure 5.3 One INFO Item Redefined into Four Items

The redefine operation can combine any set of adjacent single-byte "columns" in an INFO table, including those that span multiple items, into a psuedo-item. This is especially helpful when no single item has a unique value for each record in the table. Records in the component table of the STATSGO database, for example, can only be identified by the concatenation of the map unit ID and the component number. This is discussed in detail in Section 5.3. Redefined items appear at the end of lists of items in INFO tables. (See table definitions in Appendix A for examples.)

5.2 DATA ENTRY

This section describes the collection and mobilization of data from its original sources into Arc/Info GIS coverages and data tables. The item definitions for the resulting INFO tables are given in Appendix A.

5.2.1 TWDB Well and Nitrate Data

The first step in this study was to convert the well and nitrate measurement data from the form in which it was provided by TWDB to Arc/Info

database form. This required definition of two INFO tables, one for well data and the second for nitrate measurement data. TWDB provided the data in comma-delimited text files, with character data enclosed in single quotation marks. This format permitted the data to be entered directly into the tables using the ADD command in Arc/Info TABLES.

Selecting the Base Data. The well and nitrate measurement data from TWDB were entered into two tables, called "twdb_wells.dat" and "twdb_wells.nit", respectively. Items were added to these tables to indicate whether the records for each well or nitrate measurement should be included in the study. After the records were marked for inclusion or exclusion, according to the criteria described in Chapter 3 (see Table 3.3 and Section 3.2.4), "twdb_wells.dat" and "twdb_wells.nit" were copied to new tables, called "include.wells" and "include.nit" and the excluded records were purged from these tables leaving only the wells and nitrate measurements to be used in the study in the "include" tables.

A Programming Example. Because the process of removing excluded data from the tables illustrates some of the basic techniques used in passing information from INFO tables to external (C or FORTRAN) programs and from one INFO table to another, it will be described here in detail.

The program "testquad.aml" in Appendix B identifies wells whose latitude and longitude are not consistent with their well numbers. It also illustrates the movement of data from an INFO table to an external program and back. The steps below outline the program's procedure, and are typical of the

approach taken to an analysis that is more complex than can be carried out easily in Arc/Info.

1. From the "twdb_wells.dat" table, write the well number, latitude, and longitude of each well to a comma-delimited text file called "qtest.in".
2. To perform the actual test, run the test_quad program (see test_quad.c in Appendix B) with input from "qtest.in" and output directed to "qtest.out". The output file contains lines consisting of three comma-delimited fields: the well ID number, a one-character code ('y' or 'n') indicating that the well is or is not correctly located in its quadrangle, and a one-character code indicating whether mis-located wells are in the wrong 1_, 7.5', or 2.5' quads (values are 'd', '7', or '2').
3. Define a new INFO table ("qtest.tab") to hold the test results temporarily. The items in the table are WELLNO, QUAD_OK, and QUAD_ERR, corresponding to the fields in the test program output file.
4. Select the new table, and add records to it from the text file "qtest.out".
5. Join the tables "twdb_wells.dat" and "qtest.tab", adding the new items QUAD_OK and QUAD_ERR (with values determined by the external program) to the original table.
6. Delete the input and output text files, and the temporary INFO table.

The result of this procedure is the extension of the twdb_wells.dat table. The two new items in the table show whether a well is correctly located in the quadrangle indicated by its well number, and, if it is incorrectly located, whether

its mis-location is due to placement in the wrong quadrangle at the one-degree, 7.5-minute, or 2.5-minute level.

Although the comparison of the well number with the latitude and longitude is too complex to be carried out with the simple comparison operators provided by Arc/Info, the ability to write selected contents of INFO tables to text files, and to permanently add new items to existing INFO tables using a key item (one that has a unique value for each record) makes it possible to carry out analyses in external programs, using the INFO database functions to assure that the results are attached to the right records in the original database. The calculation of lognormal parameters for nitrate detections follows a similar, although slightly more sophisticated, procedure.

The procedure used to test the consistency of well numbers and well locations required that data from an INFO table be transferred an external program, and that the output from that program be added to the original INFO table. Testing that records in the nitrate database have corresponding records in the wells database requires reading and writing to two tables simultaneously, but does not require the tables to be joined permanently. This is accomplished using the "relate" mechanism described in Section 5.1. The aml programs "testquad.aml" and "include.aml" (listed in Appendix B) contain a complete procedure for selecting well and nitrate measurement records for exclusion from the study, based on the criteria listed in Chapter 3. Execution of these programs produces the tables "include.wells" and "include.nit", containing the primary data for the study.

Converting Nitrate-NO_3 to Nitrate-N. The final step in preparing the nitrate measurement data for use in the study was the calculation of "adjusted" nitrate concentrations. This process converted measurements from mg/l as nitrate to mg/l as nitrogen and set a uniform reporting limit for all nitrate measurements. A new item called "nit_adj" was added to the "include.nit" table. For all records included in the study reporting a nitrate concentration of 0.45 mg/l as nitrate, the value of this item was set to 0.1, to indicate a measurement at or below the reporting limit of 0.1 mg/l as nitrogen. For all other included records, the value was set to the reported nitrate concentration divided by the conversion factor 4.43.

Data Subsets for Aquifers. Two additional tables, one for wells and one for nitrate readings, were created for the five aquifers selected for closer study. These were created by copying "include.wells" and "include.nit" as "aq5.wells" and "aq5.nit". A four-character text item called "aqf" was added to "aq5.wells" to hold a code for the names of the study aquifers. Wells records with TWDB aquifer codes identified with the study aquifers were selected and appropriate values for the "aqf" item were written. The following TABLES commands illustrate this process.

```
select aq5.wells
reselect aqfcode = '124CRRZ' or aqfcode = '124WLCX'
     or aqfcode = '124CZWX' or aqfcode = '124CZWXA'
move 'CZWX' to aqf
aselect
reselect aqfcode = '218EBFZA'
move 'EBFZ' to aqf
...
```

(repeat for Hueco-Mesilla Bolson, Ogallala, and Seymour Aquifers)

```
...
```

```
aselect
reselect aqf = ''
purge
```

The resulting table contains only records for wells associated with the study aquifers, each with a simple code identifying the aquifer. These are listed in Table 5.1.

Table 5.1 Codes for Study Aquifer Identification

Aquifer Name	Code
Carrizo-Wilcox	CZWX
Edwards (Balcones Fault Zone)	EBFZ
Hueco-Mesilla Bolson	HMBL
Ogallala	OGLL
Seymour	SYMR

The "aq5.nit" table was related to the "aq5.wells" table by the shared "wellno" item. Nitrate measurement records with no corresponding well record in "aq5.wells" were identified and purged. Item definitions for "aqf.nit" are identical to those for "include.nit". Item definitions for "aqf.wells" are given in Appendix A.

5.2.2 Soil Data

The STATSGO database was received as an Arc/Info coverage with related INFO tables. It was used without alteration. Calculation of average soil thicknesses and organic material contents for STATSGO map units and 7.5' quadrangles is described in Section 5.3.

5.2.3 Precipitation Data

The precipitation maps used in this study are derived from data provided by Hydrosphere, Inc. on CD-ROMs under the name *Climatedata*. The CD-ROMs

contain Arc/Info point coverages, which locate weather reporting stations in the US. and contain summary statistics for those stations. More detailed data—daily, monthly, and annual figures for the period of record of each station—are included in tables, which must be read with Hydrosphere's proprietary software.

Because the summary data used for this study was not the same as that included with the Arc/Info coverages, the following procedure was followed to produce Thiessen polygon maps of average reported rainfall at stations in Texas and a 100-km buffer around Texas during the years from 1951 to 1980.

1. The Arc/Info coverages containing weather stations in Texas and adjoining states (New Mexico, Oklahoma, Kansas, Arkansas, and Louisiana) were joined using the Arc command "mapjoin". The resulting coverage was trimmed to a 100-km zone around Texas by applying the Arc command "clip" to the multi-state point coverage, using for the clip coverage a map created with the Arc command "buffer" applied to the outline of the STATSGO map of Texas. Finally, the Arc command "reselect" was applied to limit the coverage to precipitation stations only. This coverage was named "prec_tx".
2. The station ID number, station name, reporting year, and total annual precipitation for stations in the six-state area for each year from 1951-80 were written to a comma-delimited text file, using Hydrosphere's software. This data was entered into an INFO table called "prec.dat" (see definition in Appendix A).

3. An INFO table called "station.mean" (see definition in Appendix A), with one record for each of the stations in the coverage "prec_tx", was created to hold the summary data for the precipitation stations. The first year, last year, and maximum gap in reporting was calculated following the procedure listed as "year.aml" in Appendix B. This procedure, like the one that tests the consistency of well locations, relies on external programs to perform some analysis. Here, the external programs are written in AWK (Aho, et al. 1988), rather than C. AWK is particularly suited to once-through text file operations like this, file opening statements, variable declarations, and other overhead of C or FORTRAN are unnecessary in simple AWK programs.
4. The procedure listed as "precmean.aml" in Appendix B was followed to produce average annual precipitation figures in "station.mean." This is an example of the use of the relate/calculate method to calculate an average.
5. The "station.mean" table was joined to the polygon attribute table of "prec_tx", associating the annual averages with the station locations. Stations with gaps greater than two years in their reporting histories were dropped from the coverage.
6. A Thiessen polygon network was created from the reduced point coverage with the Arc/Info command "thiessen". The portions of this polygon coverage outside of Texas were removed using the "clip" command with

the outline of Texas from STATSGO, resulting in the polygon coverage shown in Figure 3.15.

5.2.4 Fertilizer Sales Data

Like the STATSGO soil data, the nitrate fertilizer sales data were received already in the form of Arc/Info GIS coverages. The data, as provided by the USGS, came in the form of 6 coverages, each a map of the counties of Texas with attribute data attached listing estimated fertilizer sales for a single year for the period 1986–91. This data was reorganized for use in this study.

A new INFO table called "nitrate.use" (see Appendix A for definition) was created with one record per county, and items for each year's estimated nitrogen fertilizer sales and "use" (sales in tons divided by area of county). The total estimated fertilizer sales and "use" for each county was calculated by summing and averaging the annual figures, and listed in additional items.

After the fertilizer data table was created, all but one of the original coverages were deleted. The remaining coverage was used as a county base map. Where fertilizer data was used in the study, it was attached to this base map through the use of a relate.

5.2.5 Water Utilities Division Data

Nitrate monitoring data collected by the Water Utilities Division (WUD) of the Texas Natural Resource Conservation Commission were received in the form of an Arc/Info Coverage, containing well locations and descriptions, and two data tables, one, called "poe", containing system and point of entry identifications for each water source known to WUD, the other containing, called

"nitrate", containing records of nitrate measurements collected at points of entry to the water systems. The tables were received in dBase format, translated from the Paradox database maintained by the WUD.

The dBase files were read as PC Arc/Info data files and translated into equivalent INFO tables using EXPORT in PC Arc/Info to create transferable text files, transferring the text files to the workstation with ftp, and using IMPORT in Arc/Info on the workstation to create INFO tables from the text files.

The nitrate, nitrite, and combined nitrate and nitrite concentrations reported in the "nitrate" table were recorded in text fields so that the character "<" could be used to indicate measurements below the detection limit. This makes numerical analysis of the data difficult, so two additional items were appended to the table for each concentration item: one single character field to hold the "<" characters, and a numerical field to hold the concentration value. The resulting table has item definitions listed in Appendix A for the table "nit.wrk". The item "no3fl" was set equal to "<" for all records with a "<" character in the nitrate results column. The numerical values for nitrate concentrations were added to the table by writing the record number and nitrate values to a text file, removing non-numeric characters from that file with a text editor, writing the remaining numeric values to a temporary INFO table with the ADD FROM command and joining the temporary table to "nit.wrk".

The WUD data tables were further altered in the process of comparing the nitrate measurements they contain with predictions made from the TWDB data. These alterations are described in Section 5.8.

5.2.6 Midwest Herbicide and Nitrate Data

Because of the lack of comprehensive herbicide measurement data in Texas, a comparison of nitrate and herbicide detections was made using data from Kolpin, Burkart, and Thurman (1993). The comparison is simple enough that it is fully described in Section 6.5, where the results are discussed. The preparation of the data was somewhat more complex.

The data, describing well locations, geologic settings, and construction, and the results of water quality analysis, were available only as a published report, so the values were read into a computer text file with a scanner and a character-recognition program on an Apple Macintosh microcomputer. The contents of the text file were transferred to an Excel spreadsheet where they were parsed into columns. The values in the spreadsheet were compared with the tables in the report and corrected as necessary.

Two INFO tables, one for well data and one for water-quality data, were defined (see "construction" and "quality" in Appendix A). Separate items were defined for flags, such as the "<" character to indicate concentrations below detection limits, and the numerical concentration values. The spreadsheet values were exported as comma-delimited text and transferred to the workstation, where they were loaded into the data tables using the ADD FROM command in Arc/Info's TABLES module.

Because nitrate was reported only as the sum of nitrate and nitrite, a new item for nitrate values was added and values were calculated by subtracting the nitrite concentration from the nitrite/nitrate total. Where nitrite was below the

detection level, the nitrate value was set equal to the nitrite/nitrate sum. All nitrite and nitrate values in this data set are reported in equivalent nitrogen units.

5.2.7 A Note on Map Projections

All maps used in the study were in an Albers equal-area projection with the following parameters:

Units:	Meters
Datum:	1927 North American Datum (NAD27)
1st Standard Parallel:	29_ 30' 00"
2nd Standard Parallel:	45_ 30' 00"
Central Meridian:	-96_ 00' 00"
Latitude of Origin:	23_ 00' 00"

The US Geological Survey uses this projection for its National Atlas of the United States, which many agencies use to provide base maps for a variety of thematic maps. In fact, all the map-based data used in this study (STATSGO, nitrate data, and precipitation station locations) was originally delivered in this projection, so that no re-projection of maps or GIS coverages was required for any of these data. Locations of wells and the boundaries of 7.5' quadrangles were given in unprojected latitude and longitude, and the quadrangle maps were generated in this form, then transformed into the Albers projection.

5.3 CALCULATION OF DATA DERIVED FROM STATSGO

The STATSGO database, as provided by the Soil Conservation Service, does not provide values for average soil layer thickness and average organic material content for the map units. As described in Section 3.3, these values were calculated through a process that, in effect, integrates soil parameter values through the layers of the soil components and then averages those integrated

parameters, weighted by component area, over the map units. Here, the steps required to carry out the integration and averaging process in Arc/Info will be described. Within a STATSGO soil component, the process for calculating soil thickness is simple, the process for calculating organic content is more complex. Once the parameter values are calculated for the components, the averaging process over the map units for both parameters is identical. All three procedures are described here.

For the calculation of derived data, the STATSGO map unit, component, and layer data tables were copied to new tables called "study.mapu", "study.comp", and "study.layer" to avoid corruption of the original files. Most of the items not required for soil unit identification or for calculation of the parameters of interest to this study were dropped from the new tables. The definitions of the resulting tables are listed in Appendix A.

Defining Keys for STATSGO Tables. Calculating the map unit averages for parameters listed in the component and layer tables begins with the definition of a unique identifier (a key item) for each component. Together, the map unit ID and the sequence number for a component make up such a unique identifier. Since the two are listed in adjacent fields in the data tables, they can be combined through a redefine operation, similar to that which extracted 1-degree, 7.5-minute, and 2.5-minute quadrangles from the well ID numbers in the TWDB well tables. The set of adjacent bytes that make up the map unit ID and the sequence number were redefined as an item called "mapseq" in the "study.comp" and "study.layer" tables, providing a key for relating the two tables.

Calculating Soil Layer Thickness. STATSGO lists many soil parameters, including soil depth and organic matter content, as ranges, defined by a low and high value. Item names for low values end in "L", and names for high values end in "H". For example, the items "LAYERDEPH" and "LAYERDEPL" contain high and low values for depth of a soil layer, i.e. the depth to the top and bottom of the layer. Because the thickness of the soil component is simply the maximum of the high values of the depth of the layers that make up the component, the Arc command "statistics" can calculate the component thickness in a few steps, as follows.

1. Invoke the STATISTICS command to calculate summary statistics on the layer table for each unique value of the "mapseq" item, and write the results to a new INFO file called "maxdep.dat". The syntax for this command is

   ```
   statistics study.layer mapseq maxdep.dat
   ```

   Arc then asks for the specific statistics to be calculated. The component thickness is equal to the maximum of the item "LAYERDEPH".
2. Join the table resulting from the statistics operation to the component table using the JOINITEM command with "mapseq" as the key item.
3. In TABLES, change the name of the new item in the component table from "max-layerdeph" to "soilthk" using the ALTER command.

Calculating Soil Organic Content. The process of calculating an organic matter content for each component is more complicated, because the organic matter is given as a percentage by weight, so that it must be multiplied by the

bulk density (gm/cm^3) to produce a meaningful number when integrated over the depth of the soil layer. Since both the organic matter content and the bulk density of the soil are expressed as ranges, this requires the calculation of several different values. The procedure is as follows.

1. Add items "OMM" and "BDM" to the layer table to hold values for the mid-range of organic matter and bulk density, respectively. Add items "MIN-ORG", "MID-ORG", and "MAX-ORG" to the component table to hold minimum, mid-range, and maximum values for organic content (kg/m^2) in the component.
2. In the layer table, calculate the mid-range values of organic matter fraction and bulk density as one-half the sum of the minimum and maximum values.
3. Using a relate based on the redefined "mapseq" item, from the layer table, add the product of the mid-range values of the organic matter fraction, the bulk density, and the thickness of each soil layer to the "MID-ORG" item of the corresponding component in the component table. Because each layer in the component will contribute to the sum in the component record, this has the effect of summing the products (or numerically integrating) over the layers of the component.
4. Repeat step 3 with the minimum and maximum values to establish the range of values.

The aml program "org_int.aml" (See Appendix B) was used to carry out steps 2–4.

The two procedures described above produce the soil thickness and organic content for individual soil components from the layers that make them up. A third procedure, which follows, calculates the average over a map unit of a parameter evaluated in all the components that make up that map unit.

1. Add an item to the map unit table to hold the average value for the parameter.
2. Using a relate defined on the map unit ID, from the component table add the product of the parameter value and the fraction of the map unit formed by the component to the new item in the map unit table. Because each component in the map unit will contribute to the sum in the map unit record according to the area it contributes to the map unit, this produces an area-weighted average of the component values in the map unit table.

The aml program "unit_avg.aml" (See Appendix B) carries out these steps, and also sums the component fractions of the map units as an error test. If the area percentages of the components of any map unit fail to sum to 100, the program notifies the user.

5.4 PREPARATION OF QUADRANGLE MAPS

The Texas Water Development Board does not provide a GIS coverage of the quadrangles that provide the basis for their well-numbering system. Since this study uses these quadrangles as grouping units for statistical analysis of water quality data, quadrangle maps were required both for display of statistics and for the calculation of quadrangle averages of the potential indicator variables. The aml program "build_quads7.aml" (see Appendix B) constructs an Arc/Info polygon coverage called "TWDB_7M" in geographic (unprojected

latitude and longitude) coordinates. It calls the external program "tx_7m.c" (see Appendix B) to create the coordinate file used by the Arc GENERATE command to create the coverage. The aml program uses the Arc commands GENERATE and CLEAN to build the polygon coverage, and little explanation beyond the function of these commands described by Arc/Info documentation should be required.

One subtlety, however, should be illuminated. The programs assign four-digit integers to the quads as polygon ID numbers. In the polygon attribute table (PAT), the ID is automatically assigned the name "twdb_7m-id" and the item type "B", or binary-coded integer. A new item called "quad_7.5m" of type "I", or one-byte-per-digit integer, was added to the PAT. The values in this item and the ID number are identical, but the formats are different. This new item in "I" format provides a key for relates used to link the PAT to the TWDB well and nitrate measurement tables, where the quadrangle numbers are also stored as type "I".

The quadrangle map was projected from geographic (decimal degrees) coordinates to the Albers projection used for the study. The resulting coverage was named "quads_7.5".

To identify quads with the five study aquifers, a new table called "aq_quad.dat" was created by extracting the quad numbers with the PULLITEMS command. A single-digit integer item was added to this new table for each of the five study aquifers, and one more item was added to count the number of study aquifers associated with the quad (item definitions are given in

Appendix A). The following commands set the CZWX item in "aq_quad.dat" to a value of 1 for each quad containing a well associated with the Carrizo-Wilcox. A relate called "quad" links "aq5.wells" to "aq_quad.dat" by the item "quad_7.5m".

```
select aq5.wells
reselect aqf = 'CZWX'
calc quad//czwx = 1
```

A similar set of commands set the flags for the remaining four aquifers. The sum of the aquifer flags in each record was calculated and assigned to the item "aq_cnt". Records with "aq_cnt" equal to zero were purged from the table, leaving only records of quads associated with one or more of the study aquifers. This table was used to produce the aquifer quad map in Figure 6.15.

5.5 CALCULATION OF STATISTICS

Two types of statistics were calculated for the nitrate measurement data in the TWDB database:

1. Estimates of the probabilities that a single threshold concentration level will be exceeded (discrete probabilities).
2. Estimates of the parameters of an assumed concentration probability distribution (log-normal parameter fitting).

The discrete probability and log-normal parameter estimates are calculated for groups of measurements formed by 7.5' quadrangle, and by aquifer. The mathematical meanings of these procedures have been discussed in Chapter 4; here, the details of carrying out the operations with Arc/Info and external programs will be described.

5.5.1 Discrete Probability Estimates

The various programs used to calculate estimates of the discrete probabilities all operate using essentially the same procedure. Counting wells, measurements, and measurements exceeding thresholds is carried out in TABLES by the following steps.

1. Create a table to hold the results of the calculations, with one record for each group (county, quad, or aquifer) to be considered. To support the relates that make the calculations possible, one item, used to identify the group, must be identical to an item in the well and nitrate measurement tables. An example, the table for 7.5' quadrangles, "counts.quad", is listed in Appendix A. (If necessary, an item identifying the group can be added to the measurement table by copying data from the well table. This was done for counties and aquifers, which were not included in the original definition of the measurement table.)
2. Create a relate to link the results table to the well and measurement tables.
3. Select the well table, and using the relate, add one to the item "WELL_CNT" in the results table for each record in the well table. This produces a count of the wells in each group.
4. Select the measurements table, and using the relate, add one to the item "MEAS_CNT" in the results table for each record in the well table. This produces a count of the measurements taken in each group.
5. Restrict the selection to measurements with nitrate values exceeding the detection limit, and add one to the item "DTCT_CNT" in the results table

for every record in the reduced selection. This produces a count of the measurements in each group that exceed the detection limit.

6. Repeat step 5 for the detection limits of interest to the study. Here, levels of 1 mg/l, 5 mg/l, and 10 mg/l nitrate as nitrogen were used.
7. Select the results table and calculate estimates of exceedence probabilities by dividing the number of detections above the threshold limits by the total number of measurements.

Steps 2–7 are carried out by the aml programs "count_quad.aml", "count_aq.aml", "count_aqquad.aml", and "count_county.aml" for 7.5' quadrangles, the five study aquifers, quadrangles with measurements from the study aquifers, and counties.

The above procedure calculates the best estimate of the exceedence probability for the various threshold levels, but does not provide confidence intervals on the estimates. Confidence limits are calculated using an external FORTRAN program (bino2.f in Appendix B). The program uses a cumulative binomial distribution estimation function found in a statistical function library called SCDFLIB (Brown and Lovato, 1994). The following procedure adds confidence limits to the probability estimates calculated above.

1. Create a series of temporary INFO tables to hold the results of the calculations. The items in the tables include the group (county, quad, or aquifer) identifier, upper and lower bounds for the confidence interval on the exceedence probability for the chosen threshold, and the difference between the upper and lower bounds.

2. Select the INFO table containing the measurement and threshold exceedence counts for the groups, and write a text file containing the identifier, number of measurements and number of threshold exceedences for each group. This file becomes input to the confidence limit program.
3. Call the system program (bino2) to calculate the upper and lower confidence limits. This produces a text file containing the identifier and the lower and upper confidence limits for each group.
4. Read the contents of the text output from the confidence limit program into the temporary INFO tables.
5. Add the confidence limits to the table containing the exceedence probability estimates with the JOINITEM command. The definition of the resulting table, called "bino.quad" is given in Appendix A.

The aml program "bino_quad.aml" in Appendix B carries out this procedure for all the 7.5' quadrangles with well data for the 0.1, 1, 5, and 10 mg/l thresholds .

5.5.2 Lognormal Probability Estimates

In addition to the discrete probability estimates, the best-fitting parameters for a lognormal distribution were calculated from the nitrate measurements in each quadrangle. The following procedure used to accomplish this task.

1. Sort the table "include.nit" by quadrangle number, and write the quad number and adjusted nitrate reading for each nitrate measurement to a text file called "fit.in" using the TABLES "UNLOAD" command. The text file is

used as input to a C program that carries out the next steps. Sorting the nitrate data by quad number assures that the data in the input file is grouped by quad.

2. The system program "logfit" (see "logfit.c" in Appendix B) reads the two fields from each line of the input file, appending each nitrate value (the second field) to an array. The array continues to grow as long as the same quad number is read from the first field of the line being read.
3. When a new quad number is found, the array is sorted, counted, and numbered. For each value of nitrate concentration, the log of the concentration, the value of Blom's plotting position for the highest-ranked entry with that value, and the normal variate (Z) corresponding to that plotting position are calculated.
4. When all the concentration values associated with a quad have been converted into log values and normal variates, a linear regression (using a function from Press et. al (1988)) fits the following equation to the data:

$$Z = a + b(\log(C)).$$

5. The mean of the log concentrations is calculated as -a/b, and the standard deviation of the log concentrations is calculated as 1/b.
6. Probabilities of exceeding 0.1, 1, 5, and 10 mg/l are estimated from the lognormal parameters.
7. The lognormal parameters, descriptive statistics for the regression (F, t, r^2, standard error, and significance of F), and exceedence probability estimates are written to an output file.

8. Steps 2–7 are repeated until the input file is exhausted.
9. The contents of the output file are entered into an INFO table, which can then be related to quadrangle maps, or the discrete probability estimates for plotting or comparison.

The definition of the INFO table "logfit.quad" is presented in AppendixA. The programs "logfit.aml" and "logfit.c" carry out the steps listed above, and are listed in Appendix B.

5.6 MAP PREPARATION

Each map in this document was prepared in the Arcplot module of Arc/Info. The map composition was set by a series of Arcplot commands in an aml program, the output of which could be directed either to the computer screen or to an Adobe Illustrator file. The program was rewritten and executed several times with the output directed to the screen. When a satisfactory map composition was set, the program was run a final time with output directed to an Illustator file. The aml program "gt1_plot.aml" in Appendix B, which was used to prepare Figure 6.4, is an example.

The map compositions created in Arcplot had no legends or captions. These were added by transferring the map file to an Apple Macintosh computer and adding labels and a legend with the Adobe Illustrator program. The maps themselves were not edited in this process, so that the information they contained would not be altered. Colors in the maps were set using the CMY (cyan magenta yellow) color scheme. Because both Arcplot and Illustrator permit colors to be set by numerical values on the CMY scales, it was possible to create legends in

Illustrator that exactly matched the map colors created by Arcplot. The CMY components of the shades used in the figures in Chapter 6 are listed in Table 5.2.

Table 5.2 CMY Components of Probability Map Colors

Probability	Cyan	Magenta	Yellow
> 80%	0	100	100
60–80%	0	35.3	100
40–60%	0	0	100
20–40%	39.6	19.6	80.4
<20%	100	0	100

All county maps and all quadrangle maps were created from one county and one quadrangle coverage. For example, all of the maps of quadrangles, regardless of the theme or shading scheme, were generated from the coverage quads_7.5, which was created by the process described in Section 5.4. The shading and coloring of the maps in Chapter 6 to show quad exceedence probabilities at various thresholds was done by relating the quad coverage to the counts.quad, logfit.quad, and bino.quad tables. This assured consistency between the maps and reduced the storage demands on the computer where the study data were stored. The use of related data to set a shading scheme is also illustrated in "gt1_plot.aml".

5.7 Indicator Values and Stepwise Linear Regression

The variation of the nitrate exceedence probabilities was compared to variations in the proposed indicator parameters in the 7.5' quadrangles using stepwise multiple linear regression. Average parameter values had to be calculated for each quadrangle, so that these values could be compared with the exceedence probability estimates.

Area-weighted averages of soil parameters was calculated for each quadrangle by the following procedure.

1. The STATSGO map unit coverage and the 7.5' quadrangle coverage were combined using the Arc/Info "intersect" command. This produced a polygon coverage similar to the simple example in Figure 5.4. Quadrangles are subdivided into smaller polygons by the boundaries of the STATSGO map units. (The soil polygons in the example are much simpler than actual STATSGO map units so that the proportion of the quads in each soil group can be estimated easily.)

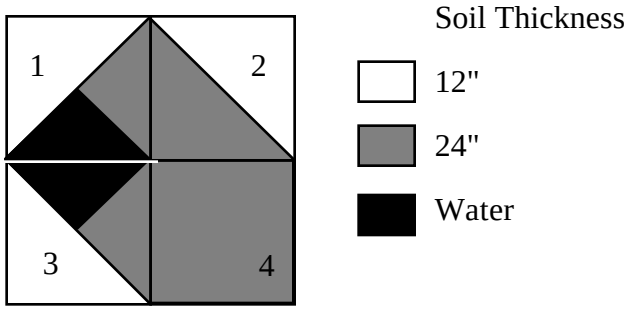


Figure 5.4 Simplified Quadrangle/Map Unit Intersection

2. The soil area in each quadrangle (i.e., the area not covered by water) was summed into a table called "params.quad". If the quadrangles in Figure5.4 have unit area, the soil areas of quads 1 and 3 are 0.75 and the soil areas of quads 2 and 4 are 1.0.
3. For each polygon, the product of the area and the parameter value were added to an item in the record of "params.quad" corresponding to the quadrangle from which the polygon was divided. For soil thickness, this

item is called "thkar." The value of "thkar" in quads 1 and 3 is 12; in quad 2 it is 18; in quad 4 it is 24.

4. The area-weighted parameter average is calculated by dividing the quadrangle sum of the area-parameter values by the soil area of the quadrangle. The value of "av-thk" for quads 1 and 3 is 16; for quad 2 it is 18, for quad 4 it is 24.

Steps 2–4 of the above procedure are carried out in the aml program "aw_avg.aml" in Appendix B.

A similar procedure was followed to calculate area-weighted averages for precipitation and nitrogen fertilizer applications.

The area-weighted-averaging process resulted in the INFO table "params.quad" which is defined in Appendix A. This table was linked to "counts.quad" with a relate, and the exceedence probability estimates and average parameter values for each quadrangle were written to an external text file. This file was transferred to a DOS computer where the stepwise linear regression was carried out using the program STATGRAPHICS, version 4.0 (STSC 1989)

5.8 EXCEEDENCE FREQUENCIES FROM WUD DATA SET

The nitrate measurement data collected by the Water Utilities Division (WUD) of the Texas Natural Resource Conservation Commission is described in Sections 3.6 and 5.2.5. In order to link the nitrate measurement data to the well locations, and to TWDB quadrangles, the following procedure was followed.

1. Working copies of the point of entry and nitrate measurement tables were made. These were named "poe.wrk" and "nit.wrk". The pws-gwt well coverage was projected from its original coordinate system into the system used for this study, resulting in a coverage called "pws".
2. The pws coverage was overlaid with the 7.5' quadrangle coverage, using the Arc command "IDENTITY". The resulting coverage was called pws-quad, and assigned a quad number to each well in the pws coverage.
3. A redefined item concatenating the water system id and the point of entry was added to the tables poe.wrk and nit.wrk and the point attribute table of pws-quad. This item, called "sysent" acts as a key for linking the three tables.
4. A temporary table called poe.surf was created by copying poe.wrk and purging all records containing a groundwater source (identified by the letter "G" as the first character in the water source entry). This was linked back to poe.wrk by a relate on the sysent item. Every record in poe.wrk that had a related record in poe.surf was purged, leaving only points of entry with no surface water sources in poe.wrk.
4. A temporary table called "wellquad" containing well IDs and quad numbers was created from pws-quad.pat using the arc command "PULLITEMS". This table was joined to poe.wrk using the Arc command "JOINITEM". After the join, all records with quad numbers equal to zero were purged from poe.wrk, leaving only records that could be linked to TWDB quads.

5. A temporary table called "poequad" containing the items for system ID, point of entry number and quad number was created from poe.wrk using "PULLITEMS". The contents of this table were sorted by system ID and point of entry number, then written to a text file called "pquad1" with the tables "UNLOAD" command. This text file was processed with the AWK program

```
awk -f, '$1 != last1 || $2 != last2
{last = $1; last2 = $2; print $0}' pquad1 > pquad2
```

resulting in a file with one entry for each point of entry, containing the identity and a quadrangle for that point of entry. (The fact that this operation could be carried out by so brief a program illustrates the utility of AWK.) The original contents of poequad were purged and replaced with the values in pquad2. This method assigns the quad of the point of entry's first well as the quad of the point. This is somewhat arbitrary if a point draws water from wells in more than one quad, but since most points draw either from a single well or wells in a single quad, this method was judged acceptable.

6. The poequad table was joined to nit.wrk, using "JOINITEM", assigning a quad number to each nitrate reading from a purely groundwater source traceable to a map location.

7. Items were added to the table "counts.quad" to hold the number of nitrate measurements, threshold exceedence counts, and threshold exceedence proportions for each quad

8. The aml program "count_wud.aml" (see Appendix B) was run to count the measurements and exceedences, and to calculate the exceedence proportions by quad for this data set.

Chapter 6: Results

This chapter contains six sections. Section 6.1 reports the results of statewide analyses of nitrate concentrations in Texas groundwater as reported in the TWDB groundwater data system. Maps and histograms in this section show the variation of the estimated probability of nitrate detection by location, discretized into 7.5' quadrangles in the horizontal dimensions only. In addition, graphs present variations in nitrate detection frequency with depth and with time throughout the state.

Section 6.2 shows much the same information for five aquifers selected for additional study. Variations in nitrate detection frequency in two dimensions, with depth, and through time are presented. In addition, the behavior of nitrate in the different aquifers is compared.

Section 6.3 shows the results of the attempt to correlate indicator variables to the variations in nitrate detection rates. Regression results for both statewide and single-aquifer data are presented.

Section 6.4 compares the nitrate detection rates calculated from the TWDB data with an independent set of nitrate measurements collected by the Water Utilities Division of the Texas Natural Resource Conservation Commission as part of their Primary Drinking Water Standards enforcement program.

Section 6.5 presents the results of a comparison of the occurrence of nitrate and herbicides as reported in the US Geological Survey's reconnaissance of groundwater in the mid-continental United States.

Section 6.6 presents a brief summary of the results.

Note that all nitrate concentrations in this chapter are given in equivalent units of elemental Nitrogen (nitrate-N). The Primary Drinking Water Standards define the maximum contaminant level (MCL) as 10 mg/l nitrate-N. The equivalent concentration in nitrate-NO3 is 44.3 mg/l.

6.1 Statewide Results

Table 6.1 shows the total number of nitrate measurements in the base data set (46,507 nitrate measurement records) that exceed four threshold concentrations. The thresholds are 10 mg/l (the MCL), 5 mg/l (half the MCL, and a trigger level for increased monitoring), 1 mg/l (selected to indicate human influence on groundwater, as described in Section 4.1), and 0.1 mg/l (the detection limit selected for this study, as described in Section 3.1). The table also lists the estimated probability of exceeding these thresholds in a measurement selected at random from a well in the State, and the upper and lower bounds on the probability estimate (90% two-sided confidence limits). These probability estimates are based on the assumption (described in Section 4.4.1) that the nitrate measurements compose a sample generated through a Bernoulli Process, resulting in a binomial distribution of threshold exceedences. The exceedence probability estimates are calculated by dividing the number of measurements exceeding the threshold by the total number of measurements. The upper and lower bounds on the estimates are calculated using the method described in Sections 4.4.1 and 5.5.1. Because the number of measurements used to calculate these estimates is large, the upper and lower bounds are close to the estimates. This is not the case when estimates are based on smaller numbers, such as those associated with a single 7.5' quadrangle.

Table 6.1 Nitrate Exceedences in Texas (46,507 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	29,643	0.6374	0.6337	0.6411
1	20,312	0.4368	0.4329	0.4405
5	7,411	0.1594	0.1566	0.1622
10	4,166	0.0896	0.0874	0.0917

Of the 4,407 7.5' quadrangles that make up the map of Texas used in this study, nitrate measurements are reported in 3554. Exceedence probabilities were estimated for these quadrangles at the four concentration thresholds by the same method as those in Table 6.1. Figures 6.2, 6.4, 6.6, and 6.8 show the spatial distribution of the resulting exceedence probabilities across Texas. An exceedence probability estimate was included in the statewide maps if twelve or more nitrate measurements are recorded for the quadrangle. As shown in Section 4.4.1, this means that for a 50% exceedence probability, the upper and lower limits of the two-sided 90% confidence interval of the probability estimate are 0.25 and 0.75, respectively. In somewhat less abstract terms, if a cell has an 50% exceedence probability estimated from twelve measurements, that cell's true exceedence probability is greater than 25% and less than 75% in nine cases out of ten. Cells with either more measurements or exceedence probabilities closer to zero or one will have narrower confidence intervals.

Histograms of the probability estimates for the quadrangles are presented in Figures 6.1, 6.3, 6.5, and 6.7. Each histogram displays two sets of bars. The taller bars show the number of quadrangles falling in the indicated probability range when all 3554 quads with measurements are counted. This would include, for example, a quad with only one measurement (which must have an estimated exceedence probability of 1 or 0). The shorter bars show quads falling in the

indicated probability range from which at least twelve measurements have been collected.

Exceedences of Detection Limit. Figures 6.1 and 6.2 show estimates of the nitrate detection (i.e., measurement in excess of 0.1 mg/l) probability in the 7.5' quadrangles.

At the detection limit of 0.1 mg/l, nitrate is safely within the range of background concentrations. Nearly a third of the quadrangles with measurements (1160 out of 3554 quads) have never reported a concentration at or below this limit and in more than a third (1320 quads), fewer than one measurement in ten has fallen at or below the detection limit.

The map in Figure 6.2 shows that, although detectable levels of nitrate are found throughout the State, measurements below the detection limit are much more common in eastern Texas. Of the 1158 mapped quadrangles, only one west of the 100th meridian (the eastern boundary of the panhandle) has a detection rate below 20%.

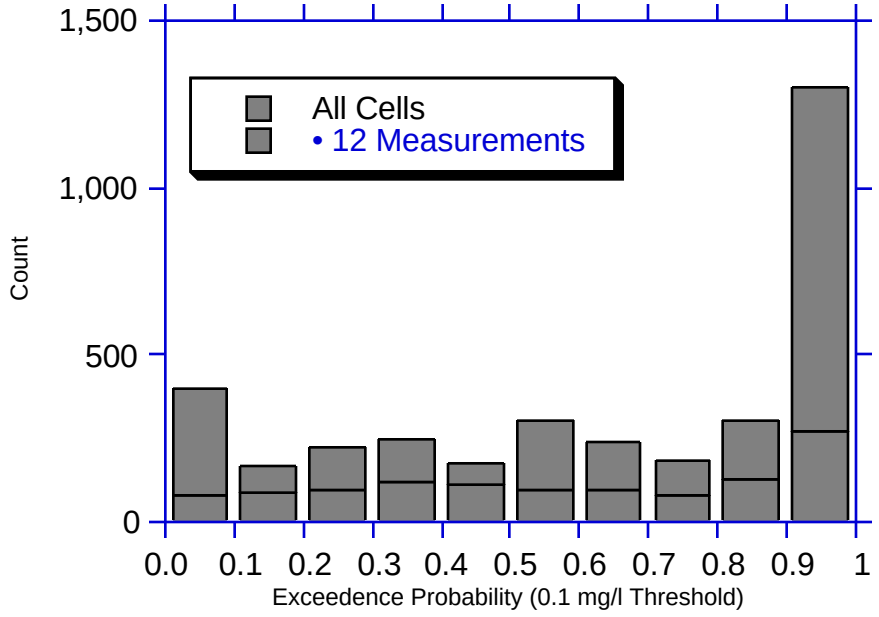


Figure 6.1 Nitrate Detection Histogram

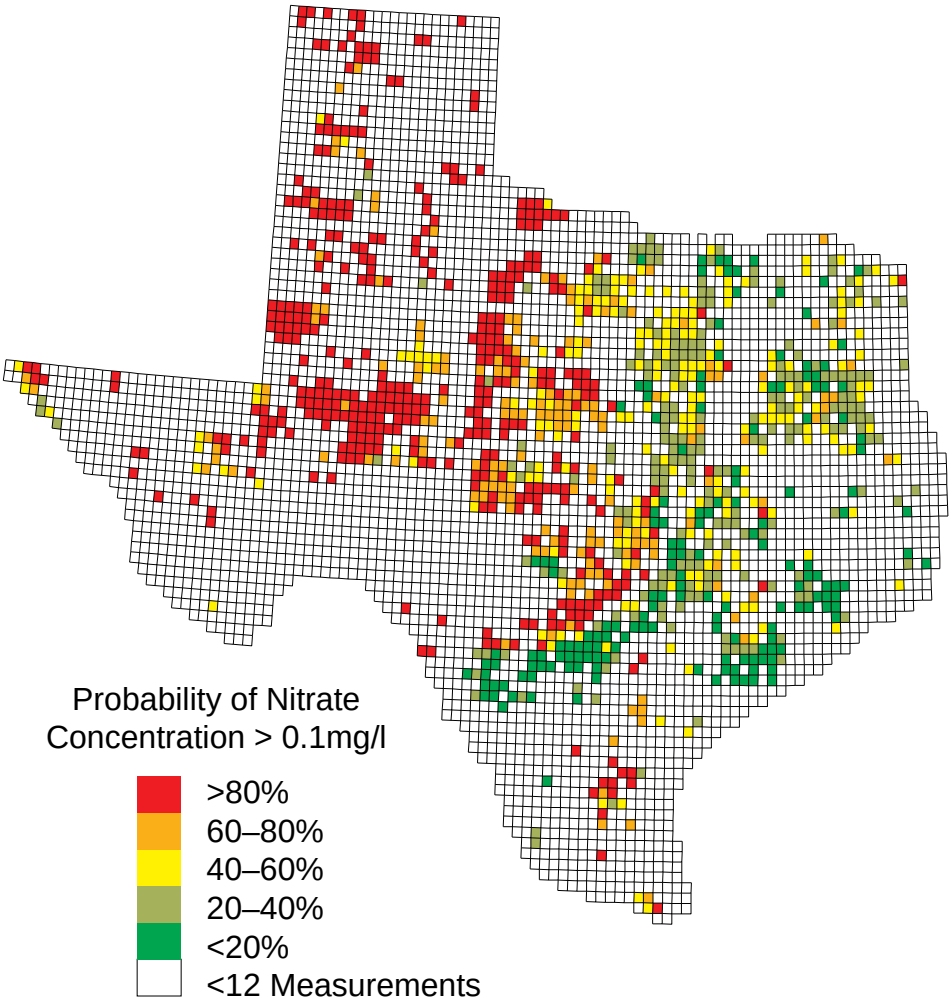


Figure 6.2 Spatial Distribution of Nitrate Detection Probabilities (Binomial Estimate)

Exceedences of 1.0 mg/l Threshold. Figures 6.3 and 6.4 show exceedence probabilities at the 1 mg/l level.

At 1 mg/l, the nitrate concentration is in an ambiguous range. Although this is considerably higher than the normal background level, concentrations of up to 3 mg/l in groundwater are frequently attributed to natural sources (Madison and Brunett 1985). At 1 mg/l, however, it is reasonable to be suspicious of human influences.

Because there are fewer exceedences of the 1 mg/l concentration threshold than of the detection limit, there are more quadrangles with near-zero exceedence probabilities. Figure 6.4 shows an increase in exceedence probability from east to west similar to that seen in the 0.1 mg/l map, but regions of high exceedence probability are more local and less regional in scope. The difference between the Carrizo-Wilcox Aquifer and the adjacent Balcones Fault Zone of the Edwards Aquifer (see Section 6.2) is quite apparent, for example. This result is similar to observations by Baker et al (1994), who noted that "River valley aquifers, sandy soils with high water tables, karst areas, and reef structures with surficial expressions are all reflected in county maps" developed as part of a voluntary well testing program.

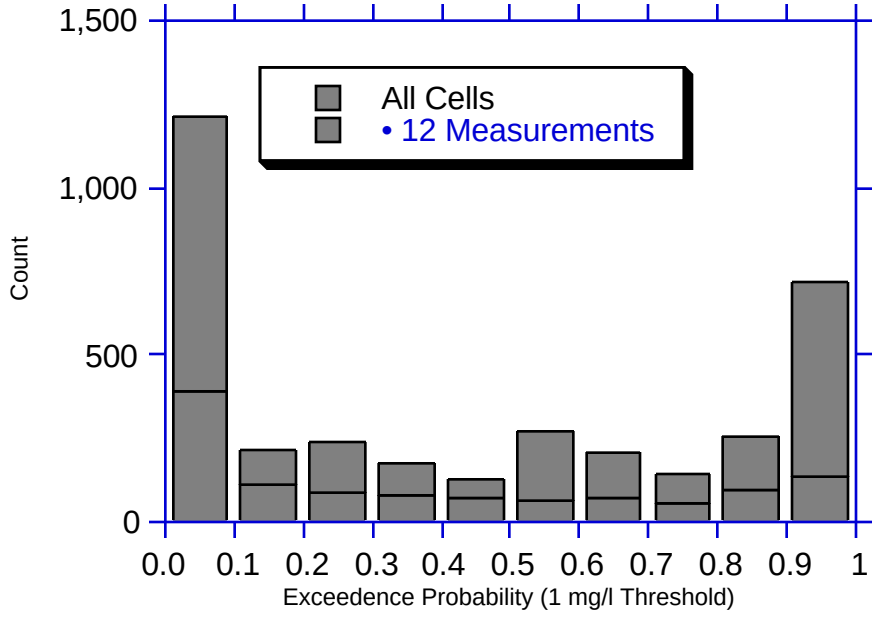


Figure 6.3 Nitrate 1 mg/l Exceedence Histogram

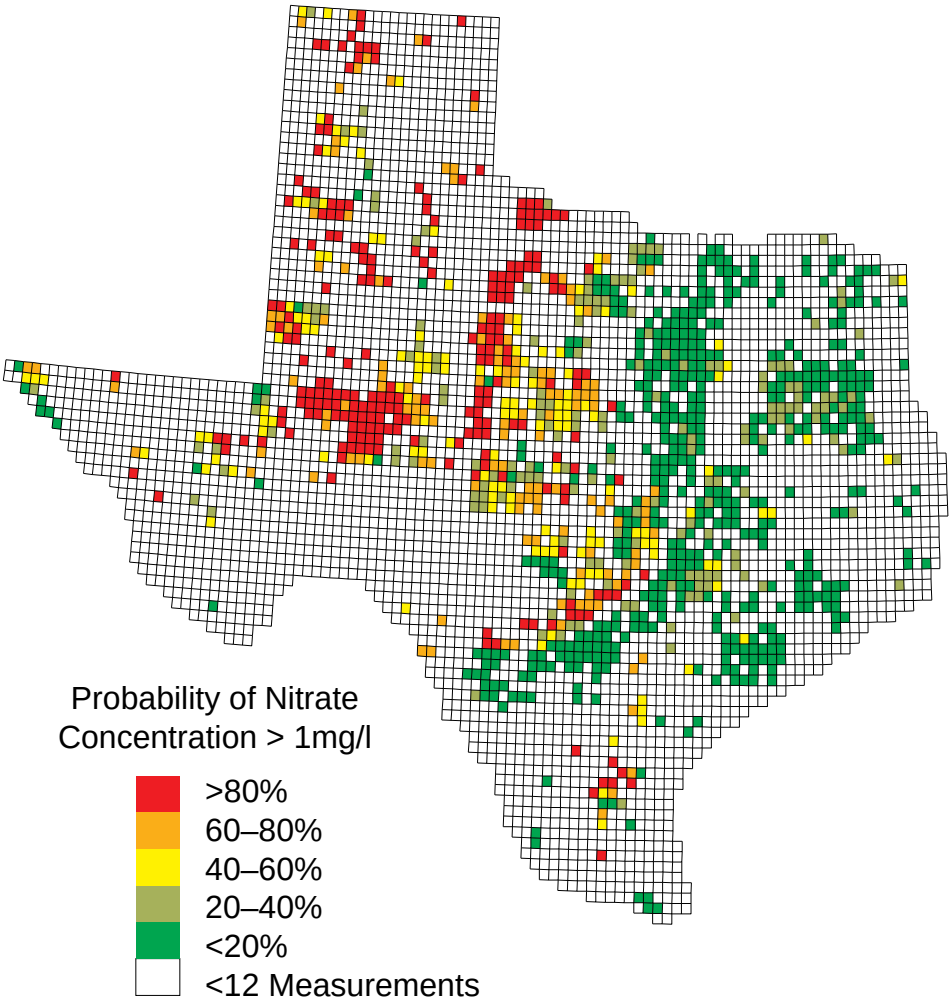


Figure 6.4 Spatial Distribution of 1 mg/l Exceedence Probabilities (Binomial Estimate)

Exceedences of 5 mg/l Threshold. Figures 6.5 and 6.6 show exceedence probabilities at the 5 mg/l threshold.

Nitrate concentrations at or above 5 mg/l due to natural sources are not unheard of, but are very uncommon. Consistent measurements of nitrate above this level clearly indicate either an extraordinarily strong natural source, or the influence of human activities. Also, this concentration is one-half of the MCL for nitrate and, although not considered high enough to endanger human health, it does trigger a switch from annual to quarterly monitoring for nitrate in public water supplies using groundwater (40 CFR 141).

Of the 1158 cells mapped, 1124 have an estimated exceedence probability of 0.0 at the 5 mg/l threshold—in only 34 of these quads has a concentration above this level been measured. The east-to-west trend of increasing exceedence probability seen in the previous maps has been replaced by a group of cells in the western part of north-central Texas, and a scattering of isolated cells mostly in the western part of the State.

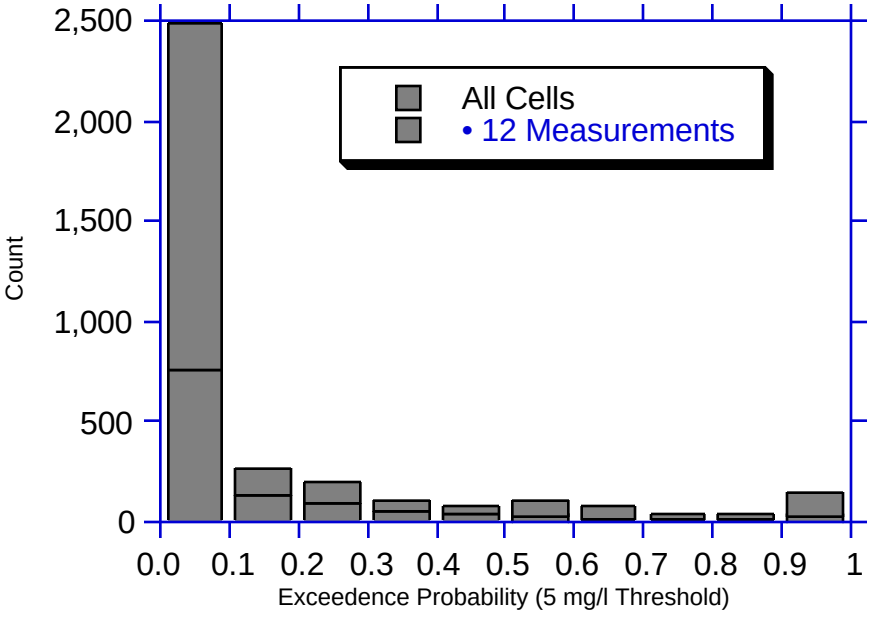


Figure 6.5 Nitrate 5 mg/l Exceedence Histogram

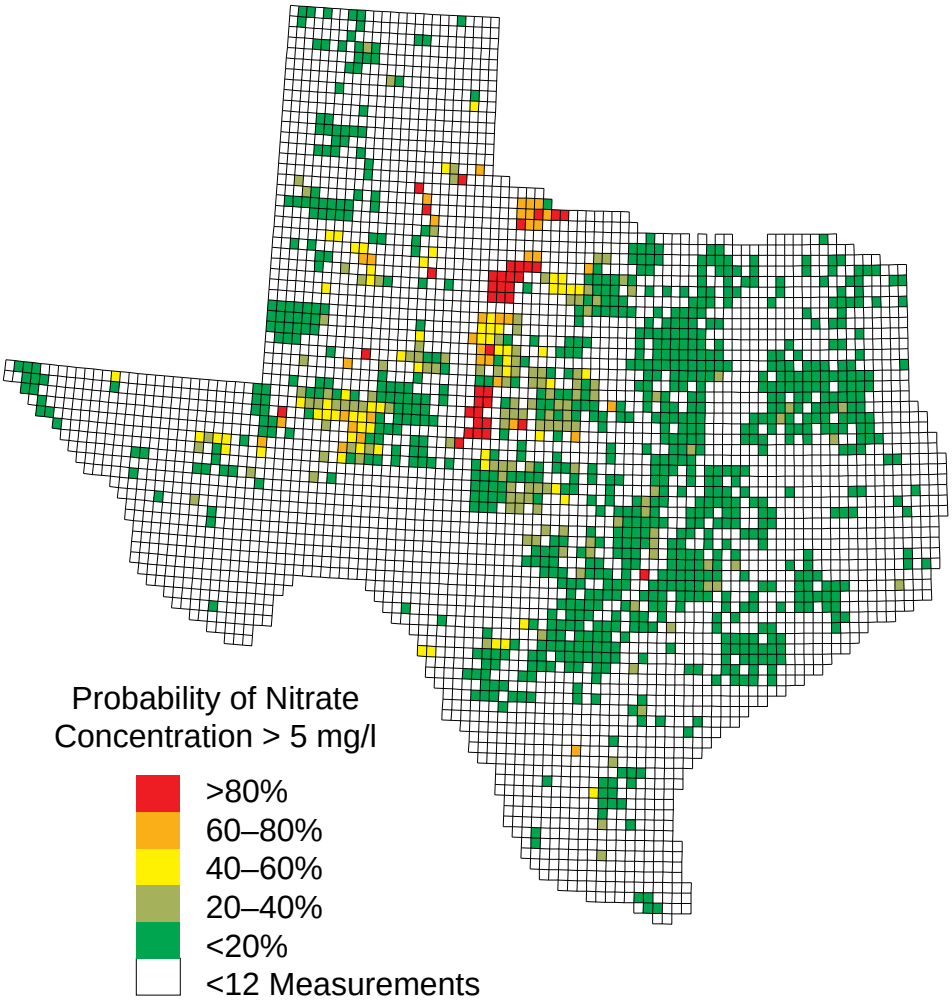


Figure 6.6 Spatial Distribution of 5 mg/l Exceedence Probabilities (Binomial Estimate)

Exceedences of 10 mg/l Threshold. Figures 6.7 and 6.8 show exceedence probabilities at the 10 mg/l level.

At 10 mg/l, nitrate is considered a human health hazard, and public water suppliers are required to notify the public and take action to reduce the nitrate concentrations when they exceed this level. Concentrations at this level are very rarely due to natural sources. The vast majority of cells with measurements (2708 of 3554) have never had a measurement exceeding this limit.

The map of 10 mg/l exceedence probabilities in figure 6.8 shows only a few quads where this high level of nitrate concentration is found often. Although nearly one in twelve measurements listed in the base data set (4,166 of 46,507) exceeds 10 mg/l, these elevated nitrate levels are very unevenly distributed in space. The only region where exceedences are found consistently, rather than in isolated quads, is in western north-central Texas in an area roughly co-incident with the extent of the Seymour Aquifer (see Section 6.2).

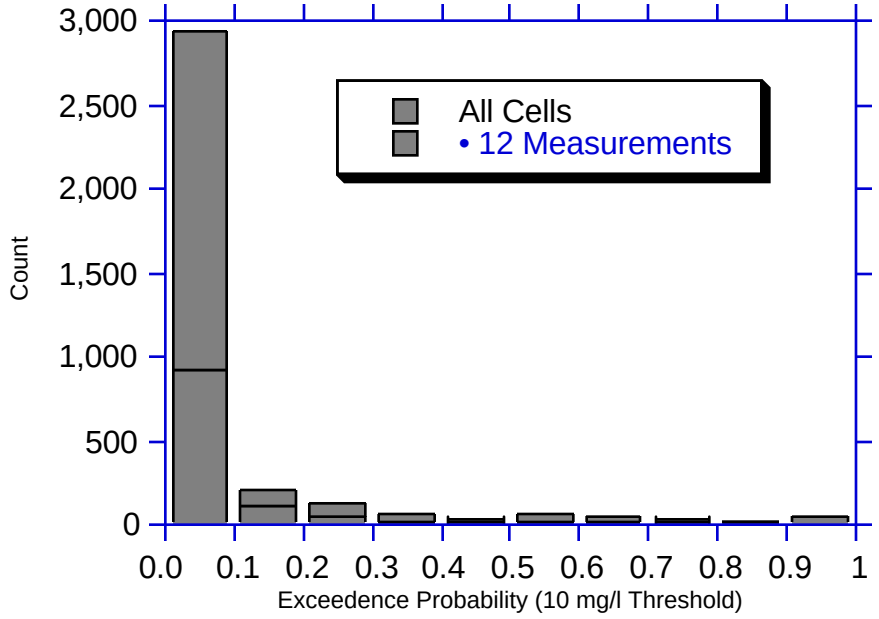


Figure 6.7 Nitrate 10 mg/l Exceedence Histogram

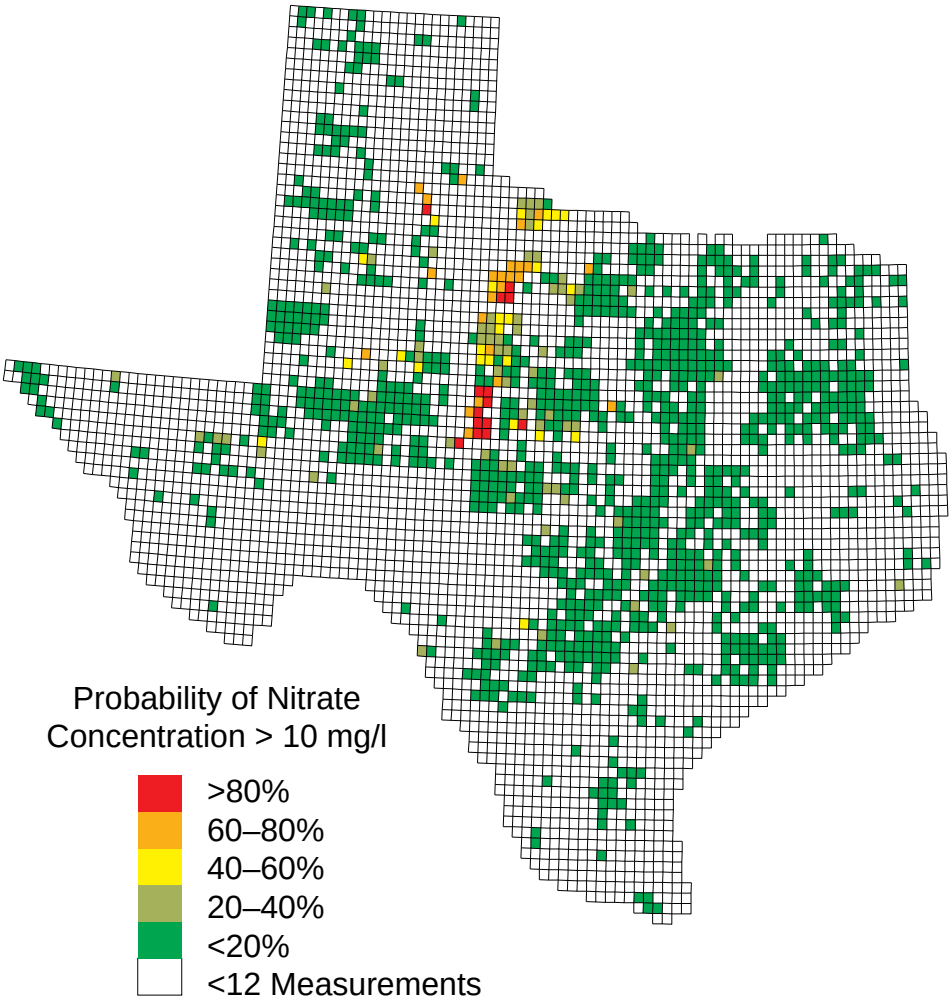


Figure 6.8 Spatial Distribution of 10 mg/l Exceedence Probabilities (Binomial Estimate)

Upper Bound Exceedence Estimates. The exceedence probabilities shown in Figures 6.1–6.8 are all the best estimates of discrete probabilities, calculated by simple division of number of exceedences by number of measurements at four thresholds. Figure 6.9 presents the 95% upper confidence limit on the binomial estimate of the 1 mg/l exceedence probability. By combining the estimated exceedence probability with a measure of the confidence in that estimate, this map presents a conservative estimate of the probability of nitrate contamination in the quadrangles. A cell has an 95% upper confidence limit value of 0.95, for example, if 100 measurements have been taken and 91 have exceeded the threshold, or if 1 measurement has been taken and that measurement did not exceed the threshold. A quad can have a low exceedence probability only if many measurements have been taken and few exceedences have been found.

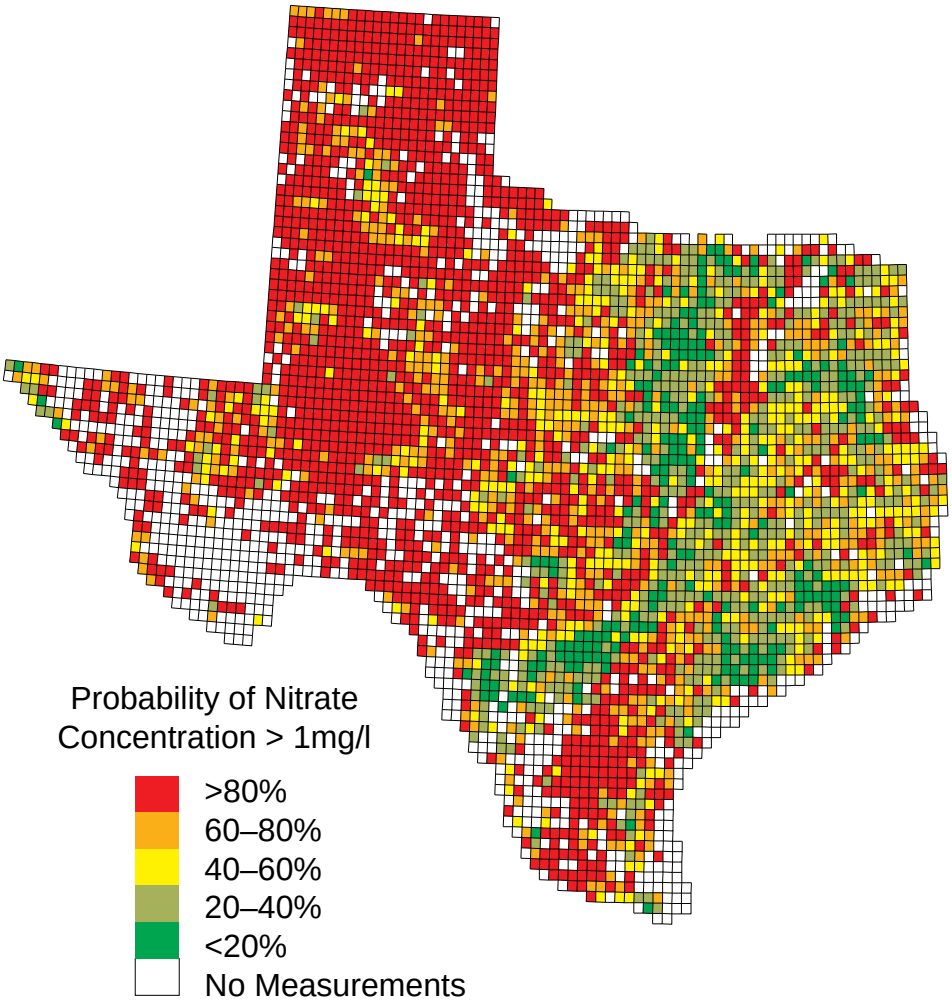


Figure 6.9 Spatial Distribution of Upper 95% Confidence Limit on 1 mg/l Exceedence Probabilities (Binomial Estimate)

Lognormal Exceedence Estimates. In addition to the discrete exceedence probabilities calculated by the binomial estimation method, parameters were calculated for the best-fitting lognormal distribution for each quadrangle with twelve or more measurements. Figure 6.10 shows the spatial distribution of the lognormal estimates of the 1 mg/l exceedence probabilities for quads with at least twelve measurements and one detection of nitrate.

To compare the lognormal distribution to the discrete probabilities, Figure 6.11 shows paired-value plots of the lognormal and discrete exceedence probability estimates at the detection limit, 1, 5, and 10 mg/l threshold concentrations. A point on one of the four graphs is located at coordinates equal to the binomial and lognormal exceedence probability estimates for one quadrangle. A point falls on the diagonal line if the two estimates are identical, above the line if the lognormal estimate is larger, and below the line if the binomial estimate is larger. In comparison to the binomial estimates, the lognormal estimates tend to be higher at the detection limit, 5 and 10 mg/l thresholds, and lower at the 1 mg/l threshold. At the higher concentration thresholds, the lognormal distribution tends to over-predict exceedences with low probabilities, and under-predict exceedences at high probabilities. One possible explanation of the differences in the predictions is that the true probability distributions have longer tails (i.e., more probability distributed to extreme high and low values) than the lognormal distribution allows.

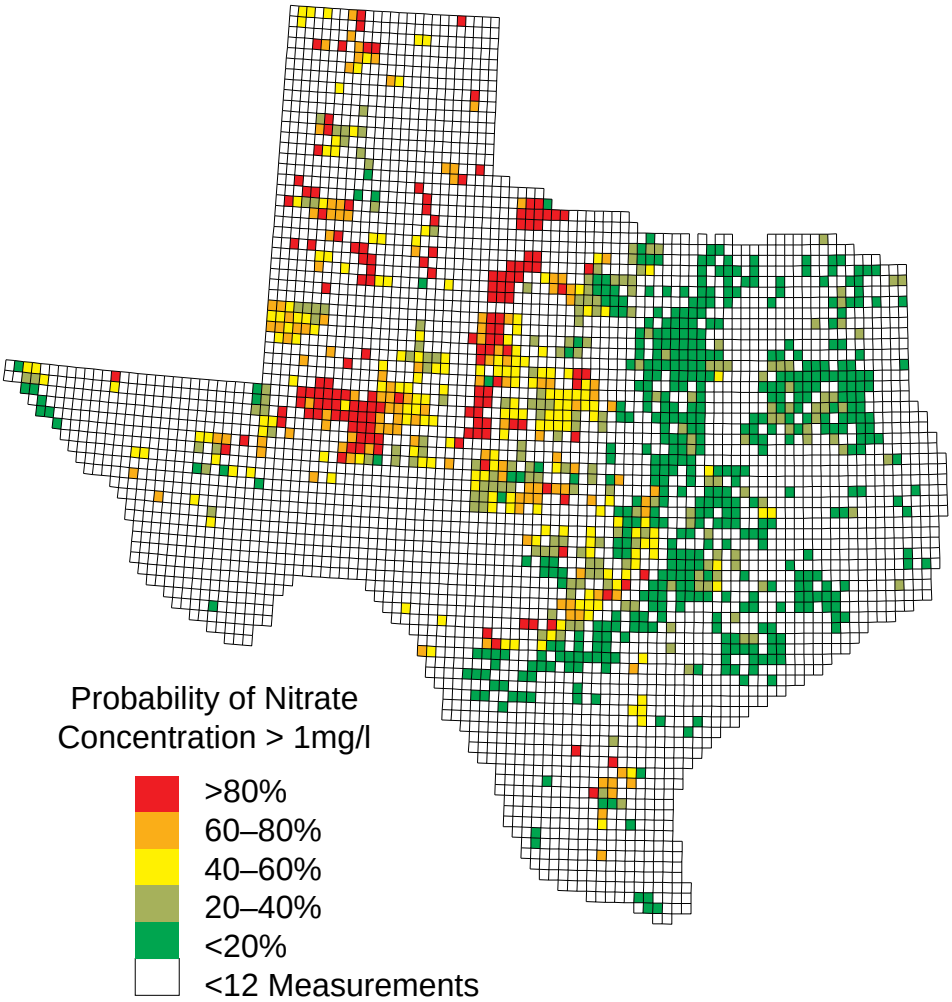


Figure 6.10 Spatial Distribution of 1 mg/l Exceedence Probabilities (Lognormal Estimate)

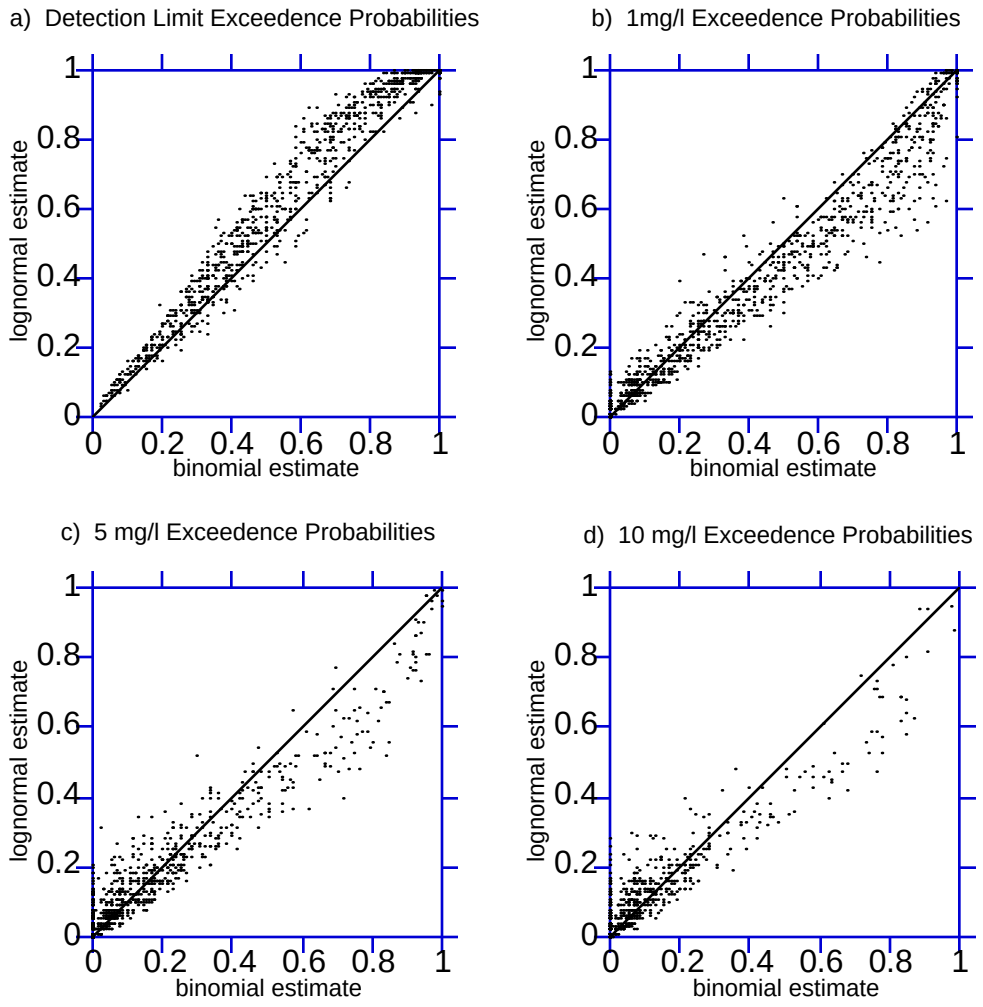


Figure 6.11 Comparison of Discrete and Lognormal Exceedence Probability Estimates

A numerical analysis of the differences between the discrete and lognormal estimates of exceedence probabilities confirms what a visual inspection of the paired-value plots in Figure 6.11 suggests. At all four threshold values, the hypothesis that the two estimates consistently differ can be confirmed with greater than 99.9% confidence using the sign test (Helsel and Hirsch 1992). The more commonly used paired-t test is inappropriate here because the differences between the two estimates are not normally distributed (tested with Probability Plot Correlation Coefficient test). The results of these tests are listed in Tables 6.2 and 6.3.

Table 6.2 Probability Plot Correlation Coefficient Test Results

Threshold	**PPCC**	α_{PPCC}
0.1 mg/l	0.988	<0.1
1.0 mg/l	0.963	<0.005
5.0 mg/l	0.948	<0.005
10.0 mg/l	0.913	<0.005

The entry "PPPC" in Table 6.2 is the correlation coefficient between the probability plotting position values (using Blom's Formula) for the binomial and lognormal estimates of the exceedence probabilities for the listed thresholds in the 1134 quadrangles with at least 12 nitrate measurements and at least one nitrate detection. The entry "α_{PPCC}" is the significance level of the test—the probability that the differences between the two estimates are normally distributed. The significance levels are expressed as upper bounds because the PPCC table in Helsel and Hirsch only has exact values for up to 100 pairs.

Table 6.3 Sign Test Results

Threshold	n	+	Z
0.1 mg/l	1004	101	-25.3
1.0 mg/l	1053	626	+6.1
5.0 mg/l	918	343	-7.6
10.0 mg/l	800	236	-11.5

The entry "n" in Table 6.3 is the number of quadrangles (out of the 1134 with both lognormal and binomial exceedence probability estimates) with different values for the two exceedence probability estimates. The entry "+" is the number of quads (out of n) in which the binomial estimate is greater than the lognormal estimate. The entry "Z" is the normal variate corresponding to the probability that the binomial estimates are consistently greater than the lognormal estimates of the exceedence probability. The normal variates are calculated by using the large-sample approximation of the sign test, as given in Helsel and Hirsch.

So far, the variation of nitrate concentration of nitrate exceedence probabilities has been limited to the two horizontal dimensions. Two more dimensions, depth and time, have yet to be considered.

Influence of Well Depth. Figure 6.12 shows the variation of the four exceedence probabilities with depth over the State. The graph was prepared by calculating the estimated probability of detecting nitrate at the threshold level (number of exceedences divided by number of measurements) for all wells at least as deep as the value shown on the horizontal axis. The values shown intersecting the left vertical axis are equal to the exceedence probabilities calculated for the 46,507 measurements in the base data set. Values were calculated at ten-foot intervals of depth. The markers on the lines of the graph are present to help distinguish the lines, not to indicate points at which values were estimated.

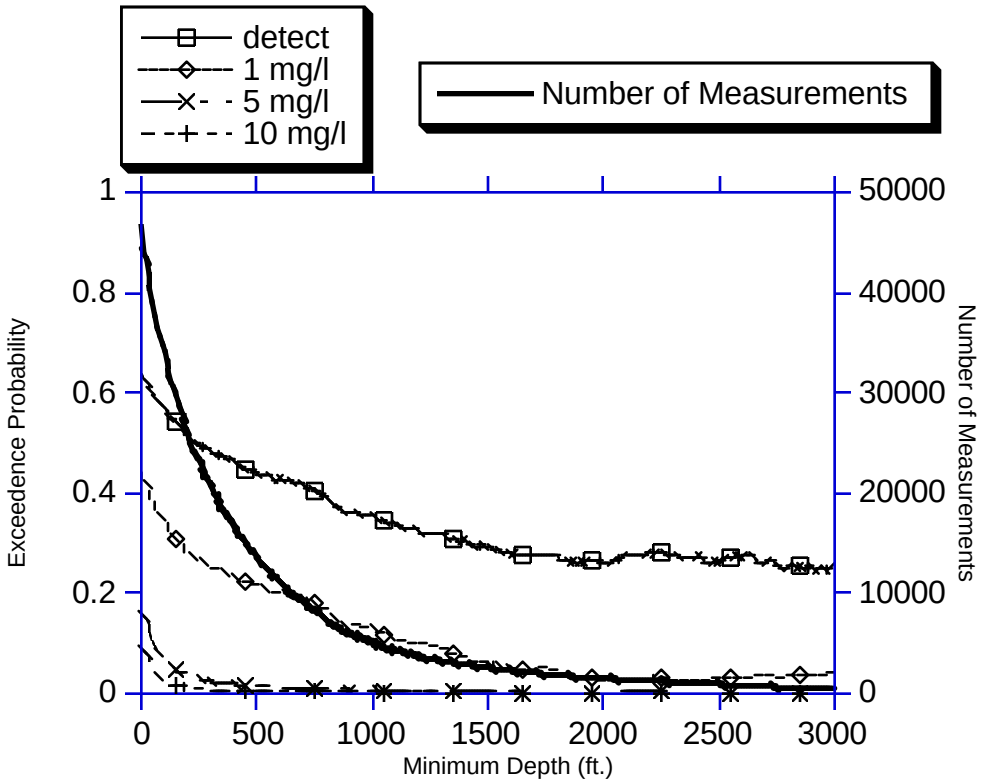


Figure 6.12 Variation of Exceedence Probabilities with Depth

A decrease in the likelihood of detecting nitrate at any threshold level is clearly visible as shallower wells are excluded from the calculation of the exceedence probabilities. This decrease is most pronounced as the shallowest wells are excluded, especially at the higher concentration thresholds. Of 4,166 measurements in exceedence of the MCL, 3,834 (about 92%) were taken from wells less than 200 feet deep.

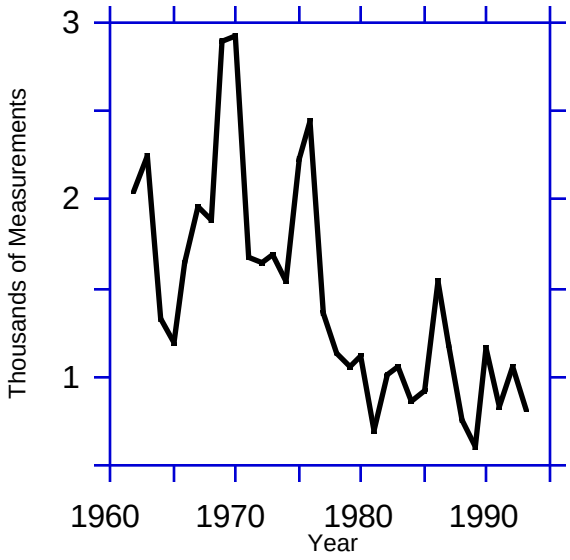


Figure 6.13 Measurements by Year in Base Data Set

Trends through Time. Figure 6.13 shows the number of measurements listed in the base data set for each year from 1962 to 1993. Figure 6.14 shows the variation of the four exceedence probabilities with the year in which the nitrate measurements were taken. In this graph, a marker is plotted for each exceedence probability calculated for the measurements collected in each year.

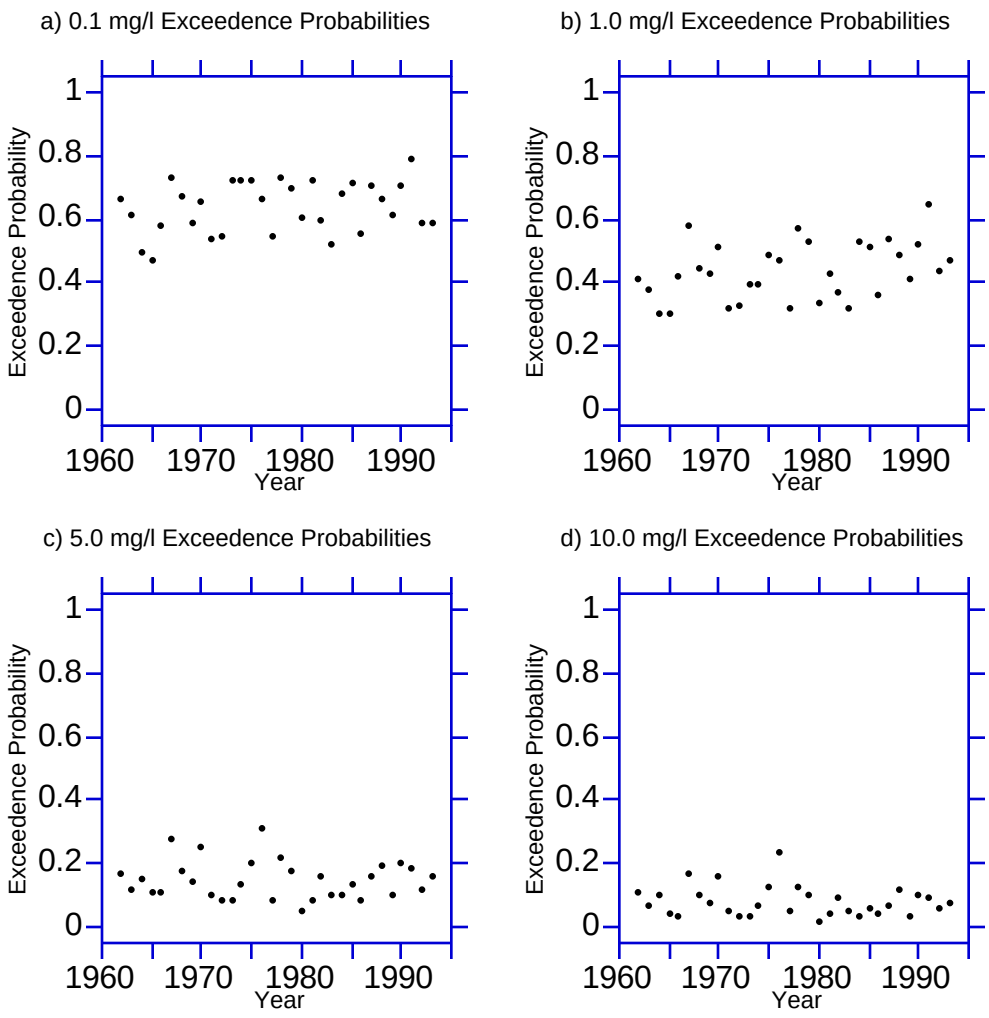


Figure 6.14 Variation of Exceedence Probabilities Over Time

In general, the variability from one year to the next (possibly reflecting changes in sampling locations) is much greater than any trend through time. Linear regression of exceedence probabilities against time confirms this for the detection limit and the 5 and 10 mg/l thresholds. The regression results for the four threshold concentrations are summarized in Table 6.4. The fitted line is measured is considered statistically significant if its t statistic is greater than 2, indicating a probability of less than 5% that the slope does not differ from zero. By this

measure, only the fitted line for the 1 mg/l threshold is significant. The slope of 0.003, indicates that the likelihood that a nitrate measurement selected at random from anywhere in the state will exceed 1 mg/l has increased by about three-tenths of a percent each year over the last 30 years. The data and the regression line for this threshold are shown in Figure 6.15.

Table 6.4 Regression Results For Threshold Exceedences through Time

Threshold	**Slope**	**t**
0.1 mg/l	0.002	1.33
1.0 mg/l	0.003	2.07
5.0 mg/l	-0.0005	0.42
10.0 mg/l	-0.0008	0.83

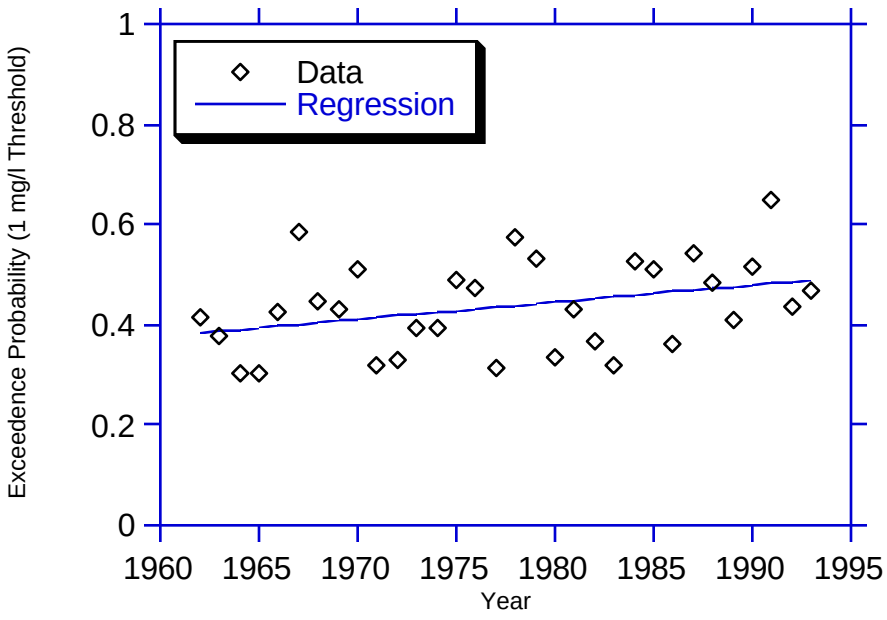


Figure 6.15 Regression of 1 mg/l Exceedence Probability Against Time

Examination of groundwater nitrate measurements statewide shows that there is considerable spatial variation in the likelihood of detecting nitrate at any

threshold level. At lower concentrations, there is a general trend of increasing exceedence probability from southeast to northwest, which becomes more localized as the threshold increases. In general, deep wells are less likely to yield high concentrations of nitrate than shallow wells. Although trends in nitrate detection through time are not strong, a significant increase with time in the likelihood of detecting nitrate at the 1 mg/l level has been found. Since increases through time, especially on a as short a time scale as thirty years, are suggestive of human influence, this tends to confirm the usefulness of the 1 mg/l threshold as an indicator of susceptibility of groundwater to human activities.

6.2 SELECTED AQUIFERS

This section reports nitrate detections in wells associated with the five aquifers selected for special study. Figure 6.16 shows the locations of the five selected aquifers on a map of Texas. The map was created by color-coding 7.5' quadrangles by the aquifer associated with wells in that quadrangle. A quadrangle was colored yellow, for example, if it contains a well associated with the Hueco-Mesilla Bolson Aquifer in the study's table of wells. Because the horizontal extent of the Carrizo-Wilcox Aquifer and the Balcones Fault Zone of the Edwards Aquifer overlap, the seven quads that contain wells in both of these aquifers were colored black. The selection of wells to associate with the aquifers is described in Section 3.2.3, which also includes a map of the TWDB's location of the aquifers' boundaries (Figure 3.5).

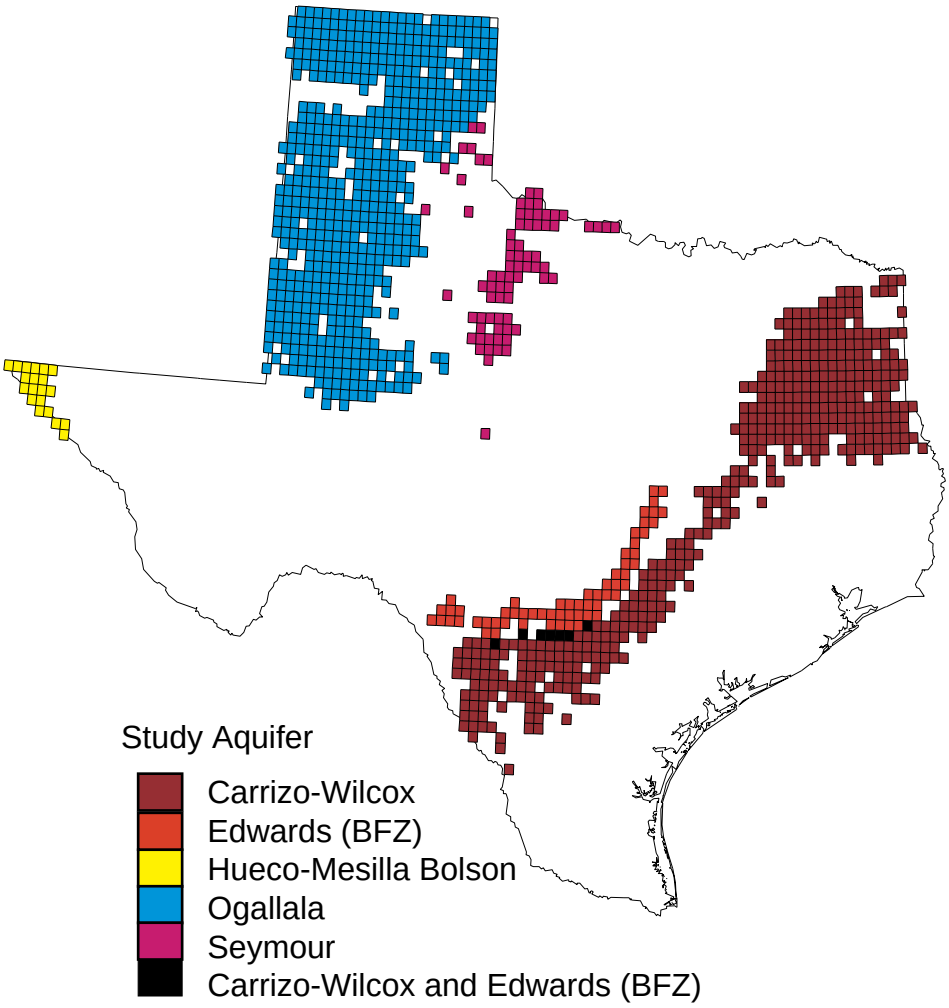


Figure 6.16 7.5' Quadrangles Associated with Study Aquifers

Table 6.5 duplicates Table 3.6, listing the number of wells and measurements associated with each aquifer, and also includes the number of 7.5' quadrangles shown for the aquifer in Figure 6.15. Note that "Edwards (BFZ)" refers to the Balcones Fault Zone of the Edwards Aquifer.

Table 6.5 Wells and Measurements in Selected Aquifers

Aquifer	**Wells**	**Measurements**	**Quadrangles**
Carrizo-Wilcox	2292	4597	433
Edwards (BFZ)	412	1691	67
Hueco-Mesilla Bolson	404	1908	20
Ogallala	3483	4430	588
Seymour	1993	2526	76

Sections 6.2.1 through 6.2.5 describe the results of a variety of analyses of nitrate measurements in each of the five study aquifers. For each aquifer, a table of exceedence probabilities, a map of the spatial distribution of the exceedence probabilities, and charts of variation of exceedence probabilities are presented. This is essentially the same information, presented in the same manner, as was given for the State as a whole in Section 6.1.

Section 6.2.6 presents summary information for all five aquifers and compares the results among them.

6.2.1 Carrizo-Wilcox Aquifer

Table 6.6 Nitrate Exceedences in the Carrizo-Wilcox Aquifer (4597 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	1124	0.2445	0.2341	0.2552
1	327	0.0711	0.0650	0.0777
5	113	0.0245	0.0209	0.0286
10	63	0.0137	0.0110	0.0169

Of the five study aquifers, the Carrizo-Wilcox is the least contaminated by nitrate. Fewer than 25% of the measurements listed in the database show even a detectable level of nitrate. The nitrate detections occur without much coherent spatial pattern within the aquifer (Figure 6.17), or with much variation with depth (Figure 6.18), although 81 of the 113 nitrate measurements exceeding 5 mg/l came from wells less than 200 feet deep. It may be significant that the quads with the highest 1 mg/l exceedence probabilities are on the western edge of the aquifer, which the TWDB identifies as an outcrop zone.

As with the State as a whole, there is more variability from year to year in nitrate detection rate than discernible trend through time (Figure 6.19). Regression of detection rates against time showed no significant trends at any threshold level.

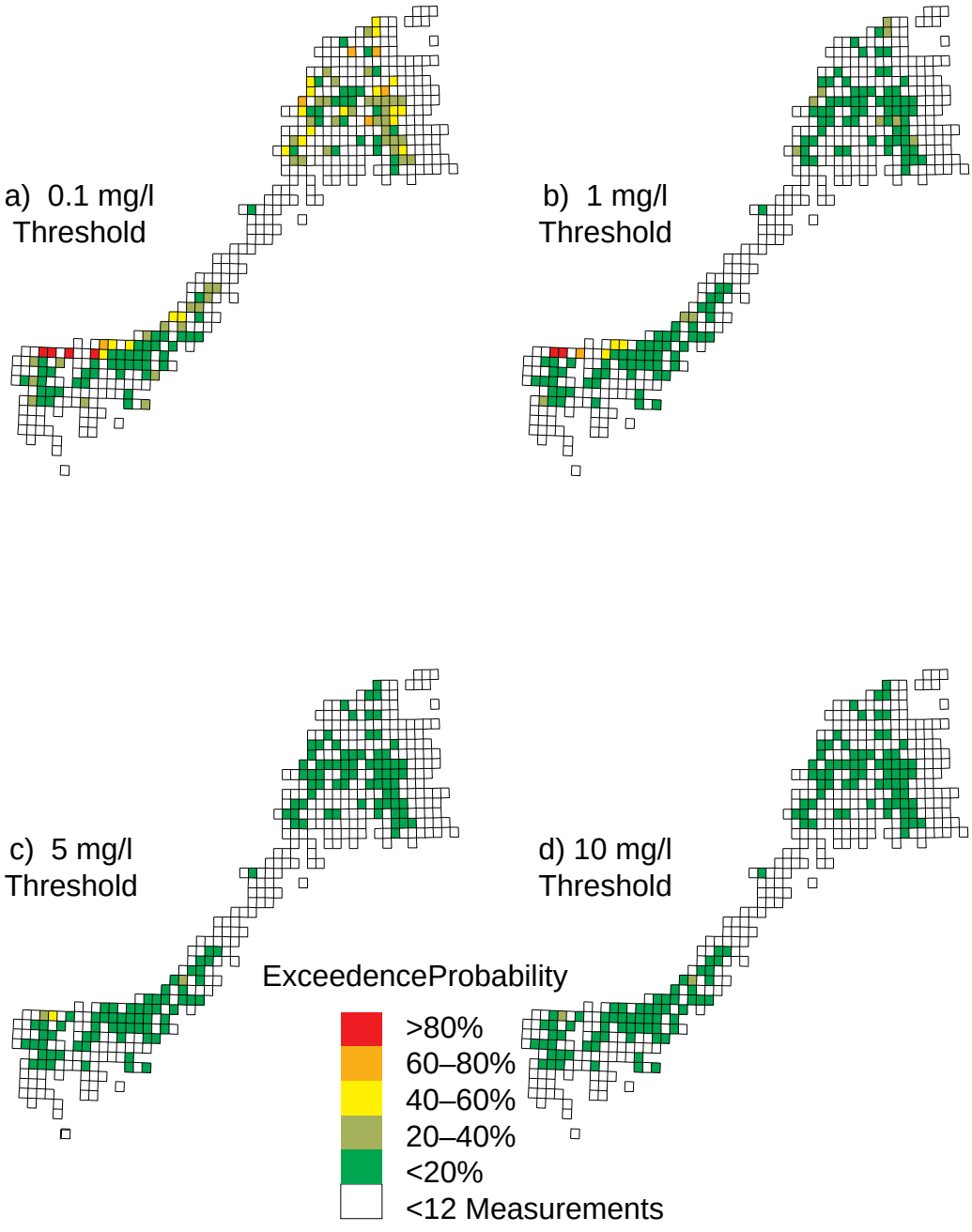


Figure 6.17 Estimated Nitrate Exceedence Probabilities by Quadrangle in the Carrizo-Wilcox Aquifer

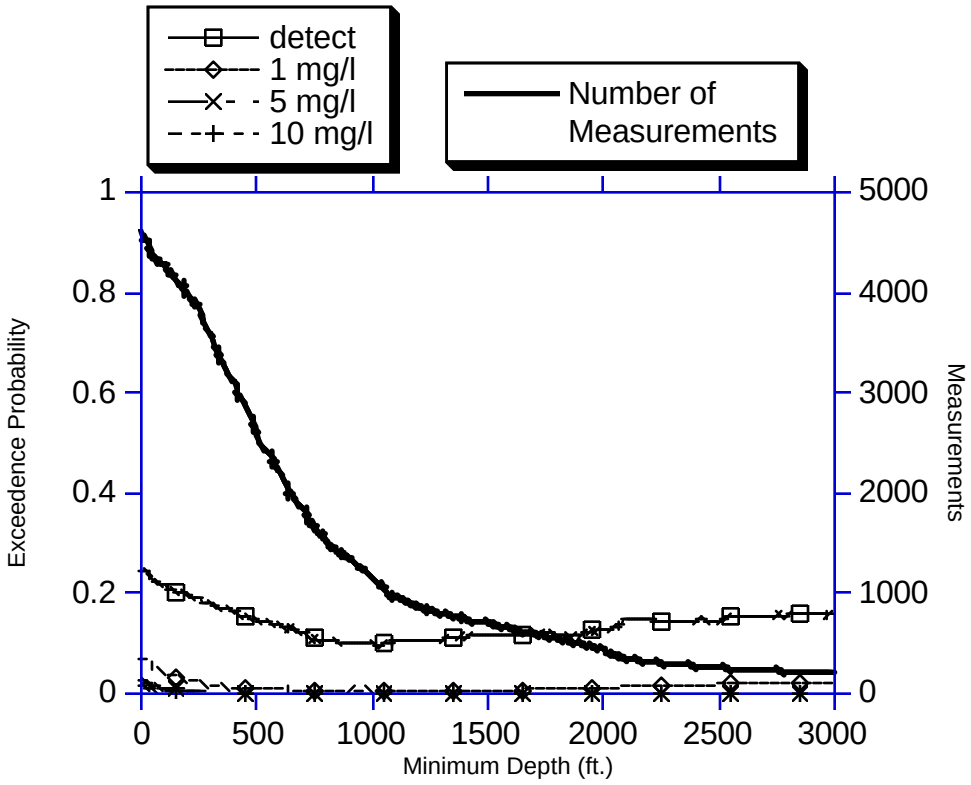


Figure 6.18 Variation of Exceedence Probabilities with Depth in the Carrizo-Wilcox Aquifer

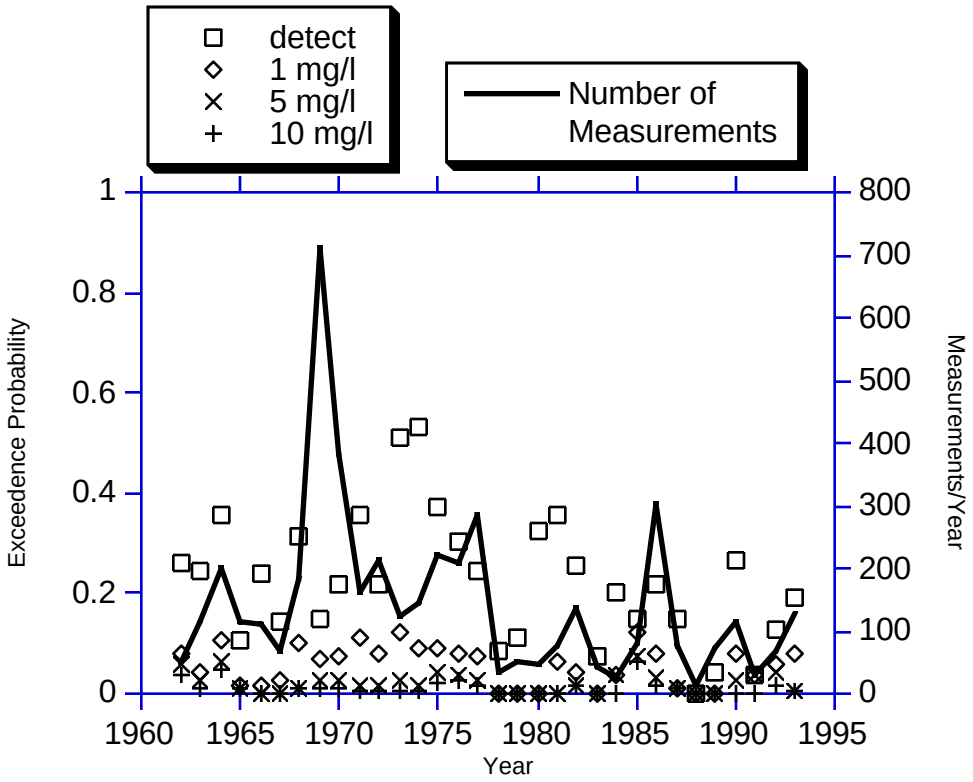


Figure 6.19 Variation of Exceedence Probabilities Over Time in the Carrizo-Wilcox Aquifer

6.2.2 Edwards Aquifer (Balcones Fault Zone)

Table 6.7 Nitrate Exceedences in the Balcones Fault Zone of the Edwards Aquifer (1691 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	1581	0.9350	0.9243	0.9445
1	1248	0.7380	0.7199	0.7556
5	13	0.0076	0.0046	0.0122
10	4	0.0024	0.0008	0.0054

Although the likelihood of *detecting* nitrate is lowest in the Carrizo-Wilcox aquifer, the likelihood of a measurement exceeding 5 mg/l is lowest in the Balcones Fault Zone of the Edwards aquifer. The map of the spatial distributions of exceedence probabilities (Figure 6.20) shows no obvious patterns in detections, but reveals a dramatic shift from high to low probabilities between the 1 mg/l and 5 mg/l thresholds. The same shift is visible when exceedence probabilities are plotted against well depth and time.

Figure 6.21 shows a slight decrease in the likelihood of detecting nitrate as deeper wells are examined, but the trend is not clear until a depth of 1,000 feet is reached. A sharp drop in nitrate detections is associated with the deepest wells (<1700 ft.), but since this is a very small number of wells, the significance of this decrease is unclear.

No significant trends through time are seen in detection rates at any threshold level. Figure 6.22 shows detection probabilities consistently close to 90%, and exceedence probabilities at the 5 and 10 mg/l level consistently close to zero. The 1 mg/l exceedence probability shows considerable variation but no consistent trend through time.

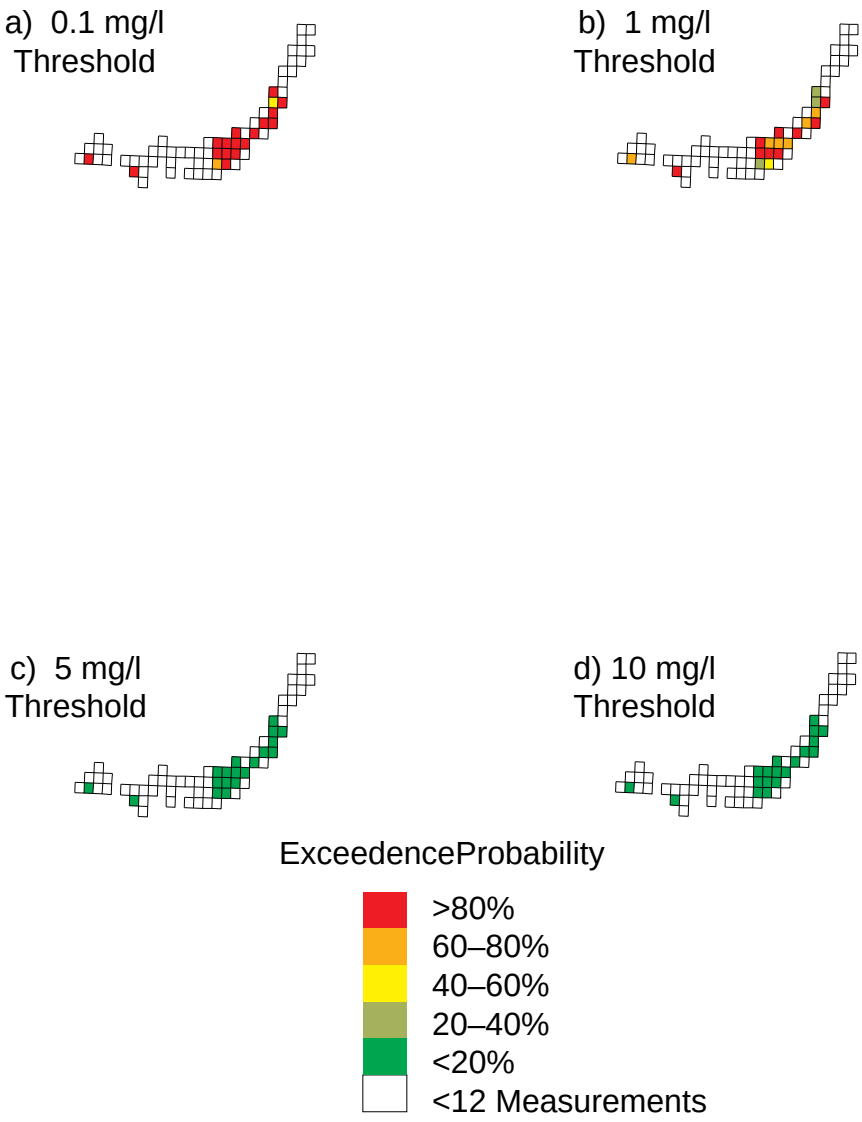


Figure 6.20 Estimated Nitrate Exceedence Probabilities by Quadrangle in the Balcones Fault Zone of the Edwards Aquifer

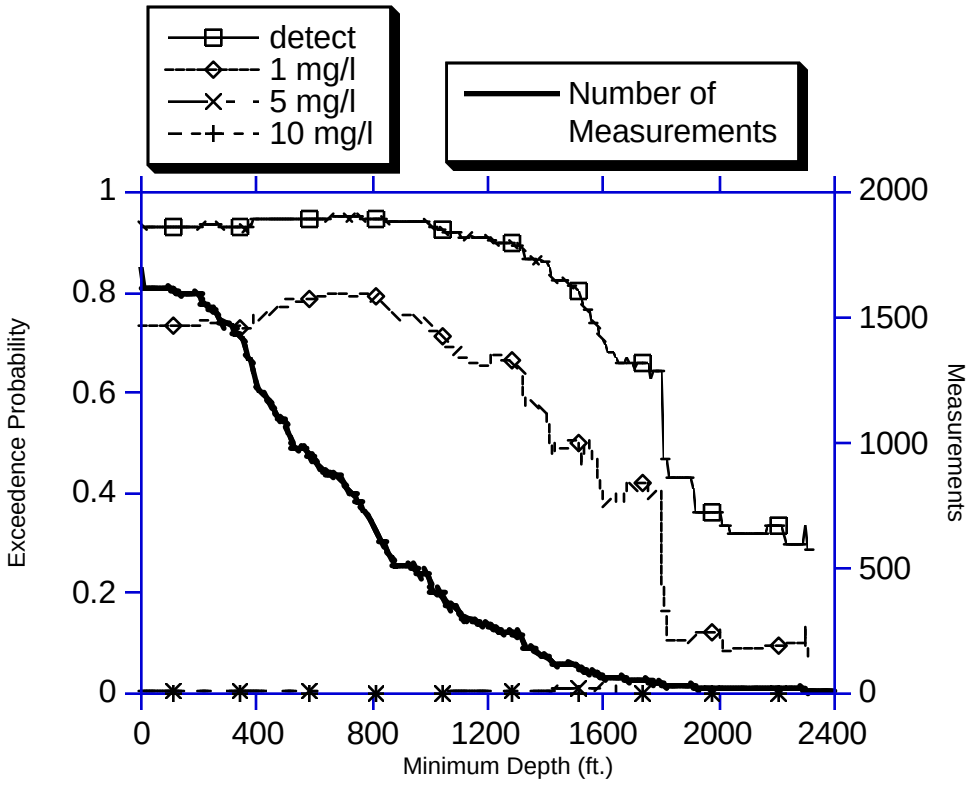


Figure 6.21 Variation of Exceedence Probabilities with Depth in the Balcones Fault Zone of the Edwards Aquifer

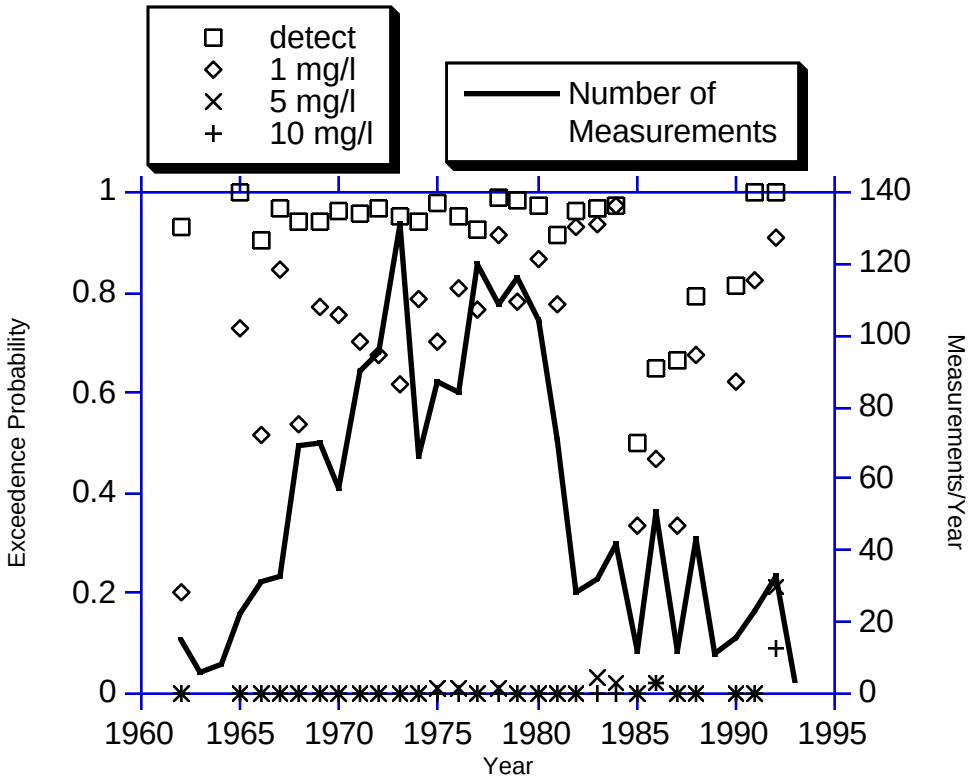


Figure 6.22 Variation of Exceedence Probabilities over Time in the Balcones Fault Zone of the Edwards Aquifer

6.2.3 Hueco-Mesilla Bolson Aquifer

Table 6.8 Nitrate Exceedences in the Hueco-Mesilla Bolson Aquifer (1908 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	1506	0.7893	0.7734	0.8046
1	869	0.4554	0.4365	0.4745
5	63	0.0330	0.0266	0.0406
10	18	0.0094	0.0061	0.0139

Because the extent of the Hueco-Mesilla Bolson Aquifer is small, the exceedence probabilities shown in Figure 6.23 have no discernible spatial pattern. As in the Edwards, detections of nitrate and exceedences of the 1 mg/l threshold are quite common, but measurements exceeding the 5 and 10 mg/l thresholds are rare. Figure 6.24 shows very little variation in exceedence probabilities with depth, the least in the five study aquifers.

Few nitrate measurements from the Hueco-Mesilla Bolson appear in the database prior to 1980, making trends through time difficult to detect. Figure6.25 might be interpreted to indicate increased exceedences of the 5 mg/l, but regression of the exceedence probabilities against time shows no statistically significant trends in exceedences of any of the threshold levels.

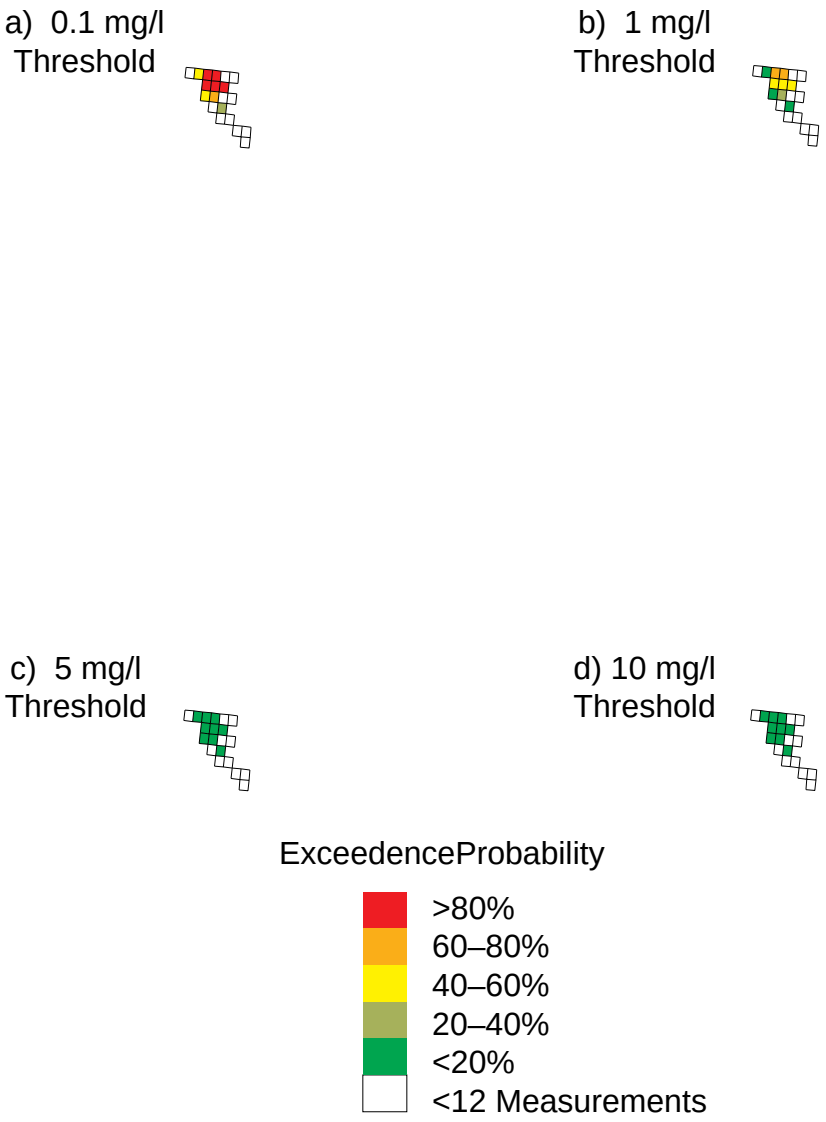


Figure 6.23 Estimated Nitrate Exceedence Probabilities by Quadrangle in the Hueco-Mesilla Bolson Aquifer

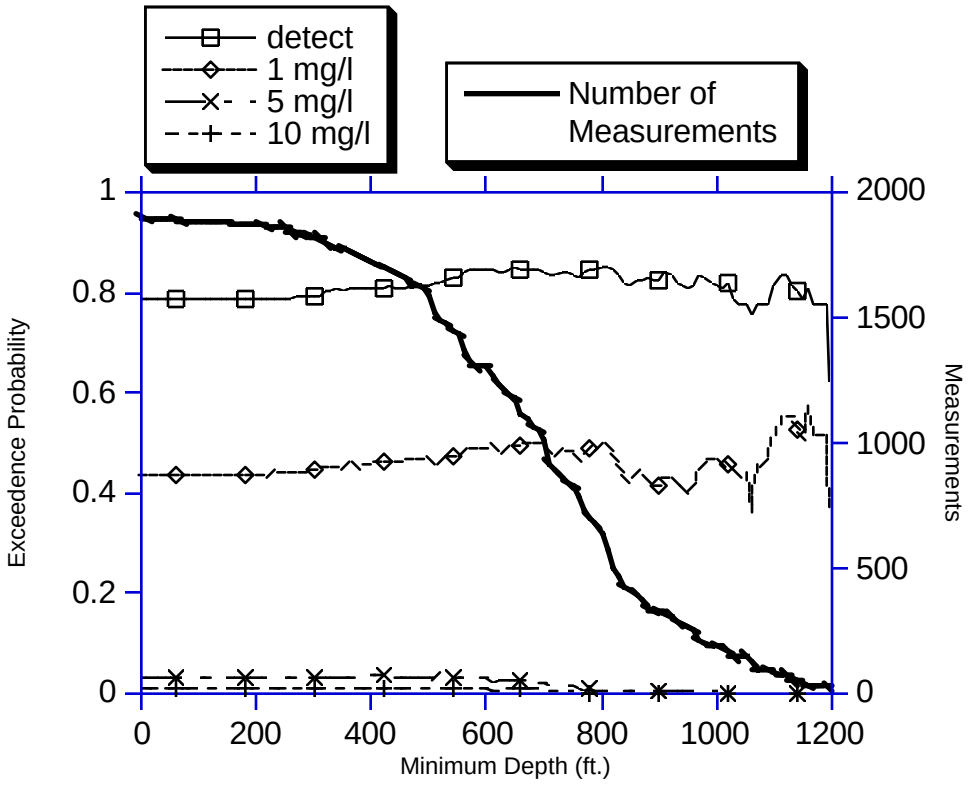


Figure 6.24 Variation of Exceedence Probabilities with Depth in the Hueco-Mesilla Bolson Aquifer

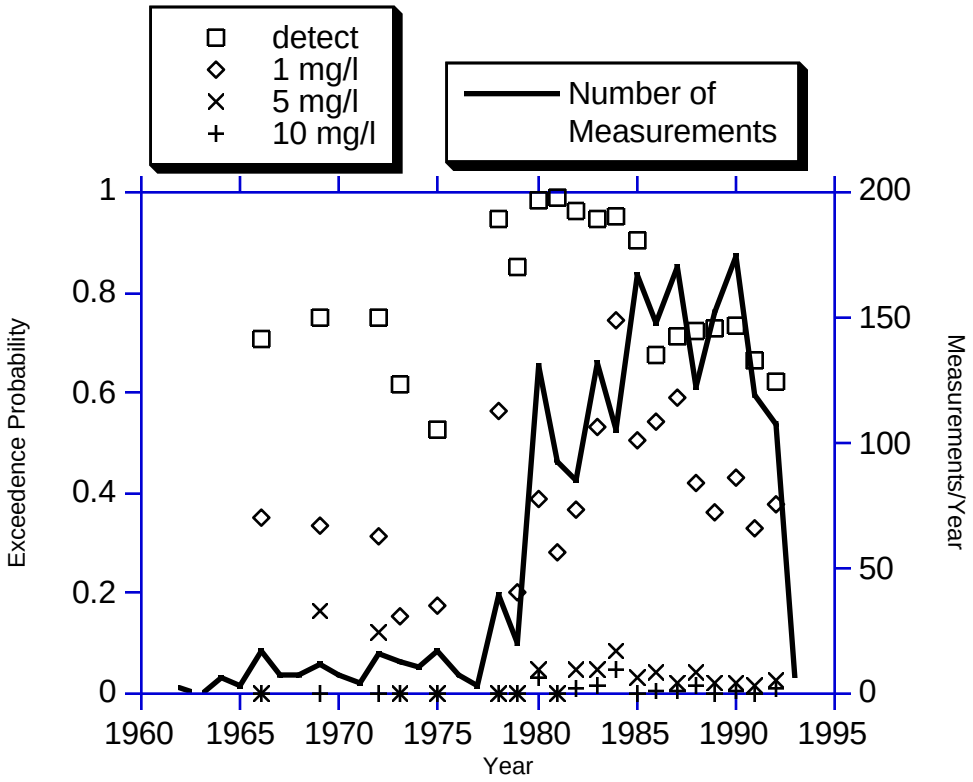


Figure 6.25 Variation of Exceedence Probabilities over Time in the Hueco-Mesilla Bolson Aquifer

6.2.4 Ogallala Aquifer

Table 6.9 Nitrate Exceedences in the Ogallala Aquifer (4430 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	4164	0.94	0.9337	0.9458
1	3235	0.7302	0.7191	0.7412
5	549	0.1239	0.1159	0.1323
10	219	0.049436	0.0441	0.0551

As in the Edwards and Hueco-Mesilla Bolson aquifers, nitrate measurements taken from the Ogallala Aquifer are very likely to exceed 1 mg/l, but much less likely to exceed 5 mg/l.

Of the five study aquifers, the Ogallala is the largest. The Texas portion of the aquifer provides water over most of the panhandle, and the aquifer extends northward through the mid-central U.S. In spite of its size, which would easily allow for trends or division into sub-regions, the map in Figure 6.26 shows variations in exceedence probabilities with no clear pattern visible. Detection rates vary, especially at the 1 mg/l threshold, but without exhibiting trend or regionalization.

Only one quad with twelve measurements or more shows a 5 mg/l exceedence probability greater than 80%. This quad was examined in more detail to see if the high rate was due to the influence of a single poorly constructed well. In fact, the 29 measurements taken in that quadrangle (number 2835, between 101_ 37' 30" and 101_ 45' west longitude and 32_ 22' 30" and 32_ 30' north latitude) come from 27 different wells. These are mostly shallow wells—none is

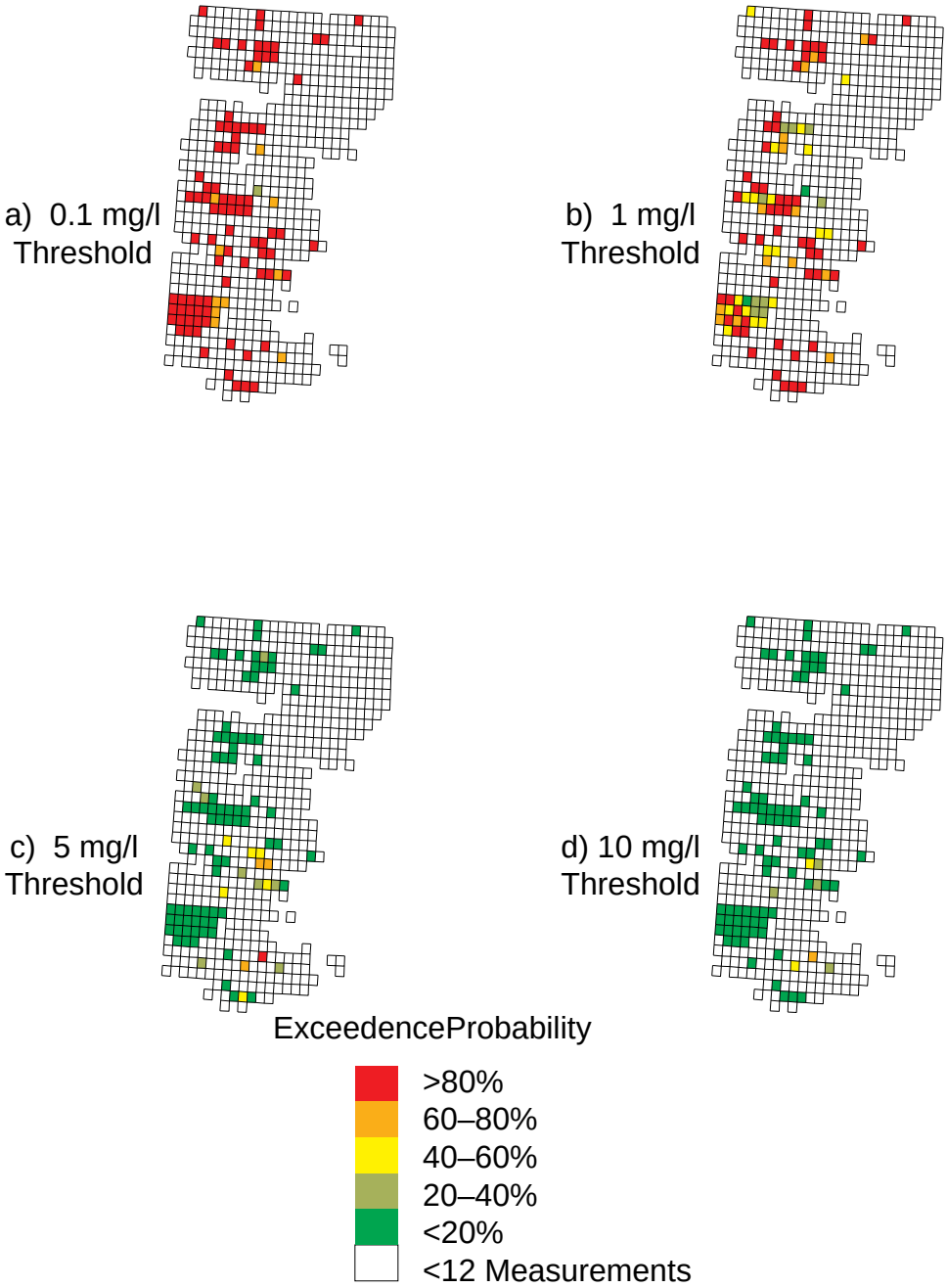


Figure 6.26 Estimated Nitrate Exceedence Probabilities by Quadrangle in the Ogallala Aquifer

deeper than 100 feet—providing water for domestic use. The region is in Martin and Howard Counties, northwest of Big Spring, in a lightly populated area containing a number of small oil fields. Apart from the shallowness of the wells, no obvious cause for the high incidence of exceedences suggests itself.

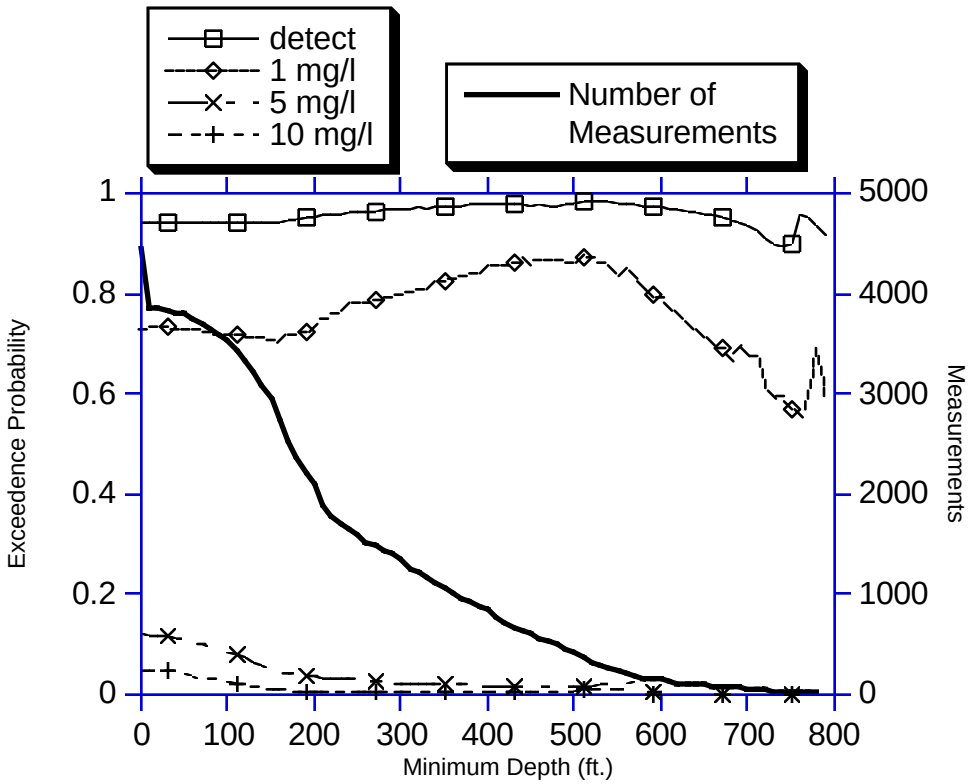


Figure 6.27 Variation of Exceedence Probabilities with Depth in the Ogallala Aquifer

Detection rates show little variation with depth in the Ogallala. Exceedences of the higher thresholds (5 and 10 mg/l) are noticeably lower in wells more than 200 feet deep, but no consistent trend with well depth is apparent in exceedences of the lower thresholds.

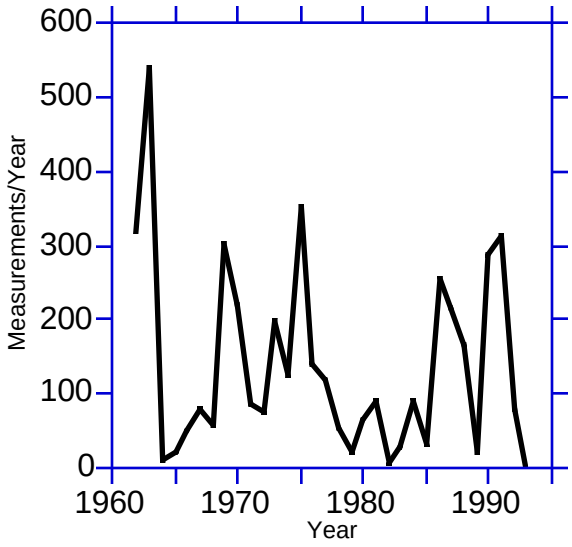


Figure 6.28 Measurements by Year in Ogallala Aquifer

Figure 6.28 shows the number of nitrate measurements per year listed in the data set for the Ogallala Aquifer. Figure 6.29 shows the variation of the four exceedence probabilities calculated for the same years. Statistically significant trends through time can be seen in three of the four exceedence probabilities. Regressions of exceedence probabilities at the detection level, 1 and 5 mg/l have t values greater than 2.0, indicating a 95% or higher probability of a consistent linear trend. Regression results are summarized in Table 6.4 (years with fewer than 12 listed measurements were excluded from the regressions). Regression lines are shown in Figure 6.29 for the three thresholds with significant trends. Probabilities of exceeding the detection limit and the 1 mg/l threshold have grown by about 0.3% per year over the period from 1962–1993, and the probability of exceeding the 5 mg/l threshold has grown by about 0.8% over the same period.

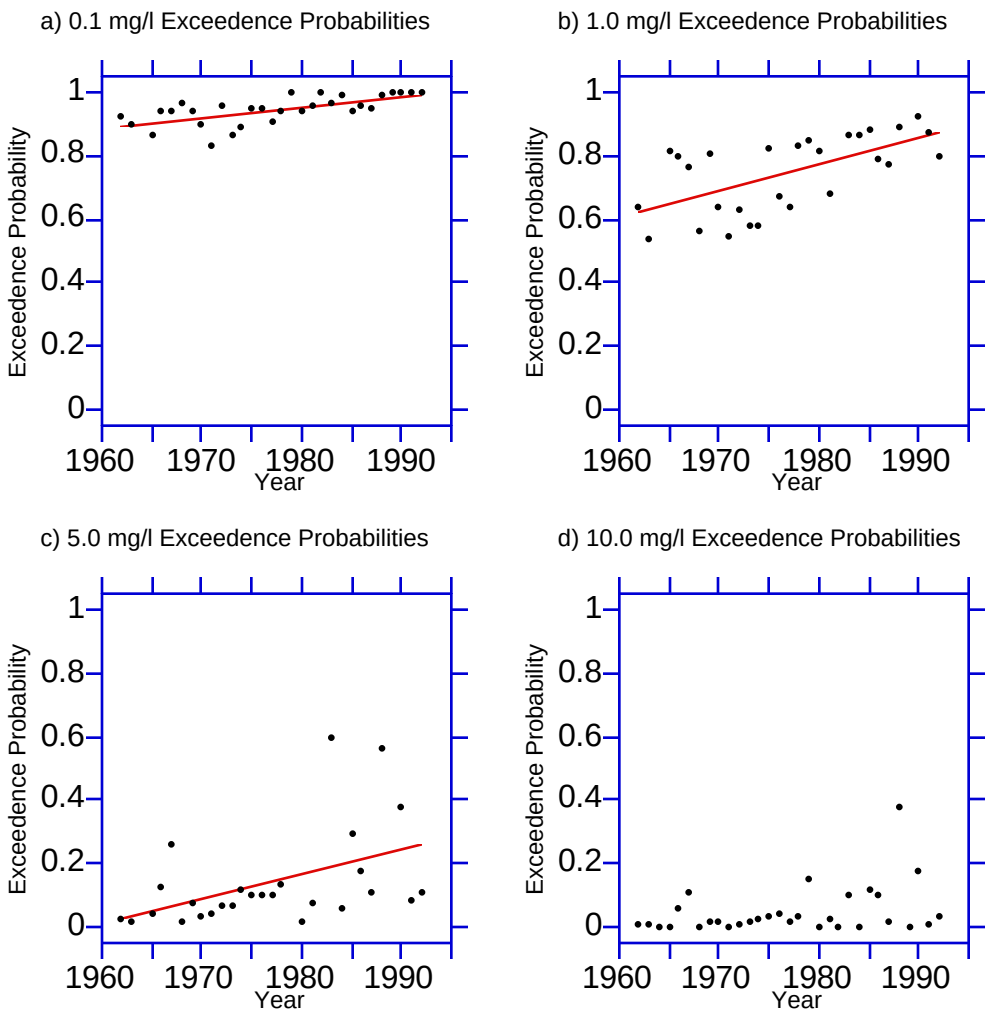


Figure 6.29 Variation of Exceedence Probabilities in the Ogallala Aquifer Over Time

Table 6.10 Regression Results For Threshold Exceedences through Time in the Ogallala Aquifer

Threshold	**Slope**	**t**
0.1 mg/l	0.003	4.94
1.0 mg/l	0.003	3.88
5.0 mg/l	0.008	2.67
10.0 mg/l	0.003	1.95

The increases in exceedence probabilities in the Ogallala point to the possibility of an accumulation of nitrates in the aquifer, which would almost certainly be due to human influences. Although the regressions for the State as a whole were barely statistically significant, the regressions in the Ogallala show an unmistakable trend through time. This may be the most convincing evidence of vulnerability revealed in this study.

6.2.5 Seymour Aquifer

Table 6.11 Nitrate Exceedences in the Seymour Aquifer(2526 Measurements)

Threshold (mg/l)	Exceedences	Exceedence Probability	Lower Bound	Upper Bound
0.1	2420	0.958	0.9508	0.9644
1	2368	0.9374	0.9289	0.9452
5	2073	0.8207	0.8076	0.8331
10	1435	0.568092	0.5517	0.5844

Of the five study aquifers, the Seymour is obviously the most highly contaminated by nitrates. Every quadrangle with twelve or more measurements from this aquifer has an estimated exceedence probability greater than 60% at the 1 mg/l threshold, and only two have exceedence probabilities below 80%. Figure 6.30 shows a slight tendency toward lower exceedence probabilities in the southern part of the aquifer at the higher thresholds, but given the small extent of the aquifer, it is unclear whether this is a significant trend.

Trends of exceedence probabilities with depth and with time in the aquifer are difficult to interpret. Figure 6.31 seems to indicate that shallower wells in the Seymour are less likely to have elevated nitrate levels than deeper wells, but given that in the study database only four wells tapping the Seymour are as deep as 150 feet, there is little room for variation with depth.

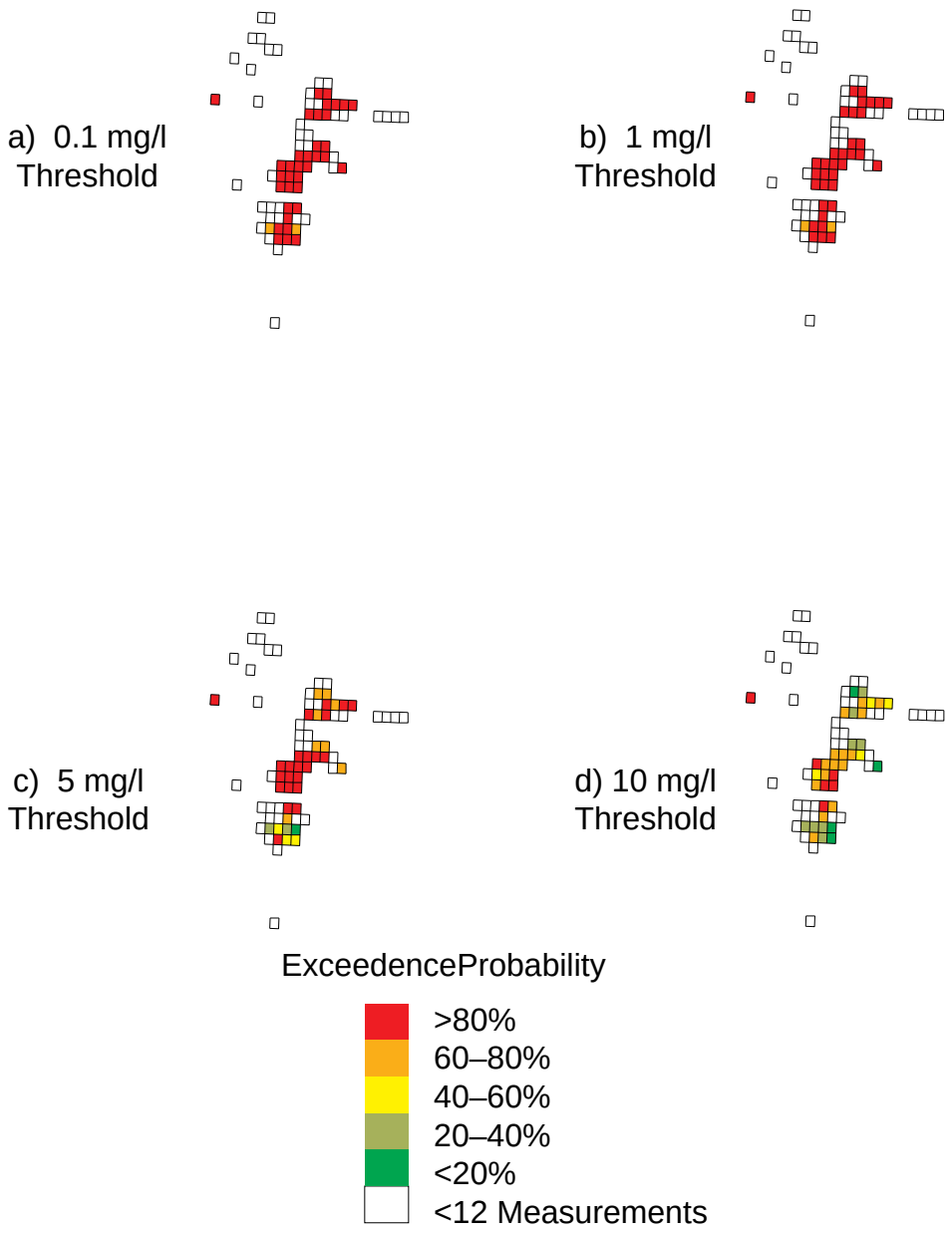


Figure 6.30 Estimated Nitrate Exceedence Probabilities by Quadrangle in the Seymour Aquifer

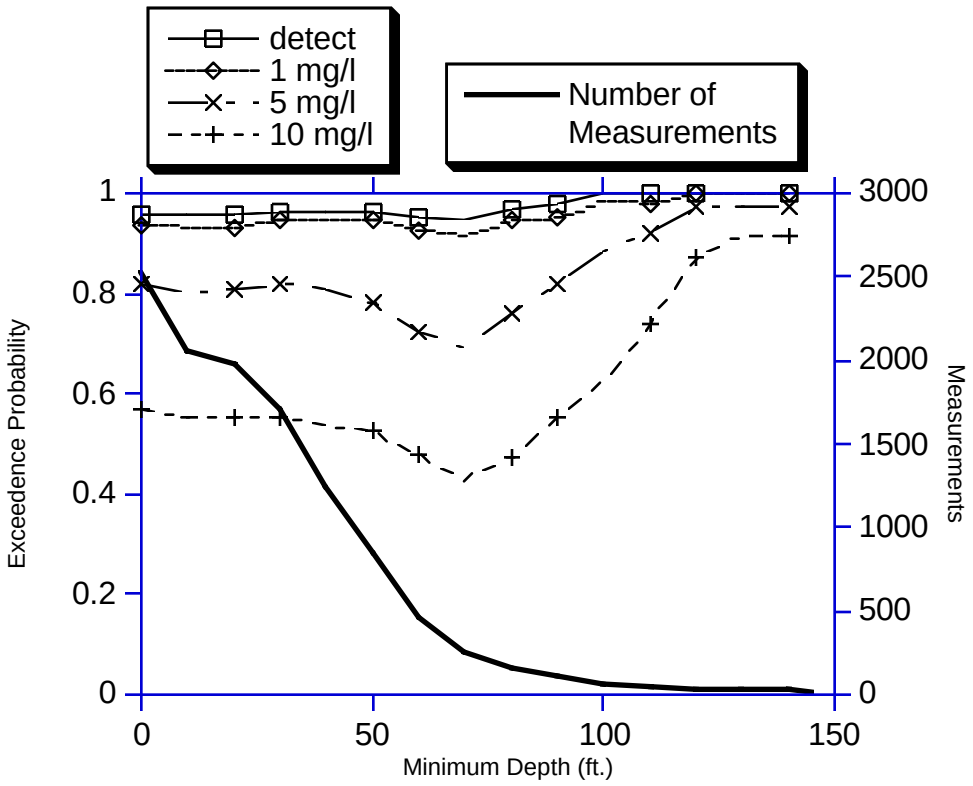


Figure 6.31 Variation of Exceedence Probabilities with Depth in the Seymour Aquifer

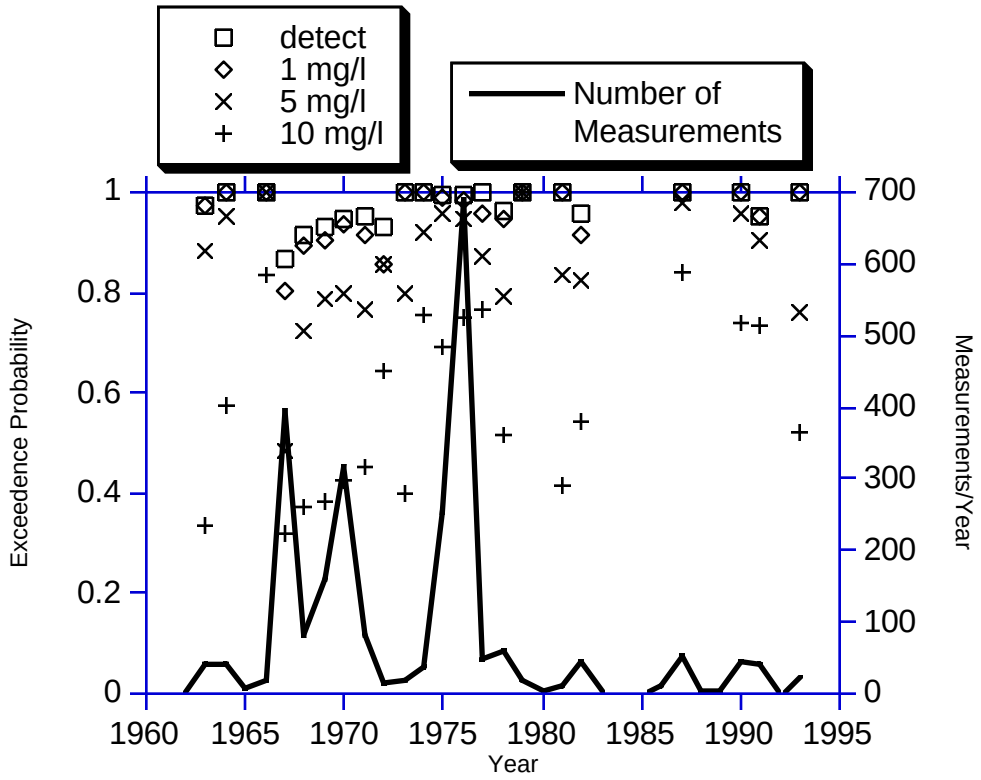


Figure 6.32 Variation of Exceedence Probabilities over Time in the Seymour Aquifer

The sampling history of the Seymour aquifer is very uneven. In only three years, (1967, 1970, and 1976) have more than 100 nitrate measurements from the Seymour been recorded and in 9 years fewer than 12 measurements were recorded; in 1984, none were recorded. Figure 6.32 may show a trend toward increasing likelihood of exceedences of the 10 mg/l threshold, but the t statistic of a regression on this probability against time is 1.95, indicating less than 95% probability that the trend is significant. Given the high incidence of exceedences at all levels, it is safe to say that the Seymour Aquifer is highly vulnerable to nitrate contamination.

6.2.6 Aquifer Summary

Figure 6.33 compares the estimated exceedence probabilities at the four thresholds for the statewide base data set and for each of the five study aquifers. The lines on the figures are provided as a visual aid and do not reflect any prediction for exceedence probabilities at intermediate thresholds. The figure reaffirms the trends discussed in the preceding sections. The Carrizo-Wilcox clearly has the lowest nitrate concentrations of the five aquifers and has lower exceedence probabilities at all thresholds than the state as a whole. The Seymour clearly has the highest concentrations, and higher exceedence probabilities at all thresholds than the state as a whole.

The Edwards (Balcones Fault Zone), the Hueco-Mesilla Bolson, and the Ogallala have intermediate values for exceedence probabilities. In these aquifers nitrate is more likely to be found at the 0.1 and 1.0 mg/l levels than in the state as a whole, but less likely to be found at the 5 and 10 mg/l levels than in the state as a whole. One possible explanation for this variation is that all three aquifers have porous compositions, which makes them very penetrable, and vulnerable to surface influences. At the same time their permeability leads to more mixing than in more tightly formed aquifers, and hence more dilution and fewer detections at high concentrations. The lack of strong trends with depth tends to confirm this possibility.

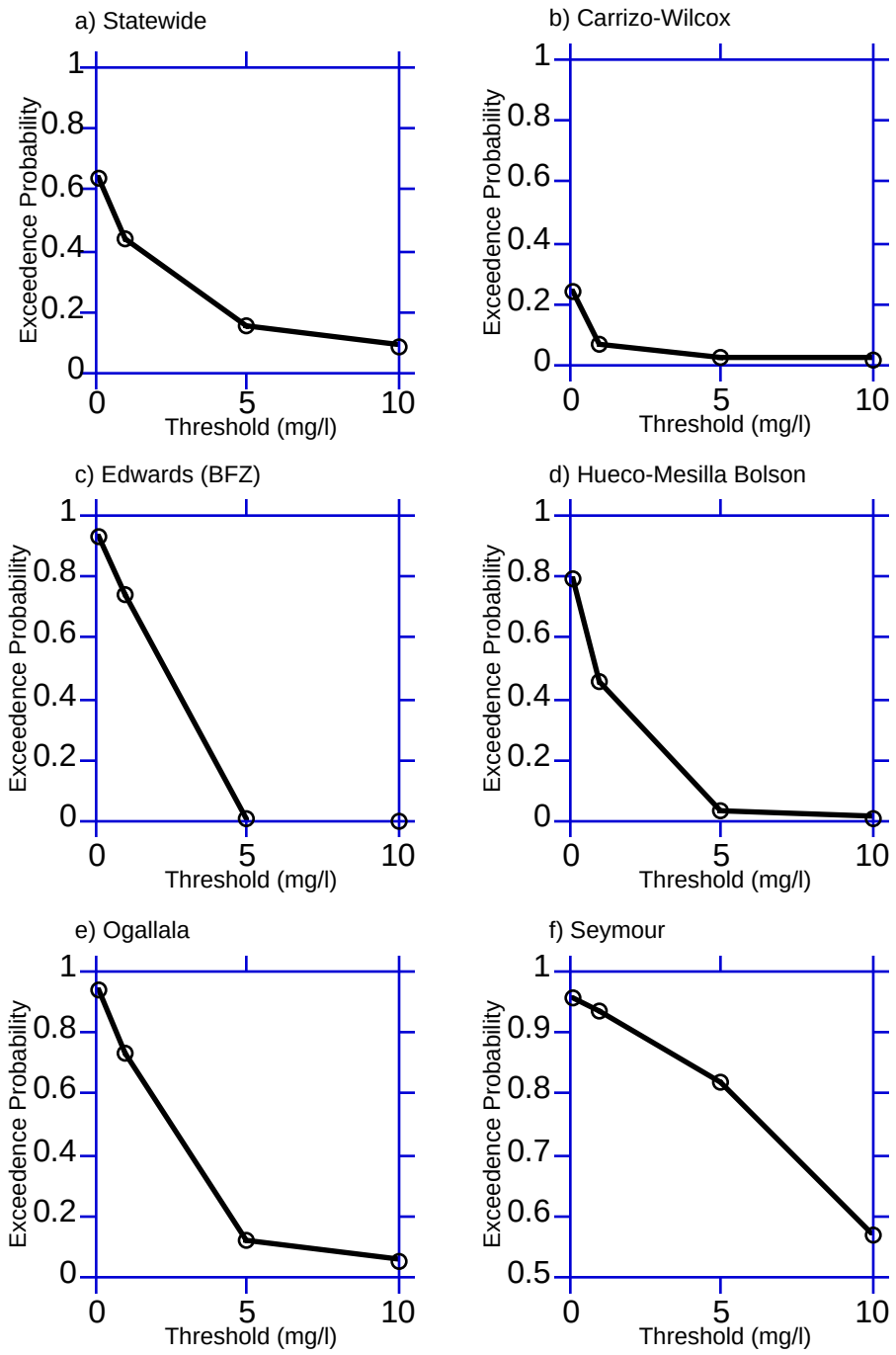
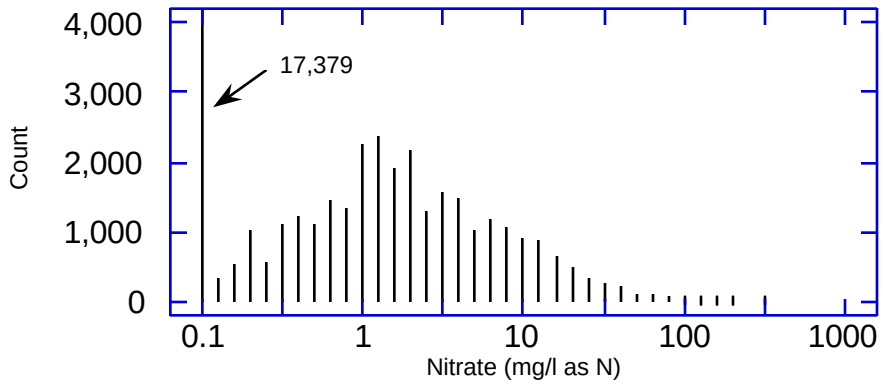


Figure 6.33 Comparison of Exceedence Probabilities Statewide and in Five Study Aquifers

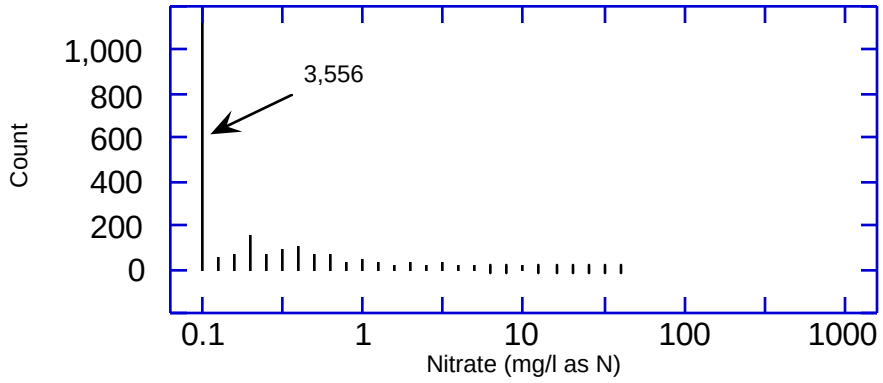
Figure 6.34 summarizes the nitrate measurements statewide and in the five aquifers as histograms. Each histogram approximates the shape of the probability distribution of nitrate concentrations in the corresponding population of water samples. Note that the concentrations are expressed as logarithms. The graphs in Figure 6.33 approximate the inverse of the cumulative probability of nitrate concentrations in the state and the aquifers. The graphs in Figure 6.34 approximate the probability densities of the state and the aquifers. The shapes of the distributions vary considerably from aquifer to aquifer. In general, the tails of the distributions (especially at the low end of the concentration range) are very long, as represented by the high numbers at the detection limit. The Seymour Aquifer comes closest to a lognormal distribution, but is very long in the tails at both ends.

Figure 6.35a summarizes the nitrate measurements in the five study aquifers in a different way, using boxplots. In a boxplot, the box contains the central 50% (between the 25th and 75th percentile) of the values in the plotted group, and the whiskers extend to the lowest and highest values within 1.5 times the width of the box. The Edwards and Seymour Aquifers show the least variation in nitrate concentrations, as illustrated by the narrowness of their boxes. Points farther from the boundaries of the box are plotted as circles or "outside" values (Helsel and Hirsch 1992) The Hueco-Mesilla Bolson has the smallest number of outside values. Note that since more than 75% of the nitrate measurements in the Carrizo-Wilcox (CZWX) are below the detection limit, the width of the box is zero, and there are no whiskers on its plot. As a result, every



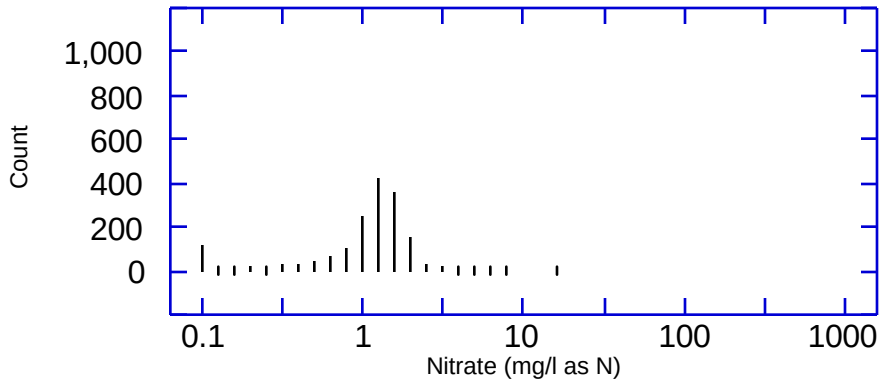
a

) Statewide Nitrate Concentration Histogram



b

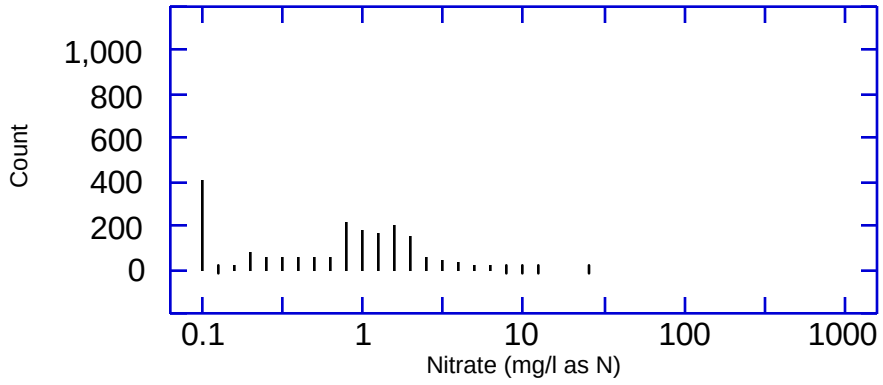
) Carrizo-Wilcox Aquifer Nitrate Concentration Histogram



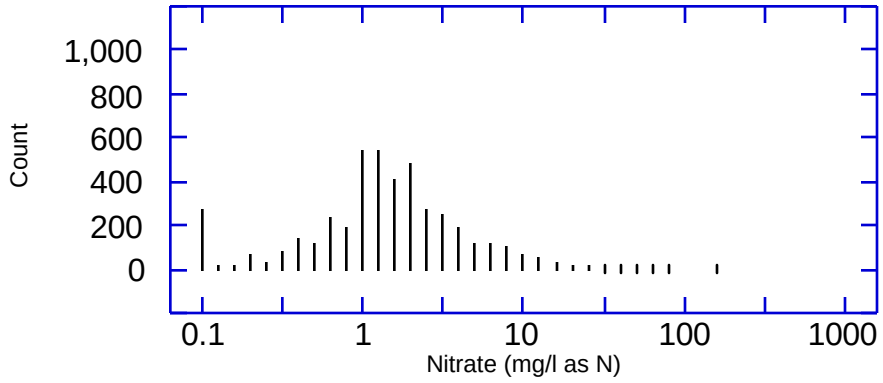
c

) Edwards Aquifer (Balcones Fault Zone) Nitrate Concentration Histogram

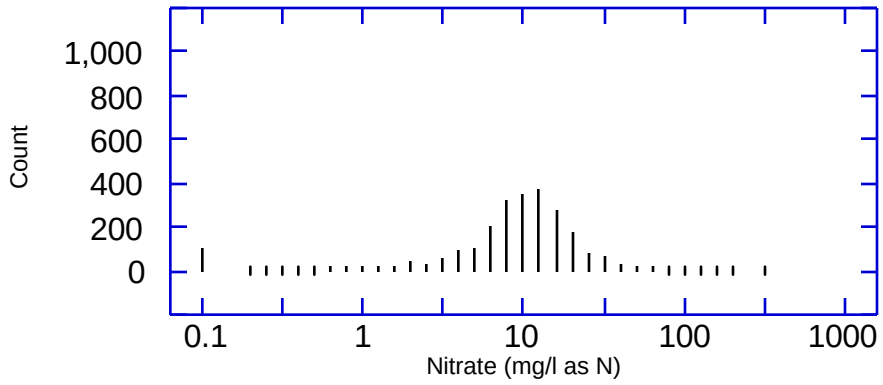
Figure 6.34 Histograms of Nitrate Concentrations in Texas and Five Study Aquifers



d) Hueco-Mesilla Bolson Aquifer Nitrate Concentration Histogram

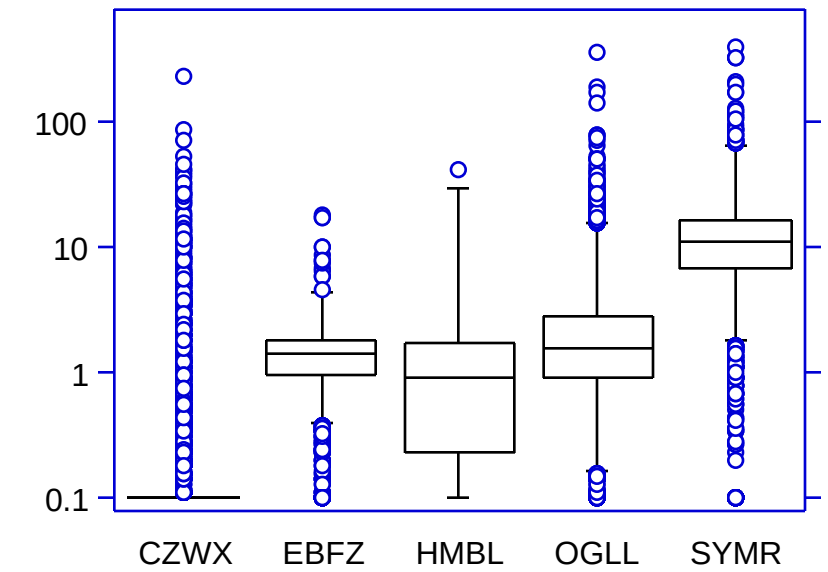


e) Ogallala Aquifer Nitrate Concentration Histogram

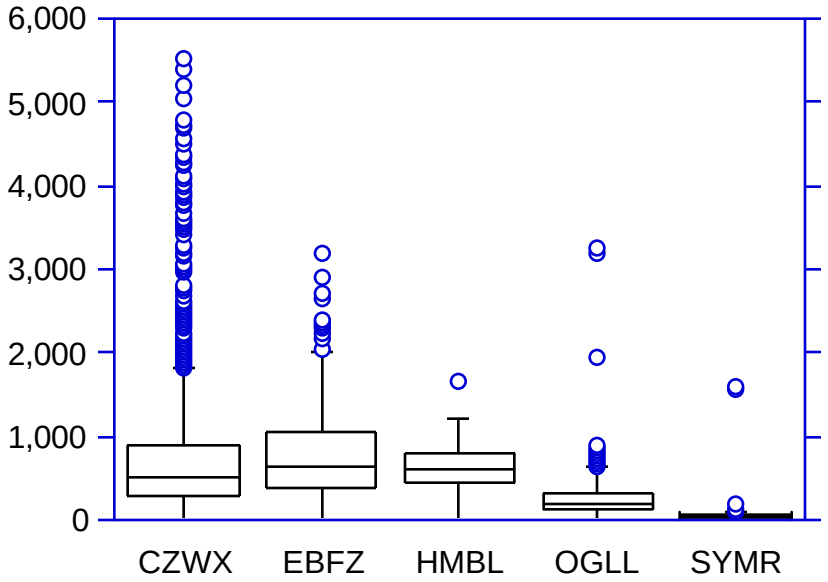


f) Seymour Aquifer Nitrate Concentration Histogram

Figure 6.34 (Continued) Histograms of Nitrate Concentrations in Texas and Five Study Aquifers



a) Nitrogen Concentration



b) Well Depths

Figure 6.35 Boxplots of Well Depths and Nitrate Concentrations in Five Study Aquifers

measurement over 0.1 mg/l is plotted as an outside value. This is another example of the effects of censored data on statistical representations.

In boxplots of normally distributed data, the boxes and whiskers are symmetrical, and roughly one point in 100 is an outside value. Although the plots for the Balcones Fault Zone of the Edwards Aquifer (EBFZ), the Ogallala Aquifer (OGLL), and the Seymour Aquifer (SYMR) are roughly symmetrical, which might indicate lognormal distribution (the plots are on a log scale), they have more outside points than a normal distribution, indicating a tail-heavy distribution. This is similar to the conclusion drawn earlier about the lack of fit of a lognormal distribution to data from single quadrangles.

Another comparison can be made from these data. A second boxplot, Figure6.35b, shows the distribution of well depths in the five aquifers. The two aquifers with the shallowest wells, the Seymour and the Ogallala, are also the ones with the highest nitrate concentrations. This observation tends to confirm the assumption that shallow groundwater is more vulnerable than shallower groundwater. However, the Edwards and Hueco-Mesilla Bolson Aquifers, which have higher detection rates than the Carrizo-Wilcox, also tend to have slightly deeper wells than the Carrizo-Wilcox. The relationship between depth and water quality remains somewhat ambiguous.

6.3 INDICATORS AND REGRESSION

In order to evaluate the predictive capacity of the potential indicator parameters, a series of stepwise multiple linear regression were performed. In each regression, an estimated exceedence probability was taken as the dependent variable, and average precipitation, average soil thickness, average soil organic matter content, and nitrogen fertilizer sale figures were taken as the independent

variables. Each of these variables was evaluated on 7.5' quadrangles across the State, as described in Chapter 5. The regressions were performed using STATGRAPHICS, a statistical and graphic data analysis package for personal computer.

In stepwise multiple linear regression, an independent variable is added to the model in the analysis if the additional information it provides is significant at a chosen confidence level. As the model is being constructed, partial F statistics are calculated for each variable not currently in the model, as though each were the next variable to be added. For a confidence level of 95%, a variable can be added to the regression if its F value is greater than 4.0. At the same time, partial F statistics are calculated for each variable already in the model, as though each were the last added to the model. If the F statistic for any variable in the model falls below the selected threshold, it is removed from the model. See Draper and Smith (1981) for a more complete discussion of this method. The F statistics for variables included and not included are combined in a single column in the following tables. The listed values are the partial F statistics for the final selected model for each exceedence probability.

In the first set of regressions, every 7.5' quadrangle with twelve or more measurements was included. These are the 1158 quadrangles that were mapped in Section 6.1. The regressions attempt to fit a model of the form

$$P_t = ß_0 + ß_1 T + ß_2 O + ß_3 R + ß_4 N \qquad (6\text{-}1)$$

where P_t is the exceedence probability in the quadrangle for threshold t, T is the soil thickness, O is the organic content of the soil, R is the average annual precipitation, and N is average annual nitrogen fertilizer sales. The results of the regressions are summarized in Table 6.12.

The purpose of the regression is primarily to identify those parameters with significant correlation to the exceedence probabilities, rather than to create a predictive model. To this end, all the variables are shown in the table, whether or not they were included in the final model.

The models resulting from the first two regressions listed in Table 6.12 include only the soil organic content and average precipitation as independent variables. In both cases, the precipitation is the more influential variable. Precipitation decreases markedly in Texas with distance from the coast, and nitrate detections increase from southeast to northwest. The regression reflects the parallels between these trends. That higher nitrate values are found where there is less precipitation runs somewhat counter to intuition, since higher recharge rates, which are driven by precipitation, are usually associated with greater vulnerability (as in DRASTIC). Possibly, higher precipitation leads to shorter residence time in the aquifers, and lower concentrations as a result. It is less surprising that higher soil organic content is associated with lower nitrate detections, since organic processes may tend to fix nitrate in the soil, preventing it from reaching groundwater.

Table 6.12 Regression Results for Quads with 12 or More Measurements

Threshold	**r^2**	**Indicators**	**Coefficient**	**Partial F**
Detection	0.414	Constant ($ß_0$)	1.201	--
		Thickness ($ß_1$)	--	0.215
		Organic ($ß_2$)	-0.0065	45.31
		Precip. ($ß_3$)	-0.0175	531.98
		Fertilizers ($ß_4$)	--	0.906
1 mg/l	0.398	Constant ($ß_0$)	1.046	--
		Thickness ($ß_1$)	--	3.26
		Organic ($ß_2$)	-0.00534	24.53
		Precip. ($ß_3$)	-0.019712	541.17
		Fertilizers ($ß_4$)	--	2.40
5 mg/l	0.154	Constant ($ß_0$)	0.1530	--
		Thickness ($ß_1$)	0.00369	39.42
		Organic ($ß_2$)	--	0.538
		Precip. ($ß_3$)	-0.00695	125.14
		Fertilizers ($ß_4$)	-0.01510	51.38
10 mg/l	0.079	Constant ($ß_0$)	0.047	--
		Thickness ($ß_1$)	0.00234	27.85
		Organic ($ß_2$)	--	0.105
		Precip. ($ß_3$)	-0.003184	46.29
		Fertilizers ($ß_4$)	0.009497	35.79

The regressions on detection and 1 mg/l exceedences have r^2 statistics of roughly 0.4, meaning that the regression equation predicts about 40% of the deviations from the mean value of the probabilities. That only two variables should predict this much of the variation is surprising. More surprising is the fact that regressions on precipitation alone yield r^2 values of 0.391 and 0.387 for detection and exceedence of 1 mg/l. The predictive capability of the first two regressions rests almost entirely on the inverse correlation between rainfall and nitrate exceedences.

The regressions on the exceedence probabilities of the higher concentrations have little meaning. Combining all available variables to produce an equation with little predictive power, they simply indicate a general lack of significant correlation between the dependent and the independent variables.

The second set of regressions, also fitting the model given in equation 6-1, was run on quadrangles containing twelve or more measurements from wells tapping the five study aquifers. These are the quads presented in the series of maps in Section 6.2. The results for these quads, summarized in Table 6.13, are very similar to those for the state as a whole. The organic material in the soil has

Table 6.13 Regression Results for Quads Associated with Study Aquifers

Threshold	**r^2**	**Indicators**	**Coefficient**	**Partial F**
Detection	0.409	Constant ($ß_0$)	1.289	--
		Thickness ($ß_1$)	--	0.668
		Organic ($ß_2$)	-0.0134	38.57
		Precip. ($ß_3$)	-0.0170	106.71
		Fertilizers ($ß_4$)	--	0.325
1 mg/l	0.387	Constant ($ß_0$)	1.130	--
		Thickness ($ß_1$)	--	0.122
		Organic ($ß_2$)	-0.0107	21.54
		Precip. ($ß_3$)	-0.0192	116.79
		Fertilizers ($ß_4$)	--	0.070
5 mg/l	0.116	Constant ($ß_0$)	0.202	--
		Thickness ($ß_1$)	0.0100	15.62
		Organic ($ß_2$)	-0.0070	11.14
		Precip. ($ß_3$)	-0.0053	11.59
		Fertilizers ($ß_4$)	-0.0254	20.01
10 mg/l	0.085	Constant ($ß_0$)	0.1600	--
		Thickness ($ß_1$)	0.0067	12.29
		Organic ($ß_2$)	-0.0039	6.15
		Precip. ($ß_3$)	-0.0031	7.03
		Fertilizers ($ß_4$)	-0.0192	19.99

more influence, but the equations contain the same independent variables and have roughly the same predictive power.

The third set of regressions, summarized in Table 6.14, is applied to the same quadrangles as the second, but now a series of dummy variables have been added, indicating the aquifer from which water was taken for the measurements and soil thickness and fertilizer sales have been dropped from consideration. The model to be fitted is thus

$$P_t = ß_0 + ß_1O + ß_2R + ß_3C + ß_4E + ß_5H + ß_6G + ß_7S$$

where O and R have the same meanings as in the equation 6-1, and C, E, H, G, and S are the dummy variables representing the Carrizo-Wilcox, Edwards(BFZ), Hueco-Mesilla Bolson, Ogallala, and Seymour Aquifers, respectively. If the measurements come from the Carrizo-Wilcox Aquifer, for example, the variable C is assigned a value of 1. G is used to represent the Ogallala aquifer because O is already used to represent soil organic content.

The results of the various regressions show that of the parameters tested, the most influential by far in determining the probability of nitrate detection or exceedence of threshold concentration is the aquifer from which the water is collected.

These regression results may be slightly misleading regarding the influence of geologic parameters relative to the other indicators. For example, although precipitation drops out of the regression when the dummy variables for the aquifers are included, this does not mean that it has no influence. The fact that the 1 mg/l exceedence probabilities in the Carrizo-Wilcox and Ogallala Aquifers differ by roughly 67% may be in part due to the difference in average rainfall over

Table 6.14 Regression Results for Quads Associated with Study Aquifers, Including Dummy Variables for Aquifers

Threshold	r^2	Indicators	Coefficient	Partial F
Detection	0.809	Constant ($ß_0$)	0.857	--
		Organic ($ß_1$)	--	0.200
		Precip. ($ß_2$)	0.00408	9.14
		CW ($ß_3$)	-0.748	711.40
		ED ($ß_4$)	--	2.118
		HM ($ß_5$)	-0.199	11.37
		OG ($ß_6$)	--	1.913
		SR ($ß_7$)	--	0.125
1 mg/l	0.787	Constant ($ß_0$)	0.736	--
		Organic ($ß_1$)	--	0.0002
		Precip. ($ß_2$)	--	0.879
		CW ($ß_3$)	-0.664	828.9
		ED ($ß_4$)	--	0.052
		HM ($ß_5$)	-0.373	34.33
		OG ($ß_6$)	--	0.052
		SR ($ß_7$)	0.187	29.21
5 mg/l	0.758	Constant ($ß_0$)	0.021	--
		Organic ($ß_1$)	--	0.566
		Precip. ($ß_2$)	--	0.0005
		CW ($ß_3$)	--	0.0004
		ED ($ß_4$)	--	0.0383
		HM ($ß_5$)	--	0.0578
		OG ($ß_6$)	0.094	28.1
		SR ($ß_7$)	0.779	939.9
10 mg/l	0.691	Constant ($ß_0$)	0.011	--
		Organic ($ß_1$)	--	0.004
		Precip. ($ß_2$)	--	0.0007
		CW ($ß_3$)	--	0.0414
		ED ($ß_4$)	--	0.0455
		HM ($ß_5$)	--	0.0013
		OG ($ß_6$)	0.031	4.35
		SR ($ß_7$)	0.545	653.4

in the parts of the state where they are located. Dummy variables for spatially distinct aquifers will subsume a great deal of spatially variable data.

A fourth set of regressions was run for the 1 mg/l threshold exceedence probability on quadrangles within single aquifers. Again, the model to be fit is given in equation 6-1. The results of the regressions are shown in Table 6.15. No model could be fit to the data from the Hueco-Mesilla Bolson Aquifer because the number of quadrangles in that aquifer is too small.

The results of the regressions show that the selected indicators have very little value within the aquifers. No significant correlations were found in the Edwards or Seymour Aquifers, and the regressions in the Carrizo-Wilcox and Ogallala Aquifers have little explanatory power, as indicated by their r^2 values.

The final conclusion to be drawn from the regressions is that a model of exceedence probabilities as good as any that can be drawn from the indicator data included in this study would apply average exceedence probabilities for each aquifer and ignore the other indicators.

Table 6.15 Regression Results for 1 mg/l threshold within Study Aquifers

Aquifer	**r^2**	**Indicators**	**Coefficient**	**Partial F**
Carrizo-Wilcox	0.041	Constant ($ß_0$)	-0.053	--
		Thickness ($ß_1$)	--	0.249
		Organic ($ß_2$)	-0.0060	6.97
		Precip. ($ß_3$)	--	0.719
		Fertilizers ($ß_4$)	--	0.004
Edwards (BFZ)	--	Constant ($ß_0$)	--	--
		Thickness ($ß_1$)	--	0.154
		Organic ($ß_2$)	--	0.223
		Precip. ($ß_3$)	--	0.092
		Fertilizers ($ß_4$)	--	0.250
Ogallala	0.0996	Constant ($ß_0$)	0.964	--
		Thickness ($ß_1$)	--	0.039
		Organic ($ß_2$)	-0.0068	4.84
		Precip. ($ß_3$)	--	0.027
		Fertilizers ($ß_4$)	-0.0201	10.39
Seymour	--	Constant ($ß_0$)	--	--
		Thickness ($ß_1$)	--	0.259
		Organic ($ß_2$)	--	0.187
		Precip. ($ß_3$)	--	0.248
		Fertilizers ($ß_4$)	--	0.264

6.4 COMPARISON WITH WATER UTILITIES DIVISION DATA

The objective of this section is to determine how well the data collected by the TWDB over a period of more than 30 years from wells constructed for many purposes predicts the likelihood of finding nitrate in samples collected in a much shorter period from wells used for public water supply. Nitrate measurements collected by the Water Utilities Division (WUD) of the Texas Natural Resource Conservation Commission as part of its Primary Drinking Water Standards enforcement effort are collected in a database maintained independently of the TWDB Groundwater Data System. Records of nitrate measurements collected between February 1993 and October 1994 were extracted from this database for comparison to the quadrangle exceedence probabilities estimated from the TWDB database.

Of 16,538 measurements recorded in the WUD database, 11,698 were collected from water systems using groundwater exclusively, and could be traced to well locations. 11,614 of these measurements could be identified with quadrangles with at least one measurement included in the analysis of the TWDB data, and 6,992 could be identified with one of the 3,554 quadrangles with 12 or more TWDB measurements (see Section 5.?).

Because the number of measurements in the WUD database is relatively small, only 132 quads have 12 or more measurement records in both databases, limiting the scope of quad-by-quad comparison of exceedence in the two databases. Figure 6.36 shows a scatter plot of this comparison for exceedences of the 0.1 mg/l threshold.

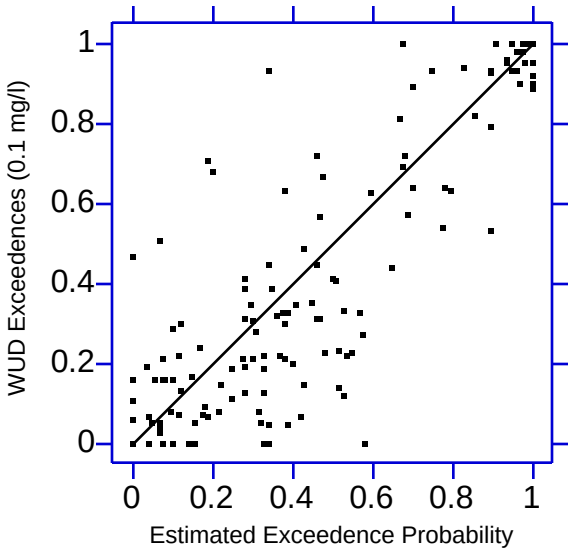


Figure 6.36 Quad-by-Quad Comparison of Estimated 0.1 mg/l Exceedence Probabilities with WUD Nitrate Measurements

To form a comparison based on all the WUD measurements, the data were aggregated by the estimated exceedence probability of the quadrangles in which the water samples were collected. The results of this comparison for the 0.1 mg/l threshold are shown in Figure 6.37. Figure 6.37a shows, for example, that of all the measurements in the WUD database collected from quads with an estimated 0.1 mg/l exceedence probability between 0.9 and 1.0, about 89% had concentrations above the threshold. The figure clearly shows a trend toward higher frequencies of nitrate detection in quads with higher estimated exceedence probabilities. The trend breaks down, however, in quadrangles with the lowest estimated exceedence probabilities. Figure 6.37b makes a similar comparison of aggregated measurements, limited to quads where the exceedence probability estimate is based on 12 or more measurements from the TWDB database. In this comparison, the agreement of estimated exceedence probabilities and exceedences

recorded in the WUD database improves, but the same break in the trend at low probabilities can be seen.

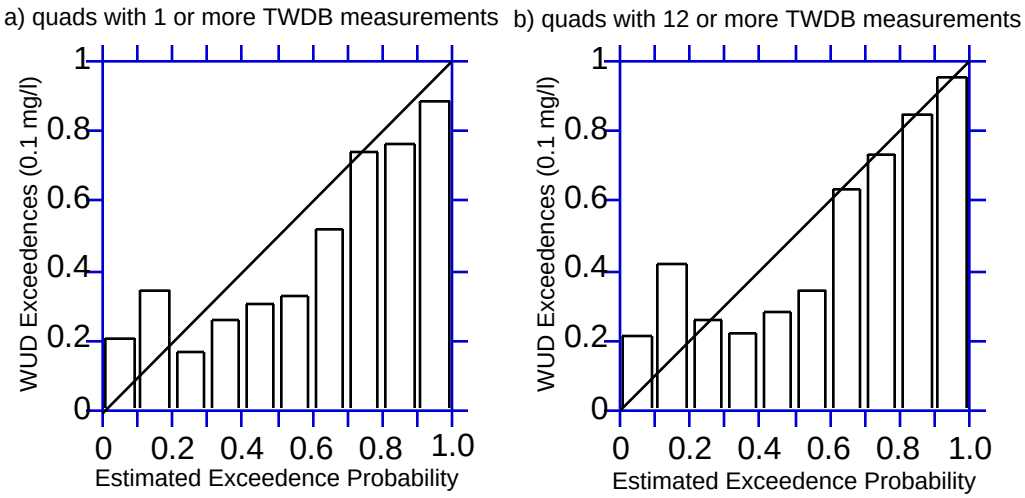


Figure 6.37 Aggregated Comparison of Estimated 0.1 mg/l Exceedence Probabilities with WUD Nitrate Measurements

Taken together, Figures 6.36 and 6.37 suggest that the TWDB data under-predict the WUD measurements about as often as they over-predict. In aggregate, the two data sets agree but there is often a considerable difference in the detection rates within a single quad. The same behavior can be seen in graphs of the same information for higher threshold levels, which are presented on the following pages in Figures 6.38 through 6.43

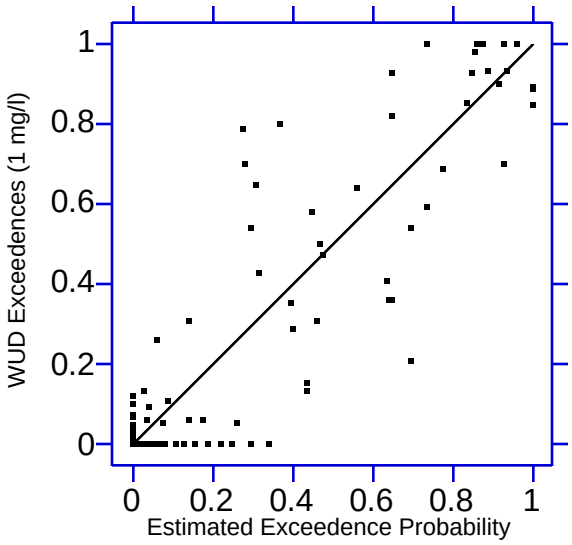


Figure 6.38 Quad-by-Quad Comparison of Estimated 1 mg/l Exceedence Probabilities with WUD Nitrate Measurements

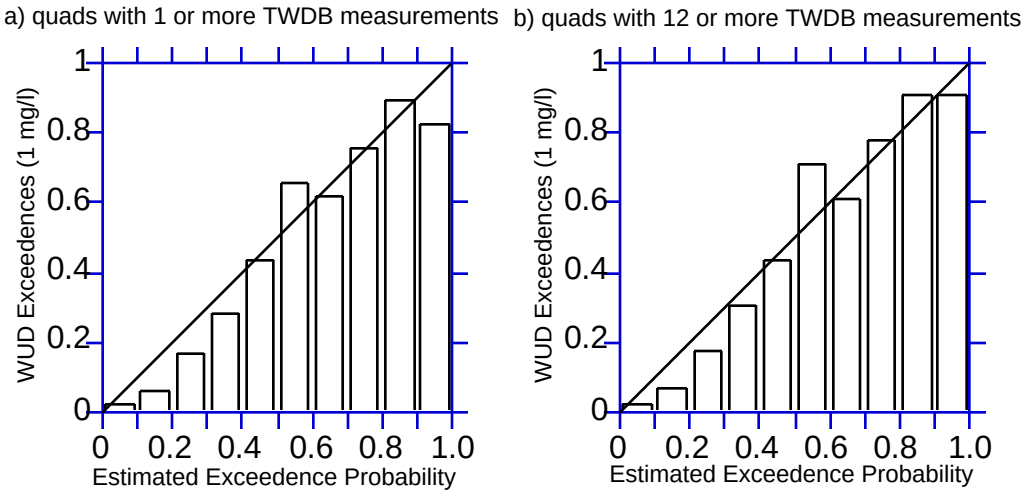


Figure 6.39 Aggregated Comparison of Estimated 1 mg/l Exceedence Probabilities with WUD Nitrate Measurements

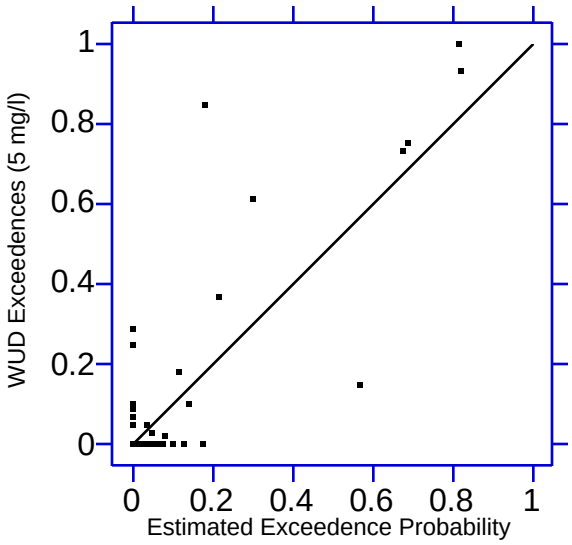


Figure 6.40 Quad-by-Quad Comparison of Estimated 5 mg/l Exceedence Probabilities with WUD Nitrate Measurements

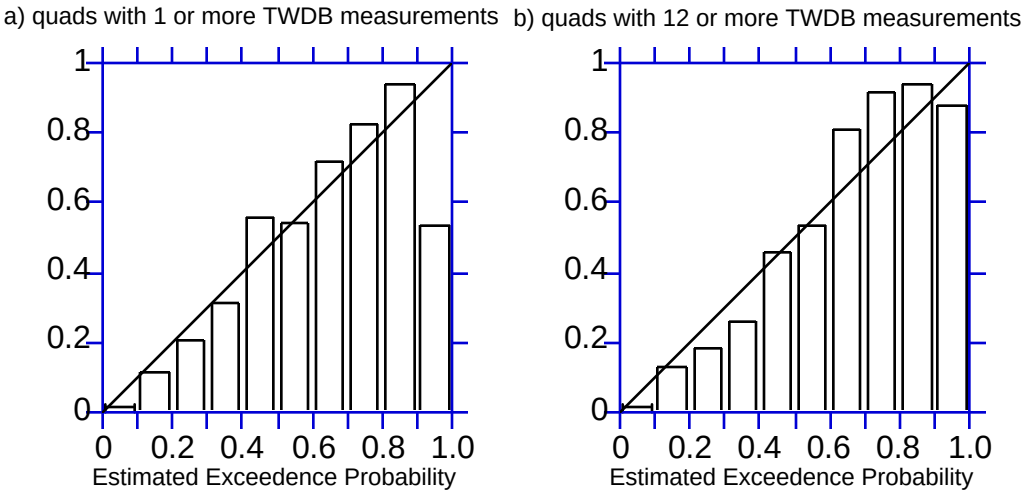


Figure 6.41 Aggregated Comparison of Estimated 5 mg/l Exceedence Probabilities with WUD Nitrate Measurements

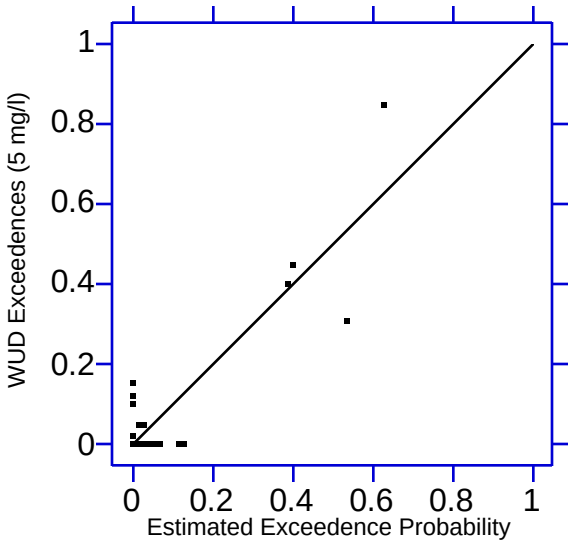


Figure 6.42 Quad-by-Quad Comparison of Estimated 10 mg/l Exceedence Probabilities with WUD Nitrate Measurements

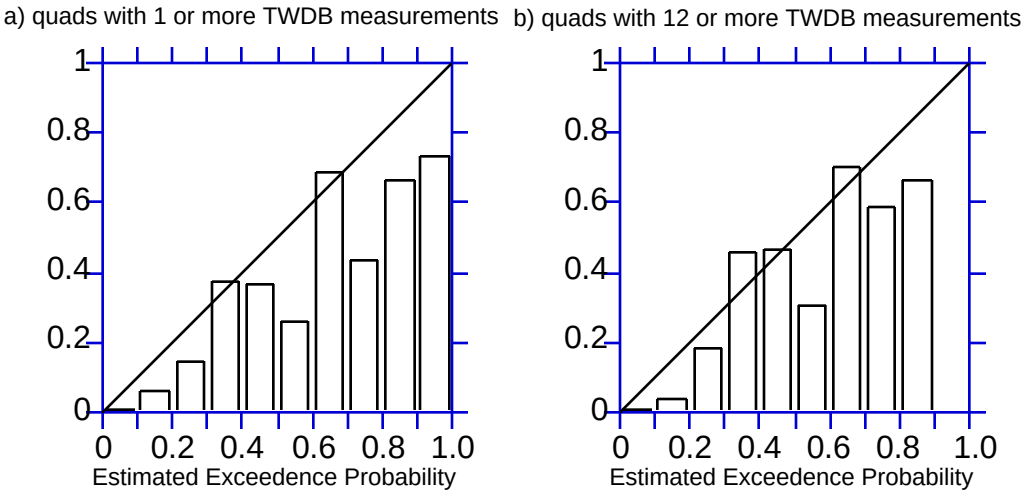


Figure 6.43 Aggregated Comparison of Estimated 10 mg/l Exceedence Probabilities with WUD Nitrate Measurements

One possible interpretation of the higher-than-predicted 0.1 mg/l exceedence rates quads with low exceedence probabilities is that there has been a gradual buildup of nitrate in groundwater systems, and that regions that in were in

equilibrium at nitrate concentrations below 0.1 mg/l, are now loaded above that level. If this were the case, however, one would expect to see this pattern repeated at the higher exceedence thresholds, especially at the 1 mg/l level, where the TWDB data shows an increase in the statewide detection rate over time.

The higher-than-predicted 5 mg/l exceedence rates and lower-than-expected 10 mg/l exceedence rates in quads with high exceedence probabilities may be due in part to the influence of drinking water regulations. More frequent sampling is required in systems where the 5 mg/l threshold is exceeded, and water sources with nitrate concentrations in excess to 10 mg/l violate the MCL and are likely to be removed from water supply systems. These factors could lead to over-sampling of water with nitrate above 5 mg/l and under-sampling of water with nitrate below 10 mg/l. No attempt was made to compensate for either of these potential biases.

6.5 NITRATE AND HERBICIDES IN MIDWEST DATA SET

Although nitrate is the only constituent studied in this work, the initial objective was to devise a system for predicting the likelihood of finding man-made agricultural chemicals in groundwater. This section addresses the question of how the occurrence of elevated levels of nitrate relates to the presence of agricultural chemicals. Because of the scarcity of herbicide data from Texas, the comparison is made using data from the herbicide and nitrate reconnaissance carried out in near-surface aquifers of the mid-continental U.S. by Kolpin, Burkart and Thurman (1992).

The report lists results of chemical analyses of 599 water samples collected from 303 wells in the mid-continental states of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota and Wisconsin. Concentrations are listed for a variety of nutrients, herbicides and

herbicide metabolites. Of interest to this work are the measurements of nitrate, seven herbicides (alachlor, atrazine, cyanazine, metolachlor, metribuzin, prometon, and simazine) and two dealkylated atrazine metabolites (deethylatrazine and deisopropylatrazine). The detection limit for nitrate is 0.05 mg/l. The detection limit for the herbicides and atrazine metabolites is 0.05 µg/l.

In general, detectable levels of herbicides are more likely to be found in water samples with elevated nitrate levels. Of 170 samples with nitrate concentrations above 3 mg/l, 84 (49%) had detectable levels of at least one herbicide or metabolite. In contrast, of 429 samples with nitrate concentrations less than or equal to 3 mg/l, 70 (16%) had detectable levels of at least one herbicide or metabolite.

However, it is also true that of 246 samples with no detectable nitrate, 22 (9%) had detectable levels of at least one herbicide or metabolite. The absence of nitrate in a well, apparently, cannot be considered a guarantee that the well is also free of herbicides—a less specific approach to the use nitrate as an indicator of herbicides is called for.

Such an approach might be based on the idea that the same conditions that lead to a high incidence of elevated nitrate levels would also lead to a high incidence of herbicide detections. A simple comparison of nitrate and herbicide concentrations in samples grouped by two geologic parameters tends to confirm this idea.

Burkart and Kolpin (1993a), in their analysis of the midwest data, found that nitrate and herbicide concentrations were higher in samples collected from unconsolidated aquifers than in samples collected from bedrock aquifers. They also found that nitrate and herbicide concentrations tend to decrease as aquifer

depth increases. (Aquifer depth is defined as the vertical distance from the land surface to the top of the aquifer material, regardless of whether the material is saturated or not.)

The matrix presented in Figure6.44 shows the number of water quality samples collected from wells falling into each of four categories based on aquifer class (bedrock or unconsolidated) and aquifer depth. The matrix also shows the number of nitrate measurements in excess of two threshold values, and number of herbicide detections in samples from the four categories. The rates of exceedence and rank of the four categories based on those rates are summarized in Table 6.16 .

	Depth ≤ 30 feet	Depth > 30 feet
Bedrock	Measurements: 113 Nitrate > 1 mg/l: 42 Nitrate > 3 mg/l: 30 Herb. Detections: 25	Measurements: 95 Nitrate > 1 mg/l: 13 Nitrate > 3 mg/l: 6 Herb. Detections: 11
Unconsolidated	Measurements: 335 Nitrate > 1 mg/l: 164 Nitrate > 3 mg/l: 120 Herb. Detections: 104	Measurements: 56 Nitrate > 1 mg/l: 26 Nitrate > 3 mg/l: 14 Herb. Detections: 14

Figure 6.44 Herbicide and Nitrate Measurements Grouped by Geologic Parameters

Table 6.16 Aquifer Categories Ranked by Nitrate and Herbicide Detection Rates

Aquifer Category	Nitrate Conc. > 1 mg/l rate	rank	Nitrate Conc. > 3 mg/l rate	rank	Herbicide Detections rate	rank
Deep Bedrock	14%	4	6%	4	12%	4
Shallow Bedrock	37%	3	26%	2	22%	3
Deep Unconsolidated	46%	2	25%	3	25%	2
Shallow Unconsolidated	49%	1	36%	1	31%	1

The results of this simple comparison are consistent with the hypothesis that conditions leading to increased vulnerability to nitrate contamination, as evidenced by high rates of elevated nitrate concentration, also lead to increased vulnerability to herbicides. This observation holds whether the threshold for elevated nitrate is set at 1 mg/l, as in this study, or at 3 mg/l, as Madison and Brunett (1985) suggest.

Although this comparison of nitrate and herbicide detections is far from conclusive, it suggests that an analysis of the occurrence of a widely measured constituent like nitrate can be used to gain insight into the occurrence of less commonly measured constituents like herbicides.

6.6 SUMMARY

The contents of this chapter have demonstrated how groundwater quality data can be regionalized with a GIS and a database management system, how that regionalized data can be analyzed statistically to classify those regions according to estimated probability of detecting excess nitrate, and how other parameters associated with those regions can be compared with the regional exceedence probabilities to form a predictive model. In addition, the regional exceedence probabilities were compared with an independent data set to test their predictive accuracy, and a simple analysis showed a possible connection between nitrate detections and vulnerability to herbicide contamination.

Sections 6.1 and 6.2 demonstrate the partitioning of the subsurface into two types of regions: the two-dimensional grid of 7.5' quadrangles and the geologic

regions of the five study aquifers. Water quality measurements are grouped by their association with these regions, and estimates of the probability that excess concentrations of nitrate will be found in the regions are calculated from those groups of measurements. The probability estimates are then used to identify the regions as more or less vulnerable to contamination by nitrate.

Section 6.3 presented the results of an attempt to generalize the results of the quadrangle exceedence probabilities by relating them to indicator variables evaluated on the same quadrangles. The regression results showed significant predictive potential only for average annual precipitation, which was inversely related to the probability of finding high nitrate concentrations, and with association of water quality measurements with specific aquifers. The only parameter associated with a source of nitrate, nitrogen fertilizer sales by county, was found to have no significant value as an indicator of nitrate exceedence probabilities.

In both Sections 6.1 and 6.2, an effort was made to identify the degree to which variations in depth and time, which cannot easily be represented in the two-dimensional domain of a GIS, influence the likelihood of finding nitrate at elevated concentrations.

Section 6.4 compared independent water samples with the nitrate exceedence probabilities presented in section 6.1. While quadrangles with higher predicted exceedences did, in aggregate, have higher frequencies of nitrate detection, there was considerable variation in individual quadrangles between predicted exceedence probabilities and frequencies of exceedence in the independent data set.

Section 6.5 shows by a simple analysis of data from the mid-continental U.S. that regions identified as vulnerable to nitrate contamination may also be vulnerable to contamination by man-made herbicides.

Chapter 7: Conclusions

This chapter summarizes the conclusions of this work, and restates the major results presented in the preceding chapters. The chapter is divided into three sections: a summary of the vulnerability assessment method, a discussion of the results the method's application to nitrate measurements in Texas, and recommendations for the future use of the method and its results.

7.1 VULNERABILITY ASSESSMENT METHOD SUMMARY

The primary result of this work is the development of a generally applicable method for assessing the vulnerability of groundwater supplies, using a geographic information system and a database of historic water quality measurements. The method is summarized by the six steps listed at the end of Chapter 2. These steps are recapped here in the specific form they were used in this work, followed by comments on their application.

7.1.1 Method Summary

1. *Select a constituent or set of constituents whose presence indicates the degree of vulnerability of a groundwater source.* The selected constituent for this study is nitrate.
2. *Identify a set of distinct mappable regions of the surface or subsurface.* Texas was divided into 7.5' quadrangles for mapping. Five aquifers were selected as an alternative set of mapping units. In the quadrangles , all measurements, regardless of well depth or aquifer association, were

grouped together for analysis. In the aquifers only measurements associated with particular geologic formations were grouped.

3. *Assemble a body of measurements of the constituent identified in step 1 that can be linked with the regions identified in step 2.* Nitrate measurement records were retrieved from the Texas Water Development Board (TWDB) Ground-Water Data System for the years 1962–1993. These measurements can be linked to quadrangles by the location of the wells from which water samples were collected for analysis. They can also be linked to aquifers through the TWDB well description database.

4. *Calculate descriptive statistics for the body of measurements linked with each region.* Exceedence probabilities for four nitrate concentration thresholds (0.1, 1.0, 5.0, and 10.0 mg/l nitrate as nitrogen) were calculated. Two methods were tried for estimation: non-parametric calculation of threshold exceedence probabilities for quadrangles based on the model of water sampling as a Bernoulli process, and fitting of all measurements in a quadrangle to a lognormal probability distribution. A minimum of twelve measurements was required for the exceedence probability estimates to be included in maps.

5. *Map the variation of the descriptive statistics from region to region.* The estimated exceedence probability for each threshold were divided into five ranges (<20%, 20–40%, 40–60%, 60–80%, >80%) and quadrangles with twelve or more measurements were color-coded according to this division.

6. *Relate the variation of the descriptive statistics to the variation of indicator parameters by forming a mathematical expression that mimics the relationship between the descriptive statistics and indicator values mapped over the same set of regions.* Stepwise multiple regression was used to form a linear expression relating quadrangle-averaged estimates of precipitation, soil thickness, soil organic content, and nitrogen fertilizer sales, to the exceedence probability estimates.

7.1.2 Comments

Selection of Constituents. In this work, nitrate was selected as the constituent to act as a surrogate for vulnerability. It was chosen because a large body of nitrate measurements is available in Texas, making statistical descriptions of its occurrence feasible, and because nitrate is commonly associated with agricultural sources, making it a potential surrogate for herbicides and other agricultural chemicals. Nitrate has many sources and is found in groundwater throughout Texas, which makes it suitable as an indicator of groundwater vulnerability, since its occurrence is not limited to regions where human activity generates concentrated sources. The drawbacks of nitrate as an indicator of vulnerability include the fact that it often occurs naturally, making it difficult to attribute high concentrations unambiguously to human influences. A similar study carried out using a constituent or constituents with no natural sources might present a clearer picture of vulnerability to human influences.

Selection of Study Regions. The selection of mappable regions should result in spatially compact regions with uniform properties. Since statistical descriptions

of the measurements in the regions lump together all the measurements from the regions, it is important the regions chosen can be adequately described by a few numbers. In effect, mathematical methods used to form the statistical descriptions of the regions assume that the regions are homogeneous, and the regions should be chosen in a way that does not violate that assumption. Neither the 7.5' quadrangles nor the aquifers used as study regions in this work fit these requirements exactly. Some of the quadrangles contain wells (and thus measurements) from several geologic formations, forming a heterogeneous population poorly suited to statistical description. Similarly, the aquifers are not spatially compact, which reduces the homogeneity of the populations of measurements they contain. In spite of these shortcomings, clear trends in vulnerability to nitrate are found between regions mapped in both sets of study regions. The division of the aquifers into quadrangles for the maps presented in Section 6.2 comes close to meeting the requirements of compactness and homogeneity, but at a cost of reducing the number of measurements in each cell.

Selection of the Database. The primary requirements for the database used to form statistical descriptions of the regions are that the data it contains should be of reliable and uniform quality and that the measurements be sufficiently plentiful to support statistical analysis. The TWDB database is certainly plentiful, although there are reasons—described in Section 3.1—to suspect some unevenness in the quality of the data it contains. The data in the Water Utilities Division data set have been subject to more rigorous and uniform quality control imposed by the provisions of the Safe Drinking Water Act, but are not yet

plentiful enough to support a study of this type, except as a test of predictions made from a larger data set.

Statistical Description of Regions. The exceedence probabilities calculated in this work are an attempt to describe the probability that threshold values of nitrate concentrations will be exceeded in the study regions. This approach to statistical description was chosen over the more common measures of central tendency and spread (like mean or median and standard deviation or inter-quartile range) because it more directly addresses the nature of regulations based on threshold concentrations such as detection limits and maximum contaminant levels (MCLs). A region is more clearly at risk if its MCL exceedence probability is high than if its average concentration is high.

Two statistical approaches to calculating exceedence probabilities were considered. One approach estimated probabilities of exceeding threshold concentrations by counting the number of exceedences in the database records associated with the regions and treating this as the result of a Bernoulli process, the other approach fits the data from 7.5' quadrangles to a lognormal distribution function. The results of the study indicate that there are few advantages to the lognormal-fit method. As graphical comparison of the two estimation methods (Figure 6.10) shows, forcing data to fit a particular distributional form incorrectly evaluates the exceedence probabilities in regions where that distribution does not fit well. A single distributional form simply lacks the flexibility to capture the range of variation in exceedence probability over a large and heterogeneous area like Texas. The computations required to fit the lognormal model are more

complex than those for the Bernoulli-process method, and the reduction of the distribution to two parameters offers little advantage. The number of measurements and exceedences of any threshold concentration in a region can be found easily through the use of a database management system and the estimation of the exceedence probability from those numbers is a simple process of division. Also, meaningful confidence intervals can be estimated for the Bernoulli-process method, and not for the lognormal-fit method.

Mapping Results. Mapping of the results of the statistical analysis makes spatial patterns in detection of nitrate very evident. This adds considerably to the value of the database for understanding variations in water quality through the State. Maps and statistics form a complementary description of the database. Summary statistics of exceedence patterns reduce a large quantity of data to a smaller, more easily interpretable set of numbers. As with the maps, the ease of interpretation comes at the expense of a loss of detail. The parallel between the maps and statistics can be extended further by analogy: Summary statistics reduce large amounts of data to a few meaningful numbers, and maps reduce large amounts of data to a few meaningful images. Proper interpretation of either maps or statistics requires an understanding of both the physical processes under study, and the mathematical or cartographic processes that produce the summary numbers and images.

Just as it is important to understand the limitations of statistical summary, which tends to obscure heterogeneity in the data, it is likewise important to understand the limitations of the images presented in the maps. Because a set of

discrete colors was used for identifying the quadrangles' exceedence probabilities, some differences are exaggerated and others minimized. Cells with 39% and 41% exceedence probabilities have different colors, while cells with 41% and 59% exceedence probabilities have the same color. The maps best serve to identify regions where consistently low or high probabilities are found, aiding in, but not replacing the interpretation of statistical analyses.

Statistical methods can be used to confirm and quantify relationships suggested by visual examination of maps. Some degree of correlation between precipitation and increased incidence of high nitrate concentrations is apparent when maps of the two are compared. The regression analysis presented in Chapter 6 confirms this relationship, and allows it to be compared to other potential indicators of nitrate contamination.

Forming a Mathematical Model. Stepwise multiple linear regression was used to form an estimate of exceedence probabilities for the 7.5' quadrangles based on values of average annual precipitation, soil thickness, soil organic content, and nitrogen fertilizer sales. The regressions showed little dependency on any of these parameters except average annual precipitation. Although this was, in some respects, the least successful aspect of this study, the lack of correlation to nitrogen fertilizer sales and soil properties is an interesting result in itself. The regression results are discussed further in Section 7.2.

Given the lack of significant correlation to the selected indicators, it is difficult to determine from the results of this work whether the method of

multiple linear regression is suitable for forming a model of groundwater vulnerability.

Convenience of Databases and GIS. Because a printed document like this one is a static object, the advantages of easily accessible on-line data are not well represented here. An example illustrates some of these advantages. In Section6.2.4, a single quadrangle in the Ogallala Aquifer with an unusually high 5 mg/l exceedence probability was described. The wells in the quad were found to be mostly shallow domestic supply wells, and the quad was found to be near the town of Big Spring amid a number of small oil fields. A regulator considering vulnerability waivers for herbicide monitoring might undertake a similar examination of the region surrounding a water supply well or well field. Once the quadrangle was identified by number (after some programming effort has been invested, this can be accomplished with a mouse point-and-click operation) all the descriptive data about wells and measurements in the quadrangle were accessible in seconds. In contrast, the information about nearby towns and oil fields required examination of paper maps and consumed about twenty minutes time.

Now that the programs used in this study have been written and tested, they can be applied with little modification to any set of water quality measurements. Modifying the programs to estimate exceedence probabilities from the Water Utilities Division data set required only a few minutes work. The process of acquiring that data set, mapping it into the existing quadrangles, and counting measurements and exceedences was the work of less than one day.

Similarly, additional indicators can easily be incorporated into the analysis, if they are available in the form of GIS coverages.

7.2 RESULTS IN TEXAS

The preceding section presented general conclusions about the methods developed for this study and their usefulness in describing the vulnerability of groundwater to contamination by nitrates, or by other constituents. This section is concerned with the results of the application of those methods to nitrate in Texas groundwater.

Interpreting Nitrate Results. Two assumptions are fundamental interpreting nitrate exceedence probabilities as an indication of groundwater vulnerability. The first of these is that the frequency of detection of elevated nitrate levels in regions as reported in a database of historic measurements collected from a variety of wells is a useful indicator of the likelihood of detecting nitrate in public water supplies in the same regions at the present and in the future. The second is that vulnerability to nitrate contamination is related to vulnerability to other contaminants.

The first assumption is confirmed by the nitrate measurements in the Water Utilities Division database. A comparison of the exceedence probabilities estimated from the TWDB data set of 46,507 records with the measurements listed in the WUD data set of 11,698 records shows that nitrate measurements taken over a short, recent period (February 1993 to October 1994) from public water supply wells conform to exceedence probabilities estimated from measurements collected over a much longer period (January 1962 to October

1993) from a more diverse set of wells. The correspondence between the two data sets is far from exact, but the recent measurements from public supply wells in the WUD database are much more likely to yield elevated nitrate levels if the wells are located in quadrangles with high exceedence probabilities, as estimated from the TWDB data set.

The second assumption, that nitrate exceedence probabilities are indicative of an intrinsic vulnerability is confirmed by two observations. The first is that the occurrence of elevated nitrate levels appears to be nearly independent of the sources of nitrate examined in this study. Nitrate in groundwater is uncorrelated with nitrogen fertilizer sales, indicating that its presence at elevated levels is due to other factors, such as the ease with which contaminants can enter the groundwater. The fact that the incidence of groundwater nitrate contamination is low in parts of east Texas where nitrogen fertilizer sales exceed four tons per square mile strongly suggests that the groundwater in that region is isolated from human influence to a much greater degree than in the Texas Panhandle, where fertilizer sales are lower and nitrate detections are more frequent.

The second confirmation comes from data collected in a reconnaissance of the Midwestern U.S. for nitrate and herbicides. A comparison of nitrate and herbicide data shows that when water quality measurements are grouped by hydrogeologic factors and these groups of measurements are ranked by the rates of detection of nitrate and herbicides, the rankings are virtually identical. Although neither of these observations forms conclusive evidence that nitrate

levels are indicators of general vulnerability, they are both consistent with that assumption.

Best Indicator. Of the four threshold concentrations, the 1 mg/l exceedence probabilities appear to be the best indicator of groundwater vulnerability. This is the level most likely to show increases over time, a more suggestive indicator than high concentrations alone. Maps of this exceedence probability also show more variability than the others, and geologic associations appear most strongly in the map of this threshold. Finally, nitrate detections at 1 mg/l correlate well to herbicide detections in the data from the Midwest reconnaissance study.

Statewide Patterns. The maps presented in Chapter 6 clearly reveal large-scale patterns in the occurrence of nitrate in Texas groundwater. Large, cohesive regions within the State can be seen to have high exceedence probabilities for nitrate at the 0.1, 1.0, 5.0, and 10.0 mg/l concentration thresholds. This strongly suggests that such regions can be identified and classified by groundwater vulnerability for regulatory purposes.

In all of the maps, the influence of geology on water quality can be plainly seen. The adjacent regions of the Balcones Fault Zone of the Edwards Aquifer and the Carrizo-Wilcox Aquifer contrast sharply at the detection level and the 1 mg/l concentration threshold. At the 5 and 10 mg/l concentration thresholds, the Seymour Aquifer is visible among the few regions where detection rates are high.

Indicators. Of the parameters examined for use as indicators of groundwater vulnerability, only average annual precipitation was tied to substantial variations

in exceedence probabilities, and this relationship was found to be opposite to expectation. Of the soil parameters, organic matter content of the soil was correlated to exceedence probabilities, but accounted for only a small part of the observed variation in those probabilities. Soil thickness and nitrogen fertilizer sales were found to have little value as indicators.

The weak link between soil parameters and exceedence probabilities may be due in part to the poor spatial resolution of the STATSGO data. The STATSGO map units are large and heterogeneous regions, with no subdivisions to indicate deviations from average soil parameter values. A map unit with an average soil thickness of fifteen inches, for example, may contain large areas with virtually no soil at all. If a quadrangle falls in such an area, the map unit average soil thickness may be a poor reflection of the actual conditions in that quad.

A similar argument can be made for the lack of correlation between nitrogen fertilizer sales and groundwater nitrate. Fertilizer sales are aggregated by county, and application of those fertilizers may be very uneven within those counties.

Average annual precipitation is subject to less local variation, exhibiting more gradual trends over the State. The size of the units used to map precipitation is more appropriate to the scale of its variation, which may account in part for its relatively high correlation to exceedence probabilities in the quadrangles.

Although there is a relationship between precipitation and nitrate exceedence probability, it is beyond the capability of statistical studies like this one to determine whether that relationship is causal. It is somewhat surprising to see that nitrate concentrations are higher where there is less rain. High recharge rates, which are driven by precipitation, are associated with increased vulnerability in DRASTIC, for example.

There is a pronounced trend in precipitation in Texas: southeast is wet; west is dry. Because of this trend, which also corresponds to important variations in geology, precipitation may be acting as a surrogate for location and aquifer structure. This explanation is supported by the lack of correlation between precipitation and exceedence probabilities *within* aquifers.

Aquifers as Indicators. Within the limited set of parameters tested in the study, geology—as represented by association of wells with the five example aquifers—appears to dominate over surface parameters such as soil properties, precipitation and fertilizer sales as an indicator of groundwater quality. A model of nitrate exceedence probability as good as any produced by the regressions in this work could be constructed by calculating average exceedence probabilities for each geologic formation and ignoring all other factors. Table 7.1 summarizes the exceedence probability estimate for the five aquifers and the State. The aquifers are listed in increasing order of 1 mg/l exceedence probability.

Table 7.1 Exceedence Probability Summary

	Threshold Exceedence Probability			
Aquifer	0.1 mg/l	1 mg/l	5 mg/l	10 mg/l
Carrizo-Wilcox	24%	7%	2%	1%
Hueco-Mesilla Bolson	79%	46%	3%	0.9%
Ogallala	94%	73%	12%	5%
Edwards (BFZ)	94%	74%	0.8%	0.2%
Seymour	96%	94%	82%	57%
Statewide	64%	44%	16%	9%

The Carrizo-Wilcox aquifer is the least likely of the five study aquifers to produce water with elevated nitrate levels, and the Seymour Aquifer is the most likely. Low exceedence probabilities were found at all thresholds in the Carrizo-Wilcox, and high exceedence probabilities were found at all thresholds in the Seymour.

More complex behaviors were seen in the Hueco-Mesilla Bolson, the Ogallala, and the Balcones Fault Zone of the Edwards. In these three aquifers, the exceedence probabilities were high at the detection level and the 1 mg/l threshold and low at the 5 and 10 mg/l threshold. A plausible explanation of this behavior can be found in the permeable structures of these aquifers (the Ogallala and Hueco-Mesilla Bolson are largely fluviatile, and the Edwards is karst). The aquifers may be vertically penetrable, increasing the vulnerability to nitrate (and other contaminants), and accounting for the high exceedence probabilities at low concentrations. At the same time, rapid horizontal motion of water through the aquifers could disperse the contaminants, preventing concentrations from reaching the higher thresholds. These aquifers also show little variation in

concentration with depth, again suggesting that constituents are well mixed and dispersed.

Variations with Depth and Time. The Texas Water Development Board's descriptions include the depth of all wells, but screened interval depths are available for only a small number of wells, and were not used in this study. A shallow well can draw water only from near the surface, but a deep well may collect groundwater along its whole depth, so well depth is a flawed indicator of water quality variations in the vertical dimension. It is true, however, that for the State as a whole, shallower wells are more likely to exhibit high nitrate concentrations (see Figure 6.12). As stated above, however, this trend is subject to variation within individual aquifers.

Similarly, while there has been an increase in the likelihood of finding nitrate in excess of 1 mg/l over time across the state, this trend can be found in only the Ogallala among the five aquifers studied here. An increase in the presence of any chemical over time is a strong indicator of vulnerability to contamination.

7.3 RECOMMENDATIONS

Recommendations resulting from this study fall into three broad categories: recommendations for use of the method, recommendations for further study, and recommendations for the use of vulnerability assessments in the regulation of groundwater.

Using the Vulnerability Assessment Method. The six steps in the method are repeated one last time, with recommendations for their application.

1. *Select a constituent or set of constituents whose presence indicates the degree of vulnerability of a groundwater source.* If the goal of the assessment is to predict the presence of a particular constituent, measurements of the constituent itself should be used, if possible. If such measurements are unavailable, another closely correlated constituent should be selected.

   If the goal is to assess a more general vulnerability to contamination by human activities, the ideal constituent would be one which has been widely measured and has no natural sources. Nitrate, because of its many sources, is not an ideal constituent for study, although it has been very widely measured. The author was unable to find a strictly anthropogenic constituent with a sufficient record of measurements. One possible way around this problem would be to combine measurements of a group of anthropogenic constituents, as was done in this report in the examination of the midwest data.

2. *Identify a set of distinct mappable regions of the surface or subsurface.* Ideally, the regions should be both homogeneous and highly populated. Because the effects of diminishing population size on confidence in estimated exceedence probabilities can be described mathematically, and the effects of heterogeneity cannot, it is better to sacrifice numbers for consistency.

If the data are sufficiently dense, measurements should be grouped into geologically homogeneous regions first, then stratified by depth or mapped by location in two dimensions.

3. *Assemble a body of measurements of the constituent identified in step 1 that can be linked with the regions identified in step 2.* The availability of data will dictate most of the study design. Although this is the third step in the method, knowledge of the available data is essential before the design of the study (steps 1 and 2) can be carried out.

4. *Calculate descriptive statistics for the body of measurements linked with each region.* Exceedence probabilities for threshold concentrations are, in the author's view, the best available quantitative measure of goundwater vulnerability. Exceedence probability estimates and confidence intervals can be calculated easily from databases, and can be compared through standard statistical methods to indicator parameters.

   As Figures 6.33 and 6.34 indicate, it is unlikely that a single probability distribution form can be used to describe the population of constituent concentrations in a body of groundwater. For this reason, estimates of exceedence probabilities should be calculated for discrete concentration thresholds using the binomial (Bernoulli process) method.

5. *Map the variation of the descriptive statistics from region to region.* Maps are an important and powerful method for communicating information about quantities that vary spatially. The maps produced in this study have provoked much more discussion and thought than tables

of results ever would have. Although it would be possible to carry out the other five steps in this method without creating maps, this step should not be omitted.

Some improvements could be made in the maps presented here, however. The division of exceedence probabilities into ranges of 0–20%, 20–40%, etc. is essentially arbitrary. It is probably of little concern to a regulator to distinguish between a 70% and a 90% probability of exceeding a maximum contaminant level. A scale that provides more resolution where exceedence probabilities are low and less resolution where they are high would be a better aid to regulatory decision-making.

6. *Relate the variation of the descriptive statistics to the variation of indicator parameters by forming a mathematical expression that mimics the relationship between the descriptive statistics and indicator values mapped over the same set of regions.* The use of multiple regression to evaluate the correlation of indicators to exceedence probabilities was inadequately tested in this study because of the lack of correlation between the chosen indicators and the presence of nitrate in groundwater. Some recommendations about potential indicators are in order to provide better tests of the linear regression method, and produce more meaningful predictions of groundwater quality.

More emphasis should be placed on *sources* of the constituent. Only nitrate fertilizer use was considered in this study, and with poor spatial resolution. Future studies of nitrate should consider sources from sewage,

from livestock production, and from natural sources—animal, vegetable, and mineral. The spatial resolution of the data and the directness with which it reflects the amount of nitrate actually available as a contamination source should be improved. The county-averaged nitrogen fertilizer sales used in this study are both spatially unfocused and causally indirect.

The STATSGO database contains values for many more soil parameters than were tested in this study. These should be investigated further. Soil permeability, for example, might be a more valuable indicator than soil thickness or organic content, which were examined here.

Because the results of this study indicate a strong dependence on geology, more detailed data on such parameters as aquifer conductivity, porosity, and depth should be used. Because such quantities vary in three dimensions, some effort will be required either to express these in the two-dimensional domain of geographic information systems, or to expand GIS to deal with three-dimensional data.

Finally, the inverse relationship between high nitrate exceedence probabilities and rainfall is very intriguing. It is possible, for example, that in east Texas, where rainfall rates are higher than in the west, more nitrate is carried away in surface runoff and removed from the groundwater system. It may be that rainfall or recharge rates are less valuable indicators of groundwater quality than the relative weights of runoff and recharge. Some effort should be directed toward developing

an index that properly expresses this relationship. The ratio of recharge to runoff might be a good place to start.

Future Work. The original goal of this study was to formulate a method for identifying groundwater sources sufficiently protected from contamination that they could be granted waivers from monitoring for man-made agricultural chemicals. This goal has not been fully met, but progress has been made toward it. In further pursuit of that goal, several steps should be taken.

The method developed in this study should be applied to measurements of herbicides and other man-made contaminants. Because nitrate is the product of both natural and anthropogenic processes, the detection of elevated nitrate levels is somewhat ambiguous as an indicator of vulnerability to human activities. If a groundwater supply shows detectable levels of atrazine, for example, there can be little question that it is vulnerable to human activities.

The 7.5' quadrangles used in this study were in part an artifact of the Texas Water Development Board's well-numbering system. Although they produce a convenient grid for exploring Texas ground water, alternative study regions should be examined. Divisions of groundwater following more physically-based boundaries, like the five aquifers studied here should be considered. Since the ultimate goal of studies like this one is to identify regions of high and low vulnerability, it is appropriate to form study regions on the basis of divisions in factors that influence vulnerability.

If studies of this type are to produce viable methods for vulnerability assessment, the most urgent need is for more complete indicator data sets. At

present, data tends to be either detailed, or widespread, but not both. The Soil Conservation Service is producing a set of soil data more detailed than STATSGO, but its coverage is still only a fraction of the country. As this data becomes available, it is possible that closer correspondence between soil parameters and water quality could be found.

Because geology appears to be the dominant influence on water quality, GIS coverages of geological parameters should be developed as part of any serious effort at GIS-based vulnerability assessment. The problems of representing the three-dimensional variations of the earth's structure in the two-dimensional domain of GIS are substantial. One possible approach would be to mimic the STATSGO data structure, identifying horizontal regions of uniform geological properties in a polygon coverage, and representing their vertical variations in tables linked to the coverage. In any case, geological databases are a necessity if GIS is to play a significant role in groundwater vulnerability assessment.

Regulatory Suggestions. The vulnerability assessments for granting waivers for monitoring of agricultural chemicals in public groundwater supplies require evaluation of individual wells. Although this study has focused on regional variations in nitrate concentrations, some recommendations are still appropriate. If a well is to be classified as protected from contamination, it is necessary to show that the well has adequate protection from backwash down the well bore. Aurelius (1989) identified well construction and pesticide mix-and-load

operations close to wellheads as important contributors to contamination where pesticides were found in Texas wells.

After well construction, the most important element in assessing the vulnerability of a well to contamination is identification of the aquifers or formations from which it draws water. The EPA's Groundwater Task-Force makes the same recommendation (USEPA, 1991), listing identification of aquifers supplying wells as a high priority for state agencies dealing with groundwater quality. If a well is poorly constructed or draws from an aquifer that has a high incidence of contamination, then other environmental factors such as soil parameters will have very little influence on the well's vulnerability.

The results of this work may have more direct bearing on programs like the EPA's Differential Protection Program, which would restrict the use of certain agricultural chemicals in sensitive areas, rather than banning their use everywhere. The maps in Chapter 6 clearly show that vulnerability to contamination by nitrate varies from region to region. If similar results can be shown for man-made contaminants such as herbicides, then there is a difference in the risks associated with using such chemicals in different regions, and a basis for regionally differing restrictions. Spatial and statistical analysis of existing groundwater contamination can help identify vulnerable regions for such regulatory purposes.

References

40 CFR 141. Code of Federal Regulations, Title 40, Part 141.

Abramowitz, M., and I. Stegun 1965. Handbook of Mathematical Functions. New York: Dover.

Aho, A., B. Kernigan, and P. Weinberger 1988. The AWK Programming Language. New York: Addison-Wesley.

Aller, L., T. Bennett, J. Lehr, R. Petty, and G. Hackett 1987. DRASTIC: A Standardized System for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings. Ada, Oklahoma: U.S. Environmental Protection Agency.

Ashworth, J. B. and R. R. Flores 1991. "Delineation Criteria for the Major and Minor Aquifer Maps of Texas." Austin, Texas: Texas Water Development Board LP-212.

Aurelius, L. 1989. Testing for Pesticide Residues in Texas Well Water. Austin, Texas: Texas Department of Agriculture.

Baker, D., L. Wallrabenstein, and R. Richards 1994. "Well Vulnerability and Agricultural Contamination: Assessments from a Voluntary Well Testing Program," Procedings of the Fourth National Conference on Pesticides: New Directions in Research, Development and Policy. D. Weigmann, ed. Blacksburg, Virginia: Virginia Water Resources Center.

Barringer, T., D. Dunnn, W, Battaglin, and E. Vowinkle 1990. "Problems and Methods Involved in Relating Land Use to Ground-Water Quality," Water Resources Bulletin, vol. 26, no. 1.

Barton, C., E. Vowinkel, J. Nawyn 1987. Preliminary Assessment of Water Qualtiy and its Relation to Hydrogeology and Land Use: Potomac-Raritan-Magothy Aquifer System, New Jersey. U.S. Geological Survey Water-Resources Investigation Report 87-4023. Trenton, New Jersey: U.S. Geoloical Survey.

Battaglin, W. A. and D. Goolsby 1995. Spatial Data in Geographic Information System Format on Agricultural Chemical Use, Land Use, and Cropping Pactices in the United States. U.S. Geological Survey Water-Resources Investigation Report 94-4176. Denver, Colorado: U.S. Geoloical Survey.

BEG (Bureau of Economic Geology) 1968. Geologic Atlas of Texas (Palestine Sheet). Austin Texas: Bureau of Economic Geology.

BEG (Bureau of Economic Geology) 1974a. Geologic Atlas of Texas (Austin Sheet). Austin Texas: Bureau of Economic Geology.

BEG (Bureau of Economic Geology) 1974b. Geologic Atlas of Texas (San Antonio Sheet). Austin Texas: Bureau of Economic Geology.

BEG (Bureau of Economic Geology) 1987. Geologic Atlas of Texas (Wichita Falls-Lawton Sheet). Austin Texas: Bureau of Economic Geology.

BEG (Bureau of Economic Geology) 1993. Geologic Atlas of Texas (Van Horn-El Paso Sheet). Austin Texas: Bureau of Economic Geology.

Blodgett, R. 1993. "Drinking Water Vulnerability Assessments in Texas," Proceedings—1993 Annual Conference, American Water Works Association.

Brown, B.W. and J. Lovato 1994. SCDFLIB: Library of Fortran Routines for Cumulative Distribution Functions, Inverses, and Other Parameters. Houston, Texas: M.D. Anderson Cancer Center. (available via anonymous FTP at odin.mda.uth.tmc.edu)

Burkart, M. and D. Kolpin 1993a. "Hydrologic and Land-Use Factors Associated with Herbicides and Nitrate in Near-Surface Aquifers," Journal of Environmental Quality, vol. 22, pp. 646-656.

Burkart, M. and D. Kolpin 1993b. "Application of a Geographic Information System in Analyzing the Occurance of Atrazine in Groundwater of the Mid-Continental United States," HydroGIS 93: Application of Geographic Information Systems in Hydrology and Water Resources, IAHS Publication No. 211.

Chen, H. and D. Druliner 1988. "Agricultural Chemical Contamination of Ground Water in Six Areas of the High Plains Aquifer, Nebraska," National Water Summary 1986—Hydrologic Events and Ground Water

Quality. Water-Supply Paper 2325. Reston, Virginia: U.S. Geological Survey.

Chow, V. T., D. R. Maidment, L. W. Mays 1988. Applied Hydrology. New York: McGraw-Hill.

Cohen, S., S. Creeger, R. Carsel, and C. Enfield 1984. "Potential for Pesticide Contamination of Groundwater Resulting from Agricultural Uses," Treatment and Disposal of Pesticide Wastes, American Chemical Society Symposium Series No. 259, R. Krueger and J. Seiber (Eds.)

Domagalski, J., and N. Dubrovsky 1992. "Pesticide Residues in Ground Water of the San Joaquin Valley, California," Journal of Hydrology, Vol. 130, pp. 299–338.

Draper, N. and H. Smith 1981. Applied Regression Analysis. New York: John Wiley and Sons.

Eliot, T. S. 1963. "A Dedication to my Wife", Collected Poems 1909–1962. New York: Harcourt Brace Jovanovitch.

ESRI (Environmental Systems Research Institue) 1990. Understanding GIS: The Arc/Info Method. Redlands, California: ESRI, Inc.

ESRI (Environmental Systems Research Institue) 1991. Arc/Info Data Model, Concepts, and Key Terms. Redlands, California: ESRI, Inc.

GAO (General Accounting Office) 1991. Goundwater Protection: Measurement of Relative Vulnerability to Pesticide Contamination. (GAO/PEMD-92-8) Washington, D.C.: General Accounting Office.

GAO (General Accounting Office) 1992. Goundwater Protection: Validity and Feasibility of EPA's Differential Protection Strategy. (GAO/PEMD-93-6) Washington, D.C.: General Accounting Office.

Haan, C. T. 1977. Statistical Methods in Hydrology. Ames, Iowa: Iowa State University Press.

Harvard University Computation Laboratory 1955. Tables of the Cumulative Binomial Probability Function. Cambridge, Massachusetts: Harvard University Press.

Helgesen, J., L. Stullken, and A. Rutledge 1987. Assessment of Nonpoint-source Contamination of the High Plains Aquifer in South Central Kansas, USGS Open-File Report 91-238. Denver, Colorado: U. S. Geological Survey.

Helsel, D. R., and R. M. Hirsch 1992. Statistical Methods in Water Resources. New York: Elsevier.

Hem, J. D. 1989. Study and Interpretation of the Chemical Characteristics of Natural Water. U.S. Geological Survey Water-Supply Paper 2245. Washington, D. C.: United States Government Printing Office.

Hydroshpere Data Products, Inc. 1994. NCDC Summary of theDay Data Guide. Boulder, CO: Hydroshpere Data Products, Inc.

Kolpin, D. and M. Burkart 1991. Work Plan for Regional Reconnaissance for Selected Herbicides and Nitrate in Ground Water of the Mid-Continental United States, 1991. U.S. Geological Survey Open-File report 91-59. Iowa City, Iowa: U.S. Geological Survey.

Kolpin, D., M. Burkart, and M. Thurman 1993. Hydrogeologic, Water-Quality, and Land-Use Data for the Reconnaissance of Herbicides and Nitrate in Near Surface Aquifers of the Midcontinental United States, 1991. U.S. Geological Survey Open-File report 93-114. Iowa City, Iowa: U.S. Geological Survey.

Korth, H. and A. Siberschatz 1991. Database System Concepts. New York: McGraw-Hill.

Madison, R. and J. Brunett 1985. "Overview of the Occurrence of Nitrate in Ground Water of the United States," National Water Summary 1985—Hydrologic Events, Selected Water-Quality Trends and Ground-Water Resources. Water-Supply Paper 2275. Reston, Virginia: U.S. Geological Survey.

Nightingale, H. and W. Bianchi 1980. "Correlation of Selected Well Water Quality Parameters with Soil and Aquifer Hydrologic Properties," Water Resources Bulletin, vol 16, no. 4.

NRC (National Research Council) 1993. Ground Water Vulnerability Assessment: Predicting Relative Contamination Potential Under Conditions of Uncertainty. Washington, D.C.: National Academy Press.

Nordstrom, P. L. and R. Quincy 1992. Ground-Water Data System Data Dictionary. Austin, Texas: Texas Water Development Board.

Nordstrom, P. L. 1994. Personal Communitcation.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling 1988. Numerical Recipes in C. New York: Cambridge University Press.

Snedecor, G. and W. Cochran, 1980. Statistical Methods (Seventh Edition). Ames, Iowa: The Iowa State University Press.

STSC, Inc. 1989. STATGRAPHICS User's Guide. Rockville, Maryland: STSC Inc.

Teso, R., T. Younglove, M. Peterson, D. Sheetks III, and R. Gallavan 1988. "Soil Taxonomy and Surveys: Classification of Areal Sensitivity to Pesticide Contamination of Groundwater", Journal of Soil and Water Conservation vol. 43, no. 4.

TWDB (Texas Water Development Board) 1992. Water for Texas Today and Tomorrow; The 1992 Update of the Texas Water Plan. Austin, Texas: Texas Water Development Board.

USEPA (United States Environmental Protection Agency) 1991. Protecting the Nation's Ground Water: EPA's Strategy for the 1990s. (WH-550G) Washington, D.C.: USEPA.

USEPA (United States Environmental Protection Agency) 1992. Pesticides in Ground Water Database, A Compilation of Monitoring Studies: 1971–1991. Washington, D.C.: USEPA.

USDA (United States Department of Agriculture) 1993. State Soil Geographic Data Base (STATSGO)—Data User's Guide. Soil Conservation Service Miscellaneous Publication Number 1492. Washington D. C.: United States Department of Agriculture.

Wendland, F., H. Albert, M. Bach and R. Schmidt 1993. Atlas zum Nitratstrom in der Bundersrepublik Deutschland. Berlin: Springer-Verlag.

Wisconsin Bureau of Water Supply 1993. "Phase II/IIB Vulnerability Assessment Procedures."

Appendix A: Database Definitions

WELL DATA
- TWDB_WELLS.DAT
- INCLUDE.WELLS
- AQ5.WELLS

NITRATE MEASUREMENT DATA
- TWDB_WELLS.NIT
- INCLUDE.NIT

PRECIPITATION DATA
- PREC.DAT
- STATION.MEAN

FERTILIZER SALES DATA
- NITRATE.USE

SOIL PARAMETER DATA
- STUDY.MAPU
- STUDY.COMP
- STUDY.LAYER

QUADRANGLE AQUIFER ASSOCIATIONS
- AQ_QUAD.DAT

DISCRETE PROBABILITIES RESULTS
- COUNTS.QUAD
- COUNTS.QUAD (extended for WUD data)
- BINO.QUAD

LOGNORMAL PROBABILITIES RESULTS
- LOGFIT.QUAD

QUAD PARAMETERS FOR REGRESSION
- PARAMS.QUAD

WUD NITRATE AND WELL DATA
- NIT.WRK
- POE.WRK

PWS-QUAD.PAT

MIDWEST NITRATE AND HERBICIDE DATA
- CONSTRUCTION
- QUALITY

WELL DATA

TWDB_WELLS.DAT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	WELLNO	7	7	I	-
8	AQFCODE	8	8	C	-
16	FIPSCODE	3	3	I	-
19	LATITUDE	7	7	I	-
26	LONGITUDE	7	7	I	-
33	LOCMETHD	1	1	I	-
34	DEPTH	6	6	I	-
40	DEPMETH	1	1	C	-
41	ALTITUDE	5	5	I	-
46	ALTMETH	1	1	C	-
47	DRILLDATE	8	8	C	-
55	PRIMEUSE	1	1	C	-
**	REDEFINED ITEMS **				
1	QUAD_2.5M	5	5	I	-
1	QUAD_7.5M	4	4	I	-
1	QUAD_1D	2	2	I	-

INCLUDE.WELLS

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	WELLNO	7	7	I	-
8	AQFCODE	8	8	C	-
16	FIPSCODE	3	3	I	-
19	LATITUDE	7	7	I	-
26	LONGITUDE	7	7	I	-
33	LOCMETHD	1	1	I	-
34	DEPTH	6	6	I	-
40	DEPMETH	1	1	C	-
41	ALTITUDE	5	5	I	-
46	ALTMETH	1	1	C	-
47	DRILLDATE	8	8	C	-
55	PRIMEUSE	1	1	C	-
56	QUAD_OK	1	1	C	-
57	QUAD_ERR	1	1	C.....	-
58	MEAS	1	1	C	-
59	INCLUDE	1	1	C	-
**	REDEFINED ITEMS **				
1	QUAD_2.5M	5	5	I	-
1	QUAD_7.5M	4	4	I	-
1	QUAD_1D	2	2	I	-

AQ5.WELLS

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	WELLNO	7	7	I	-
8	AQF	4	4	C	-
12	AQFCODE	8	8	C	-
20	FIPSCODE	3	3	I	-
23	LATITUDE	7	7	I	-
30	LONGITUDE	7	7	I	-
37	LOCMETHD	1	1	I	-
38	DEPTH	6	6	I	-
44	DEPMETH	1	1	C	-
45	ALTITUDE	5	5	I	-
50	ALTMETH	1	1	C	-
51	DRILLDATE	8	8	C	-
59	PRIMEUSE	1	1	C	-
60	QUAD_OK	1	1	C	-
61	QUAD_ERR	1	1	C	-
62	MEAS	1	1	C	-
63	INCLUDE	1	1	C	-
**	REDEFINED ITEMS **				
1	QUAD_2.5M	5	5	I	-
1	QUAD_7.5M	4	4	I	-
1	QUAD_1D	2	2	I	-

NITRATE MEASUREMENT DATA

TWDB_WELLS.NIT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	WELLNO	7	7	I	-
8	MM_DATE	2	2	B	-
10	DD_DATE	2	2	B	-
12	YY_DATE	2	4	B	-
14	RELIABILITY_REM	2	2	C	-
16	COLLECT_AGENCY	2	2	C	-
18	Q71850_FLAG	1	1	C	-
19	Q71850_NITRATE	8	9	F	2
**	REDEFINED ITEMS **				
1	QUAD_2.5M	5	5	I	-
1	QUAD_7.5M	4	4	I	-
1	QUAD_1D	2	2	I	-

INCLUDE.NIT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	WELLNO	7	7	I	-
8	MM_DATE	2	2	B	-
10	DD_DATE	2	2	B	-
12	YY_DATE	2	4	B	-
14	RELIABILITY_REM	2	2	C	-
16	COLLECT_AGENCY	2	2	C	-
18	Q71850_FLAG	1	1	C	-
19	Q71850_NITRATE	8	9	F	2
27	INCLUDE	1	1	C	-
28	NIT_ADJ	8	9	F	2
**	REDEFINED ITEMS **				
1	QUAD_2.5M	5	5	I	-
1	QUAD_7.5M	4	4	I	-
1	QUAD_1D	2	2	I	-

PRECIPITATION DATA

PREC.DAT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	STATION_ID	5	5	I	-
6	STATION_NAME	23	23	C	-
29	YEAR	4	4	I	-
33	PREC	8	6	F	2
**	REDEFINED ITEMS **				
1	UNIQUE	28	28	C	-

STATION.MEAN

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	STATION_ID	5	5	I	-
6	STATION_NAME	23	23	C	-
29	STATE	2	2	I	-
31	CNT_40	4	5	B	-
35	GAP_90	2	2	I	-
37	TOT_40	8	8	F	2
45	MEAN_40	8	18	F	6
53	DELTA-30-40	8	18	F	6
61	CNT_30	4	5	B	-
65	GAP_80	2	2	I	-
67	TOT_30	8	8	F	2
75	MEAN_30	8	8	F	2
**	REDEFINED ITEMS **				
1	UNIQUE	28	28	C	-

FERTILIZER SALES DATA

NITRATE.USE

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	FIPSCNTY	3	3	I	-
4	NTOT86	8	18	F	6
12	NTOT86.USE	4	12	F	3
16	NTOT87	8	18	F	6
24	NTOT87.USE	4	12	F	3
28	NTOT88	8	18	F	6
36	NTOT88.USE	4	12	F	3
40	NTOT89	8	18	F	6
48	NTOT89.USE	4	12	F	3
52	NTOT90	8	18	F	6
60	NTOT90.USE	4	12	F	3
64	NTOT91	8	18	F	6
72	NTOT91.USE	4	12	F	3
76	NTOT86-91	8	18	F	6
84	NTOT86-91.USE	8	18	F	6
92	NUSE86-91.RNK	3	3	I	-
95	NTOT86-91.AVUSE	8	18	F	6

SOIL PARAMETER DATA

STUDY.MAPU

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	STSSAID	5	5	C	-
6	SSAID	3	3	C	-
9	MUSYM	5	5	C	-
14	MUID	7	7	C	-
21	MUNAME	109	109	C	-
130	MUKIND	1	1	C	-
131	MLRA	4	4	C	-
135	PRIMFML	2	2	C	-
137	MUAREA	8	18	F	2
145	MUACRES	6	6	I	-
151	SUM	2	3	B	-
153	AVTHK	8	6	F	2
161	AV-MAX-ORG	8	8	F	4
169	AV-MID-ORG	8	8	F	4
177	AV-MIN-ORG	8	8	F	2

STUDY.COMP

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	STSSAID	5	5	C	-
6	MUID	7	7	C	-
13	SEQNUM	2	2	I	-
15	SOILTHK	8	18	F	6
23	MAX-ORG	8	8	F	2
31	MID-ORG	8	8	F	2
39	MIN-ORG	8	8	F	2
**	REDEFINED ITEMS	**			
6	MAPSEQ	9	9	C	-

STUDY.LAYER

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	STSSAID	5	5	C	-
6	MUID	7	7	C	-
13	SEQNUM	2	2	I	-
15	S5ID	6	6	C	-
21	LAYERNUM	1	1	I	-
22	LAYERID	2	2	I	-
24	LAYDEPL	2	2	I	-
26	LAYDEPH	2	2	I	-
28	BDL	4	4	N	2
32	BDM	5	5	N	3
37	BDH	4	4	N	2
41	OML	4	4	N	1
45	OMM	5	5	N	2
50	OMH	4	4	N	1
**	REDEFINED ITEMS	**			
6	MAPSEQ	9	9	C	-

QUADRANGLE AQUIFER ASSOCIATIONS

AQ_QUAD.DAT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	EBFZ	1	1	I	-
6	CZWX	1	1	I	-
7	OGLL	1	1	I	-
8	SYMR	1	1	I	-
9	HMBL	1	1	I	-
10	AQ_CNT	1	1	I	-

DISCRETE PROBABILITIES RESULTS

COUNTS.QUAD

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	WELL_CNT	2	4	B	-
7	MEAS_CNT	2	4	B	-
9	DTCT_CNT	2	4	B	-
11	GT1_CNT	2	4	B	-
13	GT5_CNT	2	4	B	-
15	GT10_CNT	2	4	B	-
17	DTCT_PROB	8	8	F	6
25	GT1_PROB	8	8	F	6
33	GT5_PROB	8	8	F	6
41	GT10_PROB	8	8	F	6

COUNTS.QUAD (extended for WUD data)

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	WELL_CNT	2	4	B	-
7	MEAS_CNT	2	4	B	-
9	DTCT_CNT	2	4	B	-
11	GT1_CNT	2	4	B	-
13	GT5_CNT	2	4	B	-
15	GT10_CNT	2	4	B	-
17	DTCT_PROB	8	8	F	6
25	GT1_PROB	8	8	F	6
33	GT5_PROB	8	8	F	6
41	GT10_PROB	8	8	F	6
49	WUD_MEAS	2	4	B	-
51	WUD_GT1	2	4	B	-
53	WUD_GT5	2	4	B	-
55	WUD_GT10	2	4	B	-
57	WUD_DTCT	2	4	B	-
59	WDT_PROB	8	8	F	6
67	W1_PROB	8	8	F	6
75	W5_PROB	8	8	F	6
83	W10_PROB	8	8	F	6

BINO.QUAD

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	WELL_CNT	2	4	B	-
7	MEAS_CNT	2	4	B	-
9	DTCT_CNT	2	4	B	-
11	GT1_CNT	2	4	B	-
13	GT5_CNT	2	4	B	-
15	GT10_CNT	2	4	B	-
17	DTCT_PROB	8	8	F	6
25	DTCT_LO	8	6	F	4
33	DTCT_UP	8	6	F	4
41	GT1_PROB	8	8	F	6
49	GT1_LO	8	6	F	4
57	GT1_UP	8	6	F	4
65	GT5_PROB	8	8	F	6
73	GT5_LO	8	6	F	4
81	GT5_UP	8	6	F	4
89	GT10_PROB	8	8	F	6
97	GT10_LO	8	6	F	4
105	GT10_UP	8	6	F	4

LOGNORMAL PROBABILITIES RESULTS

LOGFIT.QUAD

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	MEAS_CNT	2	4	B	-
7	MED_NIT	8	7	F	3
15	MEANLOG	8	7	F	3
23	STDLOG	8	7	F	3
31	R2	8	7	F	4
39	T_STAT	8	6	F	2
47	F_STAT	8	7	F	2
55	STDERR	8	7	F	3
63	P_FSTAT	8	8	F	6
71	MP_DTCT	8	4	F	2
79	MP_1	8	4	F	2
87	MP_5	8	4	F	2
95	MP_10	8	4	F	2

QUAD PARAMETERS FOR REGRESSION

PARAMS.QUAD

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	QUAD_7.5M	4	4	I	-
5	SOILAREA	8	18	F	5
13	THKAR	8	18	F	5
21	OMMAR	8	18	F	5
29	AVSOILTHK	8	6	F	2
37	AVSOILOMM	8	10	F	4
45	PRECAREA	8	18	F	5
53	PRCAR	8	18	F	5
61	AVPREC	8	5	F	2
69	AVNIT86-91.USE	8	18	F	6
77	CTYAR	8	8	F	2
85	NITAR	8	8	F	2

WUD Nitrate And Well Data

NIT.WRK

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	PREFIX	3	3	C	-
4	LAB	5	5	I	-
9	DATECOLL	8	10	D	-
17	TIMECOLL	8	8	C	-
25	SYS-ID	7	7	C	-
32	POE	3	3	C	-
35	SYS_NAME	34	34	C	-
69	SYS_ADDR1	34	34	C	-
103	SYS_ADDR2	34	34	C	-
137	SYS_CITY	25	25	C	-
162	SYS_ZIP	9	9	C	-
171	TESTNO3	1	1	C	-
172	TESTNO2	1	1	C	-
173	TESTNO3NO2	1	1	C	-
174	NO3RESULTS	8	8	C	-
182	NO2RESULTS	8	8	C	-
190	NO3NO2RES	8	8	C	-
198	LABCOMMENT	40	40	C	-
238	SAMPLETYPE	1	1	C	-
239	COMMENT	20	20	C	-
259	LOCATION	34	34	C	-
293	ENTRYCODE1	3	3	C	-
296	ENTRYCODE2	3	3	C	-
299	ENTRYCODE3	3	3	C	-
302	ENTRYCODE4	3	3	C	-
305	ENTRYCODE5	3	3	C	-
308	SOURCE	34	34	C	-
342	OTHER	34	34	C	-
376	STATCODE	2	2	I	-
378	PRESERVED	1	1	I	-
379	DATEIN	8	10	D	-
389	QUAD_7.5M	4	4	I	-
393	NO3FL	1	1	C	-
394	NO3	4	6	F	2
497	NO2FL	1	1	C	-
498	NO2	4	6	F	2
402	NNFL	1	1	C	-
404	NO3NO2	4	6	F	2
**	REDEFINED ITEMS **				
25	SYSENT	10	10	C	-

POE.WRK

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	PWS-ID	7	7	C	-
8	POE	3	3	C	-
11	WATERSOURCE	10	10	C	-
21	QUAD_7.5M	4	4	I	-
**	REDEFINED ITEMS **				
1	SYSENT	10	10	C	-

PWS-QUAD.PAT

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	AREA	8	18	F	5
9	PERIMETER	8	18	F	5
17	PWS-QUAD#	4	5	B	-
21	PWS-QUAD-ID	4	5	B	-
25	PWS#	4	5	B	-
29	PWS-ID	4	5	B	-
33	PWSID	7	7	C	-
40	POE	3	3	C	-
43	WATERSOURCE	10	10	C	-
53	OWNERSDES	15	15	C	-
68	STATEWELL	7	7	C	-
75	LATITUDE	6	6	C	-
81	LONGITUDE	7	7	C	-
88	LOCACC	1	1	C	-
89	LOCAGEN	1	1	C	-
90	LOCMETH	10	10	C	-
100	DATUM	2	2	C	-
102	SPATREF	1	1	C	-
103	FIPS	3	3	C	-
106	QUADS	8	8	C	-
114	WELLSTAT	1	1	C	-
115	DEPTHAGEN	1	1	C	-
116	DEPTHSOURC	1	1	C	-
117	AQUIFER	8	8	C	-
125	AQUIAGEN	1	1	C	-
126	AQUIFMETH	1	1	C	-
127	AQUITYPE	1	1	C	-
128	AQUIPORO	1	1	C	-
129	REMARKS	1	1	C	-
130	INITIALS	3	3	C	-
* 133	QUADS_7.5#	4	5	B	-
* 137	QUADS_7.5-ID	4	5	B	-
* 141	QUAD_7.5M	4	4	I	-
**	REDEFINED ITEMS **				
33	SYSENT	10	10	C	-

*—item added by overlying with coverage quads_7.5

MIDWEST NITRATE AND HERBICIDE DATA

CONSTRUCTION

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	SITE_ID	5	5	C	-
6	LATITUDE	4	8	B	-
10	LONGITUDE	4	8	B	-
14	CONST_YEAR	4	4	I	-
18	WELL_DEPTH	2	4	B	-
20	OPEN_INT_TOP_DPH	2	4	B	-
22	OPEN_INT_BOT_DPH	2	4	B	-
24	PRIMARY_USE	1	1	C	-
25	AQ_CLASS	1	1	C	-
26	AQ_TYPE	1	1	C	-
27	AQ_MATERIAL	8	8	C	-
35	DPTH_AQ_TOP	2	4	B	-

QUALITY

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	SITE_ID	5	5	C	-
6	DUP_FLAG	1	1	C	-
7	SAMPLE_DATE	8	10	D	-
15	WATER_LEV_FLAG	1	1	C	-
16	WATER_LEVEL	2	4	B	-
18	SPEC_COND	2	4	B	-
20	PH	4	5	F	1
24	DISS_O	4	5	F	1
28	NITRITE_FLAG	1	1	C	-
29	NITRITE	4	6	F	2
33	NITRITE+ATE_FLAG	1	1	C	-
34	NITRITE+NITRATE	4	6	F	2
38	NITRATE	4	6	F	2
42	AMMONIUM_FLAG	1	1	C	-
43	AMMONIUM	4	6	F	2
47	PHOSPHORUS_FLAG	1	1	C	-
48	PHOSPHORUS_ORTHO	4	6	F	2
52	ALACHLOR_FLAG	1	1	C	-
53	ALACHLOR	4	6	F	2
57	ATRAZINE_FLAG	1	1	C	-
58	ATRAZINE	4	6	F	2
62	CYANAZINE_FLAG	1	1	C	-
63	CYANAZINE	4	6	F	2
67	D_E_ATRZN_FLAG	1	1	C	-
68	DEETHYLATRAZINE	4	6	F	2
72	D_IPL_ATRZN_FLAG	1	1	C	-
73	DEISOPROPYLATRZN	4	6	F	2
77	METOLACHLOR_FLAG	1	1	C	-
78	METOLACHLOR	4	6	F	2
82	METRIBUZIN_FLAG	1	1	C	-
83	METRIBUZIN	4	6	F	2
87	PROMETON_FLAG	1	1	C	-
88	PROMETON	4	6	F	2
92	SIMAZINE_FLAG	1	1	C	-
93	SIMAZINE	4	6	F	2

Appendix B: Computer Programs

This appendix contains the computer programs, some written in Arc/Info's macro language (aml), and others in C, FORTRAN, or AWK, that organize the data used in the study and calculate the reported statistics. The following list groups the programs according to their function and gives their names. The actual code follows, in the order listed.

WELL AND MEASUREMENT RECORD SELECTION
- test_quad.aml
- test_quad.c
- include.aml

PRECIPITATION CALCULATIONS
- years.aml
- firstyear
- lastyear
- gap
- precmean.aml

SOIL PARAMETER CALCULATIONS
- org_int.aml
- unit_avg.aml

QUADRANGLE GENERATION
- build_quads7.aml
- tx_7m.c

DISCRETE PROBABILITY CALCULATIONS
- count_quad.aml
- count_aq.aml
- count_aqquad.aml
- count_wud.aml
- bino2.f
- bino_quad.aml

LOGNORMAL PARAMETER CALCULATIONS

logfit.c

fit_quad.aml

MAP CREATION

gt1_plot.aml

QUADRANGLE PARAMETER AVERAGES

aw_avg.aml

```
/*  testquad.aml -- determines whether the number of a well listed
/*  in the TWDB database is consistent with the well's location
/*  data.

tables

  /*  define an info table to hold the results of the test
  define qtest.tab
    wellno 7 7 i
    quad_ok 1 1 c
    quad_err 1 1 c
    ~

  /*  create a text file with the necessary data for the test
  select twdb_wells.dat
    unload qtest.in wellno latitude longitude

  /*  execute the C program the performs the test
  SYSTEM test_quad qtest.in qtest.out

  /*  transfer the results to the new info table
  select qtest.tab
    add from qtest.out

q stop

/*  Expand the well data table with the results of the test.
JOINITEM twdb_wells.dat qtest.tab twdb_wells.dat wellno primeuse

/*  delete the text files and the temporary info table
&sv delstat := [DELETE -FILE qtest.in]
&sv delstat := [DELETE -FILE qtest.out]
&sv delstat := [DELETE -INFO qtest.tab]

&return
```

```
/*  test_quad.c -- tests whether a location, given by latitude and
*   longitude, is consistent with the TWDB number assigned to a
*   well */

/*  USAGE:  test_quad infile outfile */

#include <stdio.h>

/* define an array holding numbers and coordinates for the SE
*  corners of the 89 1-degree quads enclosing Texas, numbered
*  in accordance with TWDB wells */

int secor[89][3] = { {1, 103, 36},{2, 102, 36},{3, 101, 36},
{4, 100, 36},{5, 100, 35},{6, 101, 35},{7, 102, 35},
{8, 103, 35},{9, 103, 34},{10, 102, 34},{11, 101, 34},
{12, 100, 34},{13, 99, 34},{14, 98, 34},{15, 97, 34},
{16, 94, 33},{17, 95, 33},{18, 96, 33},{19, 97, 33},
{20, 98, 33},{21, 99, 33},{22, 100, 33},{23, 101, 33},
{24, 102, 33},{25, 103, 33},{26, 103, 32},{27, 102, 32},
{28, 101, 32},{29, 100, 32},{30, 99, 32},{31, 98, 32},
{32, 97, 32},{33, 96, 32},{34, 95, 32},{35, 94, 32},
{36, 93, 31},{37, 94, 31},{38, 95, 31},{39, 96, 31},{40, 97, 31},
{41, 98, 31},{42, 99, 31},{43, 100, 31},{44, 101, 31},
{45, 102, 31},{46, 103, 31},{47, 104, 31},{48, 105, 31},
{49, 106, 31},{50, 105, 30},{51, 104, 30},{52, 103, 30},
{53, 102, 30},{54, 101, 30},{55, 100, 30},{56, 99, 30},
{57, 98, 30},{58, 97, 30},{59, 96, 30},{60, 95, 30},{61, 94, 30},
{62, 93, 30},{63, 93, 29},{64, 94, 29},{65, 95, 29},
{66, 96, 29},{67, 97, 29},{68, 98, 29},{69, 99, 29},
{70, 100, 29},{71, 101, 29},{72, 102, 29},{73, 103, 29},
{74, 104, 29},{75, 103, 28},{76, 100, 28},{77, 99, 28},
{78, 98, 28},{79, 97, 28},{80, 96, 28},{81, 95, 28},
{82, 96, 27},{83, 97, 27},{84, 98, 27},{85, 99, 27},
{86, 99, 26},{87, 98, 26},{88, 97, 26},{89, 97, 25}};

main(argc, argv)
  int  argc;
  char  *argv[];
{
  int wellno, latitude, longitude, scancnt;
  int qdeg, q7, q2, remain;
  int latd, latm, lats, latmss, latmssmin, latmssmax;
  int lngd, lngm, lngs, lngmss, lngmssmin, lngmssmax;
  char loc_ok = 'y', errtype = 'n', inst[30];
  FILE *datafp, *outfp;

  /*  usage message for careless users */
  if(argc != 3){
    fprintf(stderr, "\nUSAGE:  %s infile outfile\n", argv[0]);
    exit(1);}

  /*  open the data file READ ONLY */
  if((datafp = fopen(argv[1], "r")) == NULL){
    fprintf(stderr, "\nunable to open data file:  %s.\n",
      argv[1]);
    exit(1);}
```

```
/*  open the output file */
if((outfp = fopen(argv[2], "w")) == NULL){
  fprintf(stderr, "\nunable to open output file:  %s.\n",
    argv[2]);
  fclose(datafp);
  exit(1);}

/*  read a line from the input file*/
while((scancnt = fscanf(datafp, "%s", inst)) != EOF){

  /*  if the line is properly formatted, extract and test
      the data  */
  if(sscanf(inst,"%d,%d,%d",
     &wellno, &latitude, &longitude) == 3){

    /*  parse the well number into degree quad, 7.5 min quad,
        and 2.5 min quad numbers */
    qdeg = wellno / 100000;
    remain = wellno % 100000;
    q7 = remain / 1000;
    remain = remain % 1000;
    q2 = remain / 100;
    remain = remain % 100;

    /*  parse latitude into degrees minutes seconds */
    latd = latitude / 10000;
    remain = latitude % 10000;
    latm = remain / 100;
    lats = remain % 100;

    /*  parse longitude into degrees minutes seconds  */
    lngd = longitude / 10000;
    remain = longitude % 10000;
    lngm = remain / 100;
    lngs = remain % 100;

    /*  add minutes and seconds as seconds */
    latmss = latm * 60 + lats;
    lngmss = lngm * 60 + lngs;

    /*  identify inconsistent 1-degree quad numbers */
    if(latd != secor[qdeg-1][2] || lngd != secor[qdeg-1][1]){
      loc_ok = 'n';
      errtype = 'd';}

    /*  define range of lat and long for 7.5 minute quad
         nb:  7.5 min = 450 sec */
    latmssmin = ((64 - q7) / 8) * 450;
    latmssmax = latmssmin + 450;
    lngmssmin = ((64 - q7) % 8) * 450;
    lngmssmax = lngmssmin + 450;

    /*  identify inconsistent 7.5 minute quad numbers */
    if(loc_ok == 'y'){
    if(latmss < latmssmin || latmss > latmssmax ||
       lngmss < lngmssmin || lngmss > lngmssmax){
      loc_ok = 'n';
```

```
        errtype = '7';}}

      /*  define range of lat and long for 2.5 minute quad
           nb:  2.5 min = 150 sec */
      latmssmin = latmssmin + ((9 - q2) / 3) * 150;
      latmssmax = latmssmin + 150;
      lngmssmin = lngmssmin + ((9 - q2) % 3) * 150;
      lngmssmax = lngmssmin + 150;

      /*  identify inconsistent 2.5 minute quad numbers */
      if(loc_ok == 'y'){
      if(latmss < latmssmin || latmss > latmssmax ||
         lngmss < lngmssmin || lngmss > lngmssmax){
        loc_ok = 'n';
        errtype = '2';}}

      /* print results to output file */
      fprintf(outfp, "%d,%c,%c\n", wellno, loc_ok, errtype);

      /*  reinitialize output variables */
      loc_ok = 'y';
      errtype = 'n';
    }  /* end of test loop */
  }  /* end of file read loop */

  fclose(datafp);
  fclose(outfp);

}  /*  end of program test_quad.c */
```

```
/*  include.aml -- selects those records to be included in TWDB
/*  nitrate study.  Should be run after testquad.aml

/*  set up relate to gain access to well records from nitrate
/* measurement table.
relate add
  well
  twdb_wells.dat
  INFO
  wellno
  wellno
  ordered
  rw
~

tables

  /*  add items to well and measurement tables to indicate
  /*  inclusion in study, and wells with included measurements
  additem twdb_wells.dat meas 1 1 c
  additem twdb_wells.dat include 1 1 c
  additem twdb_wells.nit include 1 1 c

  /*  sort the well table by well number to speed up relate
  sel twdb_wells.dat
  sort wellno

  /*  set MEAS field to 'n' (records corresponding to wells with
  /*  included measurements will be changed later in the program.)
  move 'n' to meas

  /*  select the nitrate table and restrict the selection to
  /*  records with remarks indicating poor reliability
  sel twdb_wells.nit
  res reliability_rem = '01' or reliability_rem = '02' ~
      or reliability_rem = '03'

  /*  add to selection records with no corresponding well records
  asel wellno ne well//wellno

  /*  add to selection records for measurements from
  /*  mis-located wells
  asel well//quad_ok = 'n'

  /*  add to selection records with thresholds above
  /*  0.1 mg/l (as N) or 0.45 (as nitrate)
  asel q71850_flag = '<' and q71850_nitrate gt 0.45

  /*  add to selection records with "greater than" flag
  asel q71850_flag = '>'

  /*  add to selection records for samples prior to 1962
  asel yy_date lt 1962

  /*  mark selected records for exclusion
  move 'n' to include
```

```
  /*  invert selection and mark those records for inclusion
  nsel
  move 'y' to include

  /*  mark well records corresponding to included measurements
  move 'y' to well//meas

  /*  mark well records to be included in study
  sel twdb_wells.dat
  res quad_ok = 'y' and meas = 'y'
  move 'y' to include
  nsel
  move 'n' to include

  /*  create table include.nit
  copy twdb_wells.nit include.nit
  sel include.nit
  res include = 'n'
  purge
    y

  /*  create table include.wells
  copy twdb_wells.dat include.wells
  sel include.wells
  res include = 'n'
  purge
    y


q stop
relate drop
  wells
  ~

&return
```

```
/*  years.aml -- procedure followed to add items to info table
/*  station.mean indicating first year and last year of data
/*  reported and maximum gap in recording period.

tables

sel prec.dat
  sort station_id station_name year
  unload gap.dat station_id station_name year
  sys awk -f firstyear gap.dat > first.out
  sys awk -f lastyear gap.dat > last.out
  sys awk -f maxgap gap.dat > gap.out

define first.in
  station_id 5 5 i
  station_name 23 23 c
  firstyear 4 4 i
~
redefine
  1
  unique
  28
  28
  c
~
add from first.out

define last.in
  station_id 5 5 i
  station_name 23 23 c
  lastyear 4 4 i
~
redefine
  1
  unique
  28
  28
  c
~
add from last.out

define gap.in
  station_id 5 5 i
  station_name 23 23 c
  gap 2 2 i
~
redefine
  1
  unique
  28
  28
  c
~
add from gap.out

q stop
```

```
joinitem station.mean gap.in station.mean unique op_40
joinitem station.mean last.in station.mean unique op_40
joinitem station.mean first.in station.mean unique op_40

&return
```

```
#  firstyear -- awk script to find first reporting year in annual
#  reports from stations identified by ID number and name.
#  Expects input in form of comma-delimited fields containing
#  station ID, station name, and reporting year.

BEGIN{FS=OFS=",";
  station = "";
  staname = "";}

$1 != station || $2 != staname{print $1,$2,$3;
  station = $1;
  staname = $2;}
```

```
#  lastyear -- awk script to find last reporting year in annual
#  reports from stations identified by ID number and name.
#  Expects input in form of comma-delimited fields containing
#  station ID, station name, and reporting year.

BEGIN{FS=OFS=",";
  lastyear = 0
  station = ""
  staname = ""}

$1 != station || $2 != staname{print station,staname,lastyear;
  station = $1
  staname = $2
  lastyear = $3}

$1 == station && $2 == staname{lastyear = $3;}
END{print station,staname,lastyear}
```

```
#  maxgap -- awk script to find maximum gap in annual reports
#  from stations identified by ID number and name.  Expects
#  input in form of comma-delimited fields containing
#  station ID, station name, and reporting year.

BEGIN{FS=OFS=",";
  gap = 0;
  count = 0;
  lastyear = 0
  station = ""
  staname = ""}

$1 != station || $2 != staname{print station,staname,count;
  count = 0;
  station = $1
  staname = $2
  lastyear = $3}

$1 == station && $2 == staname && $3 == (lastyear + 1)
  {lastyear = $3;}

$1 == station && $2 == staname && $3 != (lastyear + 1)
  {gap = $3 - lastyear - 1;
    if(gap > count) count = gap;
    lastyear = $3;}

END{print station,staname,count}
```

```
/*  precmean.aml -- procedure followed to calculate mean annual
/*  precipitation reported at databsae stations from 1951 to 1980

tables

relate add
  station
  station.mean
  info
  unique
  unique
  ordered
  rw
~

sel prec.dat
  calc station//cnt_40 = station//cnt_40 + 1
  calc station//tot_40 = station//tot_40 + prec
  res year lt 1981
  calc station//cnt_30 = station//cnt_30 + 1
  calc station//tot_30 = station//tot_30 + prec

sel station.mean
  res cnt_40 ne 0
  calc mean_40 = tot_40 / cnt_40
  asel
  res cnt_30 ne 0
  calc mean_30 = tot_30 / cnt_30

q stop
&return
```

```
/*  org_int.aml
/*
/*  Calculates integral of organic material in the soil layer
/*  table for each component.

relate add
  comp
  study.comp
  info
  mapseq
  mapseq
  ordered
  rw
~

tables

sel study.comp
  sort mapseq
  calc min-org = 0
  calc max-org = 0
  calc avg-org = 0

sel study.layer

/*  caculate average organic content (pct.)
calc omm = ( omh + oml ) / 2

/*  caculate average bulk density (g/cm^3)
calc bdm = ( bdh + bdl ) / 2

/*  multiply average organic content by average bulk density and
/*  add the result to the average organic field in the component
/*  table.  (0.254 converts inches to cm, g/cm^2 to kg/m^2, and
/*  percents to decimals.)
calc comp//avg-org = comp//avg-org + omm * bdm ~
  * ( laydeph - laydepl ) * 0.254

/*  multiply max organic content by max bulk density and add the
/*  result to the max organic field in the component table.
calc comp//max-org = comp//max-org + omh * bdh ~
  * ( laydeph - laydepl ) * 0.254

/*  multiply min organic content by min bulk density and add the
/*  result to the min organic field in the component table.
calc comp//min-org = comp//min-org + oml * bdl ~
  * ( laydeph - laydepl ) * 0.254

q stop

relate drop
  comp
  ~

&return
```

```
/*  unit_avg.aml
/*
/*  Calculates average of a numerical item in the soil
/*  component table for each map unit.  (executed from
/*  within TABLES)

/*  Read the name of the component item to be averaged, the
/*  destination mapunit item, and, optionally, the definition of
/*  the mapunit item if it does not already exist in the table
&args compitem muitem add itdef:rest

relate add
  mapu
  study.mapu
  info
  muid
  muid
  ordered
  rw
~

/* option to add new item for unit average
&if [TRANSLATE %add%] = 'ADD' &then ~
  additem study.mapu %muitem% [UNQUOTE %itdef%]

sel study.mapu
  sort muid
  calc sum = 0
  /* if new item option not exercised, set item value to zero
  &if [NULL %add%] &then ~
    calc %muitem% = 0

sel study.comp
  calc mapu//%muitem% = mapu//%muitem% + ~
    %compitem% * comppct / 100
  calc mapu//sum = mapu//sum + comppct

sel study.mapu
  res sum ne 100
  list muid %muitem% sum

sel

relate drop
  mapu
  ~

&return
```

```
/*  build_quads7.aml -- builds a polygon coverage in geographic
/*  (DD) co-ordinates of the 7.5 minute quadrangles used for well
/*  numbering by the Texas Water Development Board.

/*  make this a double-precision coverage
precision double

/*  run the C program that generates the quad coordinates
&sys tx7m > tx7m.gen

/*  put the coordinates into a new polygon coverage
generate twdb_7m
  input tx7m.gen
  polys
  quit

/*  tidy up the coverage to get rid of double-listed coordinates
clean twdb_7m

/*  set up the quadrangle ID numbers used for relates to well
data.
tables
  additem twdb_7m.pat quad_7.5m 4 4 i
  sel twdb_7m.pat
  calculate quad_7.5m = twdb_7m-id
  redefine
    25
    quad_1d
    2
    2
    I
  ~
q stop

/* delete the generate data file
&sv delstat := [DELETE tx7m.gen]

&return
```

```
/*  tx7m.c -- generates 7.5 minute cells in decimal degrees (DD)
*   numbered according to the TWDB well-numbering scheme.  Output
*   is an Arc/Info generate file.  */

#include <stdio.h>

/*  set up data matrix for 1-dgree quad locations */
int data[13][4] = {
  {4, 1, -104, 37},  /* One-degree quads in Texas fall in 13    */
  {4, -1, -101, 36}, /* rows. Numbers for quads run west-to-east*/
  {7, 1, -104, 35},  /* and east-to-west in alternating rows.   */
  {10, -1, -95, 34}, /* The matrix "data"   contains 13 four-   */
  {10, 1, -104, 33}, /* column rows.  In each row, the first    */
  {14, -1, -94, 32}, /* column is the number of cells in the    */
  {13, 1, -106, 31}, /* corresponding row of one-degree quads.  */
  {12, -1, -94, 30}, /* The second column is the direction of   */
  {1, 1, -104, 29},  /* the numbering (1 for west-to-east, -1   */
  {6, 1, -101, 29},  /* for east-to-west) in the row.  The third*/
  {4, -1, -97, 28},  /* and fourth columns are the latitude and */
  {3, 1, -100, 27},  /* longitude of the northwest corner of the*/
  {1, 1, -98, 26}};  /* lowest-numbered cell  in the row.  */
                     /* There are 89 one-degree quads. */

main(){
  int dcell = 1;
  int row;
  int cnt, mcell;
  int xcnr, ycnr;
  float xctr, yctr;
  float offset = 0.0625, csize = 0.125;

  /* outer loop for rows of one-degree quads */
  for(row = 0; row < 13; row++){

    /* 1st nested loop for individual one-degree quads */
    for(cnt = 0; cnt < data[row][0]; cnt++){

      /* locate the northwest corner of the one-degree quad */
      xcnr = data[row][2] + data[row][1]*cnt;
      ycnr = data[row][3];

      /* 2nd nested loop for 7.5 minute quads */
      for(mcell = 0; mcell < 64; mcell++){

        /* center of 7.5 minute quad is a half-cell south and east
of
           the quad's NW corner.  NW corner located by integer
division
           of 7.5-minute quad number. Integer quotient is number
of rows
           down from top; remainder is number of columns over from
           edge.  */
        xctr = xcnr + (mcell%8)*csize + offset;
        yctr = ycnr - (mcell/8)*csize - offset;

        /* print cell number and center co-ordinates, followed by
           cell corner co-ordinates.  */
```

```
        printf("%d, %f, %f\n", 100*dcell + mcell + 1, xctr, yctr);
        printf("  %f, %f\n", xcnr + (mcell%8)*csize, ycnr -
               (mcell/8)*csize);
        printf("  %f, %f\n", xcnr + (mcell%8)*csize + csize,
               ycnr - (mcell/8)*csize);
        printf("  %f, %f\n", xcnr + (mcell%8)*csize + csize,
               ycnr - (mcell/8)*csize - csize);
        printf("  %f, %f\n", xcnr + (mcell%8)*csize,
               ycnr - (mcell/8)*csize -csize);

        /* print "end" to close polygon */
        printf("%s\n", "end");
      } /* end 7.5-minute quad loop */

      /* step to next one-degree quad */
      dcell++;
    } /* close one-degree quad loop */
  } /* close row loop */

  /* print "end" to close generate file */
  printf("%s\n", "end");
}
```

```
/*****************************************************************
/*
/*  count_quad.aml -- counts number of wells, number of nitrate
/*  measurements, number of nitrate measuements exceeding 0.1, 1,
/*  5, and 10 mg/l (N) thresholds for counties in Texas, based on
/*  TWDB data.
/*
/*    CALLED BY:  user
/*
/*    CALLS:  none
/*
/*    USAGE:  &r count_quad
/*
/*****************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    none
/*
/*  GLOBAL
/*    none
/*
/*****************************************************************

tables

  relate add
    quad
    counts.quad
    info
    quad_7.5m
    quad_7.5m
    ordered
    rw
  ~

  /*  initialize the counts and sort the quad data table
    sel counts.quad
    calc well_cnt = 0
    calc meas_cnt = 0
    calc dtct_cnt = 0
    calc gt1_cnt = 0
    calc gt5_cnt = 0
    calc gt10_cnt = 0
    calc dtct_prob = 0
    calc gt1_prob = 0
    calc gt5_prob = 0
    calc gt10_prob = 0
    sort quad_7.5m

  sel include.wells
  calc quad//well_cnt = quad//well_cnt + 1

  sel include.nit
  calc quad//meas_cnt = quad//meas_cnt + 1
```

```
  res nit_adj gt 0.1
  calc quad//dtct_cnt = quad//dtct_cnt + 1

  res nit_adj gt 1.0
  calc quad//gt1_cnt = quad//gt1_cnt + 1

  res nit_adj gt 5.0
  calc quad//gt5_cnt = quad//gt5_cnt + 1

  res nit_adj gt 10.0
  calc quad//gt10_cnt = quad//gt10_cnt + 1

  sel counts.quad
  res meas_cnt ne 0
  calc dtct_prob = dtct_cnt / meas_cnt
  calc gt1_prob = gt1_cnt / meas_cnt
  calc gt5_prob = gt5_cnt / meas_cnt
  calc gt10_prob = gt10_cnt / meas_cnt

  relate drop
    quad
    ~

q stop
&return
```

```
/*****************************************************************
/*
/*  count_aq.aml -- counts number of wells, number of nitrate
/*  measurements, number of nitrate  measuements exceeding 0.1, 1,
/*  5, and 10 mg/l (N) thresholds for study aquifers, based on
/*  TWDB data.
/*
/*    CALLED BY:  user
/*
/*    CALLS:  none
/*
/*    USAGE:  &r count_aq
/*
/*****************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    none
/*
/*  GLOBAL
/*    none
/*
/*****************************************************************

tables

  relate add
    aq
    counts.aq5
    info
    aqf
    aqf
    ordered
    rw
  ~

  /*  initialize the counts and sort the aquifer data table
    sel counts.aq5
    calc well_cnt = 0
    calc meas_cnt = 0
    calc dtct_cnt = 0
    calc gt1_cnt = 0
    calc gt5_cnt = 0
    calc gt10_cnt = 0
    calc dtct_prob = 0
    calc gt1_prob = 0
    calc gt5_prob = 0
    calc gt10_prob = 0
    sort aqf

  sel aq5.wells
  calc aq//well_cnt = aq//well_cnt + 1

  sel aq5.nit
  calc aq//meas_cnt = aq//meas_cnt + 1
```

```
  res nit_adj gt 0.1
  calc aq//dtct_cnt = aq//dtct_cnt + 1

  res nit_adj gt 1.0
  calc aq//gt1_cnt = aq//gt1_cnt + 1

  res nit_adj gt 5.0
  calc aq//gt5_cnt = aq//gt5_cnt + 1

  res nit_adj gt 10.0
  calc aq//gt10_cnt = aq//gt10_cnt + 1

  sel counts.aq5
  res meas_cnt ne 0
  calc dtct_prob = dtct_cnt / meas_cnt
  calc gt1_prob = gt1_cnt / meas_cnt
  calc gt5_prob = gt5_cnt / meas_cnt
  calc gt10_prob = gt10_cnt / meas_cnt

q stop
&return
```

```
/*****************************************************************
/*
/*  count_aqquad.aml -- counts number of wells, number of nitrate
/*  measurements, number of nitrate  measuements exceeding the
/*  0.1, 1, 5, and 10 mg/l (N) thresholds for 7.5-minute quads in
/*  five study aquifers in Texas, based on TWDB data.
/*
/*    CALLED BY:  user
/*
/*    CALLS:  none
/*
/*    USAGE:  &r count_aqquad
/*
/*****************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    none
/*
/*  GLOBAL
/*    none
/*
/*****************************************************************

tables

  relate add
    aq_q
      aq_quad.dat
      info
      quad_7.5m
      quad_7.5m
      ordered
      rw
    ed_q
      ed_quad.dat
      info
      quad_7.5m
      quad_7.5m
      ordered
      rw
    cw_q
      cw_quad.dat
      info
      quad_7.5m
      quad_7.5m
      ordered
      rw
    og_q
      og_quad.dat
      info
      quad_7.5m
      quad_7.5m
      ordered
      rw
    hm_q
```

```
    hm_quad.dat
    info
    quad_7.5m
    quad_7.5m
    ordered
    rw
  sr_q
    sr_quad.dat
    info
    quad_7.5m
    quad_7.5m
    ordered
    rw
~

sel aq_quad.dat
  sort quad_7.5m
  calc ebfz = 0
  calc czwx = 0
  calc ogll = 0
  calc symr = 0
  calc symr = 0
  calc aq_cnt = 0

sel ed_quad.dat
  sort quad_7.5m
  calc well_cnt = 0
  calc meas_cnt = 0
  calc dtct_cnt = 0
  calc gt1_cnt = 0
  calc gt5_cnt = 0
  calc gt10_cnt = 0
  calc dtct_prob = 0
  calc gt1_prob = 0
  calc gt5_prob = 0
  calc gt10_prob = 0

sel cw_quad.dat
  sort quad_7.5m
  calc well_cnt = 0
  calc meas_cnt = 0
  calc dtct_cnt = 0
  calc gt1_cnt = 0
  calc gt5_cnt = 0
  calc gt10_cnt = 0
  calc dtct_prob = 0
  calc gt1_prob = 0
  calc gt5_prob = 0
  calc gt10_prob = 0

sel og_quad.dat
  sort quad_7.5m
  calc well_cnt = 0
  calc meas_cnt = 0
  calc dtct_cnt = 0
  calc gt1_cnt = 0
  calc gt5_cnt = 0
```

```
  calc gt10_cnt = 0
  calc dtct_prob = 0
  calc gt1_prob = 0
  calc gt5_prob = 0
  calc gt10_prob = 0

sel hm_quad.dat
  sort quad_7.5m
  calc well_cnt = 0
  calc meas_cnt = 0
  calc dtct_cnt = 0
  calc gt1_cnt = 0
  calc gt5_cnt = 0
  calc gt10_cnt = 0
  calc dtct_prob = 0
  calc gt1_prob = 0
  calc gt5_prob = 0
  calc gt10_prob = 0

sel sr_quad.dat
  sort quad_7.5m
  calc well_cnt = 0
  calc meas_cnt = 0
  calc dtct_cnt = 0
  calc gt1_cnt = 0
  calc gt5_cnt = 0
  calc gt10_cnt = 0
  calc dtct_prob = 0
  calc gt1_prob = 0
  calc gt5_prob = 0
  calc gt10_prob = 0

/*  set aquifer flags on quads where wells located, and count
/*  the number of wells in each quad in the aquifer tables.
sel aq5.wells
  res aqf = 'EBFZ'
    calc ed_q//well_cnt = ed_q//well_cnt + 1
    calc aq_q//ebfz = 1
  nsel
  res aqf = 'CZWX'
    calc cw_q//well_cnt = cw_q//well_cnt + 1
    calc aq_q//czwx = 1
  nsel
  res aqf = 'OGLL'
    calc og_q//well_cnt = og_q//well_cnt + 1
    calc aq_q//ogll = 1
  nsel
  res aqf = 'SYMR'
    calc sr_q//well_cnt = sr_q//well_cnt + 1
    calc aq_q//symr = 1
  nsel
  res aqf = 'HMBL'
    calc hm_q//well_cnt = hm_q//well_cnt + 1
    calc aq_q//hmbl = 1

/*  Count number of measurements and exceedences of detection
/*  limit,1, 5, and 10 mg/l (N) thresholds.
```

```
sel aq5.nit
  res aqf = 'EBFZ'
    calc ed_quad//meas_cnt = ed_quad//meas_cnt + 1
    res nit_adj gt 0.1
      calc ed_quad//dtct_cnt = ed_quad//dtct_cnt + 1
    res nit_adj gt 1.0
      calc ed_quad//gt1_cnt = ed_quad//gt1_cnt + 1
    res nit_adj gt 5.0
      calc ed_quad//gt5_cnt = ed_quad//gt5_cnt + 1
    res nit_adj gt 10.0
      calc ed_quad//gt10_cnt = ed_quad//gt10_cnt + 1

  asel
  res aqf = 'CZWX'
    calc cw_quad//meas_cnt = cw_quad//meas_cnt + 1
    res nit_adj gt 0.1
      calc cw_quad//dtct_cnt = cw_quad//dtct_cnt + 1
    res nit_adj gt 1.0
      calc cw_quad//gt1_cnt = cw_quad//gt1_cnt + 1
    res nit_adj gt 5.0
      calc cw_quad//gt5_cnt = cw_quad//gt5_cnt + 1
    res nit_adj gt 10.0
      calc cw_quad//gt10_cnt = cw_quad//gt10_cnt + 1

  asel
  res aqf = 'OGLL'
    calc og_q//meas_cnt = og_q//meas_cnt + 1
    res nit_adj gt 0.1
      calc og_q//dtct_cnt = og_q//dtct_cnt + 1
    res nit_adj gt 1.0
      calc og_q//gt1_cnt = og_q//gt1_cnt + 1
    res nit_adj gt 5.0
      calc og_q//gt5_cnt = og_q//gt5_cnt + 1
    res nit_adj gt 10.0
      calc og_q//gt10_cnt = og_q//gt10_cnt + 1

  asel
  res aqf = 'HMBL'
    calc hm_q//meas_cnt = hm_q//meas_cnt + 1
    res nit_adj gt 0.1
      calc hm_q//dtct_cnt = hm_q//dtct_cnt + 1
    res nit_adj gt 1.0
      calc hm_q//gt1_cnt = hm_q//gt1_cnt + 1
    res nit_adj gt 5.0
      calc hm_q//gt5_cnt = hm_q//gt5_cnt + 1
    res nit_adj gt 10.0
      calc hm_q//gt10_cnt = hm_q//gt10_cnt + 1

  asel
  res aqf = 'SYMR'
    calc sr_q//meas_cnt = sr_q//meas_cnt + 1
    res nit_adj gt 0.1
      calc sr_q//dtct_cnt = sr_q//dtct_cnt + 1
    res nit_adj gt 1.0
      calc sr_q//gt1_cnt = sr_q//gt1_cnt + 1
    res nit_adj gt 5.0
      calc sr_q//gt5_cnt = sr_q//gt5_cnt + 1
```

```
      res nit_adj gt 10.0
        calc sr_q//gt10_cnt = sr_q//gt10_cnt + 1

  /*  Count the number of aquifers present in each quad
  sel aq_quad.dat
    calc aq_cnt = ebfz + czwx + ogll + symr + hmbl

  /*  Calculate the estimated probabilities of exceeding 0.1, 1,
  /*  5, and 10 mg/l (N) thresholds.

  sel ed_quad.dat
    res meas_cnt ne 0
      calc dtct_prob = dtct_cnt / meas_cnt
      calc gt1_prob = gt1_cnt / meas_cnt
      calc gt5_prob = gt5_cnt / meas_cnt
      calc gt10_prob = gt10_cnt / meas_cnt

  sel cw_quad.dat
    res meas_cnt ne 0
      calc dtct_prob = dtct_cnt / meas_cnt
      calc gt1_prob = gt1_cnt / meas_cnt
      calc gt5_prob = gt5_cnt / meas_cnt
      calc gt10_prob = gt10_cnt / meas_cnt

  sel og_quad.dat
    res meas_cnt ne 0
      calc dtct_prob = dtct_cnt / meas_cnt
      calc gt1_prob = gt1_cnt / meas_cnt
      calc gt5_prob = gt5_cnt / meas_cnt
      calc gt10_prob = gt10_cnt / meas_cnt

  sel hm_quad.dat
    res meas_cnt ne 0
      calc dtct_prob = dtct_cnt / meas_cnt
      calc gt1_prob = gt1_cnt / meas_cnt
      calc gt5_prob = gt5_cnt / meas_cnt
      calc gt10_prob = gt10_cnt / meas_cnt

  sel sr_quad.dat
    res meas_cnt ne 0
      calc dtct_prob = dtct_cnt / meas_cnt
      calc gt1_prob = gt1_cnt / meas_cnt
      calc gt5_prob = gt5_cnt / meas_cnt
      calc gt10_prob = gt10_cnt / meas_cnt

q stop

relate drop
  quad
  ed_quad
  cw_quad


~&return
```

```
/*****************************************************************
/*
/*  count_quad.aml -- counts number of wells, number of nitrate
/*  measurements, number of nitrate measuements exceeding 0.1, 1,
/*  5, and 10 mg/l (N) thresholds for counties in Texas, based on
/*  WUD data.
/*
/*    CALLED BY:  user
/*
/*    CALLS:  none
/*
/*    USAGE:  &r count_quad
/*
/******************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    none
/*
/*  GLOBAL
/*    none
/*
/******************************************************************

tables

  relate add
    quad
    counts.quad
    info
    quad_7.5m
    quad_7.5m
    ordered
    rw
  ~

  /*  initialize the counts and sort the quad data table
    sel counts.quad
    calc wud_meas = 0
    calc wud_dtct = 0
    calc wud_gt1 = 0
    calc wud_gt5 = 0
    calc wud_gt10 = 0
    calc wdt_prob = 0
    calc w1_prob = 0
    calc w5_prob = 0
    calc w10_prob = 0
    sort quad_7.5m

  sel nit.wrk
  calc quad//wud_meas = quad//wud_meas + 1

  res no3 gt 0.1
  calc quad//wud_dtct = quad//wud_dtct + 1

  res no3 gt 1.0
```

```
  calc quad//wud_gt1 = quad//wud_gt1 + 1

  res no3 gt 5.0
  calc quad//wud_gt5 = quad//wud_gt5 + 1

  res no3 gt 10.0
  calc quad//wud_gt10 = quad//wud_gt10 + 1

  sel counts.quad
  res wud_meas ne 0
  calc wdt_prob = wud_dtct / wud_meas
  calc w1_prob = wud_gt1 / wud_meas
  calc w5_prob = wud_gt5 / wud_meas
  calc w10_prob = wud_gt10 / wud_meas

  relate drop
    quad
    ~

q stop
&return
```

```
      PROGRAM BINO2
C*****************************************************************
C  bino2.f -- Uses a cumulative binomial probability function to
C  find confidence limits on the single-event probability for a
C  series of trials.  Built around a binomial distribution
C  approximator in the SCDFLIB fortran library, developed at
C  M. D. Anderson Cancer Center.
C*****************************************************************
C  USAGE:  bino2 confidence_level
C    Where confidence level is a number between 0 and 1.
C*****************************************************************

      REAL XN, S, GAMMA1, GAMMA2, PP, PUP, PLO, BOUND
      INTEGER SWITCH, STATUS, IDW, XNW, SW
      CHARACTER*10 ARG1
      CHARACTER*16 IDENT

C  Open input and output files.
      OPEN(7, FILE='bino2.fmt', STATUS='OLD')
      OPEN(8, FILE='bino2.in', STATUS='OLD')
      OPEN(9, FILE='bino2.out', STATUS='UNKNOWN')

C  Read the confidence level off the command line, and calculate
C  the one-sided equivalent of the two-sided confidence level.
      CALL GETARG(1, ARG1)
      READ(ARG1, '(F6.4)') GAMMA2
      GAMMA1 = (1 + GAMMA2)/2.

C  Set the switch variable for the SDFBIN function to 4, so it
C  will find the unknown binomial parameter.
      SWITCH = 4

C  Read the field widths from the Arc/Info unload format file.
      READ(7, '(3I2)') IDW, XNW, SW

C  Read an identifier, number of trials, and number of successes
C  from a line in the input file.
10    CONTINUE
      READ(8, '(A<IDW>, F<XNW>, F<SW>)', END=9999) IDENT, XN, S

C  Special case for all successes.
      IF (S .EQ. XN) THEN
        PUP = 1.0
        PLO = (1 - GAMMA2)**(1./XN)
        STATUS = 0
      END IF

C  Special case for no successes.
      IF (S .EQ. 0.) THEN
        PLO = 0.0
        PUP = 1 - (1 - GAMMA2)**(1./XN)
        STATUS = 0
      END IF

C  General case.
      IF ((S .NE. 0) .AND. (S .NE. XN)) THEN
        CALL SDFBIN(SWITCH, GAMMA1, (S-1), XN, PLO, STATUS, BOUND)
```

```
         CALL SDFBIN(SWITCH, GAMMA1, (XN-S-1), XN, PP, STATUS,
     &     BOUND)
         PUP = 1 - PP
       END IF

C  Print results.
       IF(STATUS .EQ. 0)
     &  WRITE(9, '(A<IDW>, 2(F6, X), F6.4, 2(X, F6.4))')
     &  IDENT, XN, S, GAMMA2, PLO, PUP
       IF(STATUS .NE. 0)
     &  WRITE(9, '(A<IDW>, 2(F6, X), 3(F6.4, X), I2, X, F3.1)')
     &  IDENT, XN, S, GAMMA2, PLO, PUP, STATUS, BOUND
       GOTO 10


9999   CONTINUE
       CLOSE(7)
       CLOSE(8)
       CLOSE(9)
       STOP
       END
```

```
/*****************************************************************
/*  bino_quad.aml -- calculates confidence intervals for
/*  probability of detecting nitrate in concentrations above 0.1,
/*  1, 5, and 10 mg/l.  Calls system program bino2, which
/*  calculates two-sided confidence intervals.
/*
/*  No arguments.  No variables.
/*****************************************************************

TABLES  /*  database operations

  /*  define temporary INFO tables to hold program outputs
  DEFINE dtct.temp
    QUAD_7.5M 4 4 I
    DTCT_LO 8 6 F 4
    DTCT_UP 8 6 F 4
    DTCT_GAP 8 6 F 4
  ~
  DEFINE gt1.temp
    QUAD_7.5M 4 4 I
    GT1_LO 8 6 F 4
    GT1_UP 8 6 F 4
    GT1_GAP 8 6 F 4
  ~
  DEFINE gt5.temp
    QUAD_7.5M 4 4 I
    GT5_LO 8 6 F 4
    GT5_UP 8 6 F 4
    GT5_GAP 8 6 F 4
  ~
  DEFINE gt10.temp
    QUAD_7.5M 4 4 I
    GT10_LO 8 6 F 4
    GT10_UP 8 6 F 4
    GT10_GAP 8 6 F 4
  ~

  /* select table of measurement and detection counts
  SELECT counts.quad

  /*  write text file of quad numbers, number of measurements and
  /*  number of nitrate detections (> 0.1 mg/l).
  UNLOAD bino2.in quad_7.5m meas_cnt dtct_cnt ~
    COLUMNAR bino2.fmt INIT

  /*  call FORTRAN program to calculate confidence intervals
  SYSTEM bino2 0.90

  /*  post-process output file to create INFO input file
  /*  containing  quad number, upper & lower confidence limits,
  /*  and difference between upper & lower limits.
  SYSTEM awk '{gap = $6 - $5; print $1","$5","$6","gap}' ~
    bino2.out > dtct.csv

  /*  remove bino2 input & output files
  SYSTEM yes|rm bino2.in bino2.fmt bino2.out
```

```
  /*  repeat for 1, 5, and 10 mg/l detection limits.
  UNLOAD bino2.in quad_7.5m meas_cnt gt1_cnt ~
    COLUMNAR bino2.fmt INIT
  SYSTEM bino2 0.90
  SYSTEM awk '{gap = $6 - $5; print $1","$5","$6","gap}' ~
    bino2.out > gt1.csv
  SYSTEM yes|rm bino2.in bino2.fmt bino2.out
  UNLOAD bino2.in quad_7.5m meas_cnt gt5_cnt ~
    COLUMNAR bino2.fmt INIT
  SYSTEM bino2 0.90
  SYSTEM awk '{gap = $6 - $5; print $1","$5","$6","gap}' ~
    bino2.out > gt5.csv
  SYSTEM yes|rm bino2.in bino2.fmt bino2.out
  UNLOAD bino2.in quad_7.5m meas_cnt gt10_cnt ~
    COLUMNAR bino2.fmt INIT
  SYSTEM bino2 0.90
  SYSTEM awk '{gap = $6 - $5; print $1","$5","$6","gap}' ~
    bino2.out > gt10.csv
  SYSTEM yes|rm bino2.in bino2.fmt bino2.out

  /*  write results into new INFO tables
  SELECT DTCT.TEMP
  ADD FROM dtct.csv
  SELECT GT1.TEMP
  ADD FROM gt1.csv
  SELECT GT5.TEMP
  ADD FROM gt5.csv
  SELECT GT10.TEMP
  ADD FROM gt10.csv

  /*  duplicate INFO table for addition of confidence limit data
  COPY counts.quad bino.quad

q stop  /*  exit tables for last step.

/*  expand INFO table with columns containing confidence limits.
JOINITEM bino.quad dtct.temp bino.quad quad_7.5m dtct_prob
JOINITEM bino.quad gt1.temp bino.quad quad_7.5m gt1_prob
JOINITEM bino.quad gt5.temp bino.quad quad_7.5m gt5_prob
JOINITEM bino.quad gt10.temp bino.quad quad_7.5m gt10_prob

&return
```

```
/* logfit.c
*  5/10/94
*  revised 9/14/94, changed from natural to common logs.
*  Tom Evans
*  Civil Engineering Department, University of Texas at Austin
*/
/*****************************************************************
*  logfit.c -- fits data in a two-column comma-delimited input
*  file to a series of lognormal distributions.  The first column
*  of the input file should contain a string to identify a group
*  that the data in the second column belong to.  All records
*  associated with a single group should appear on consecutive
*  lines.
*
*  The output file will contain one record for each group
*  identifier of the input file.  These records consist of the
*  identifier followed by the mean, standard deviation,
*  coefficient of variation, and regression parameters.
*
*  This program uses functions from _Numerical Recipes in C_
*  (Press et al., 1988), identified in comments as NR.
*
*  USAGE:  logfit infile outfile
*
*  NOTE ON VARIABLE NAMES:
*     floating point variable names end in 'f'
*     double-precision floating point variable names end in 'lf'
*     integer variable names end in 'i'
*     double-precision integer variable names end in 'li'
*     pointer variable names end in 'p'
*           ???fp is a pointer to a file
*           ???ip is a pointer to an integer
*           ???fpp is a pointer to an float
*
*****************************************************************/
#include <stdio.h>
#include <math.h>

#define MAXCHAR 80         /* Max acceptable input line length */
#define MAX_VECTOR_LENGTH 5000  /* Max length of data vector */

float *vector(); /* NR's variable-offset float vector creator */
void mdian1();   /* NR median-finding function */
int read_line (); /* reads a line of text from a file */
void fit_logn();  /* fits data to a lognormal distribution */
float normz();  /* normal variate for excedance probability */
float normp();  /* exceedance probability for a normal variate */
float betai();  /* NR's incomplete beta function */

/*****************************************************************
*  MAIN PROGRAM
*****************************************************************/

main(argc, argv)
  int  argc;
  char  *argv[];
{
```

```
float dataf, *datavfp; /* datum from file, data vector */
float medf, meanlogf, stdlogf, R2, t, F, Se, sigF;
                       /* calculated by fit_log function */
float Pd, P1, P5, P10;  /* model probabilities */
int ni;  /* number of records in input file */
int cnti = 1; /* number of values in data vector */
char inst[MAXCHAR + 1], idst[10], idnewst[10], tempst[10];
FILE *datafp, *outfp;

/*  usage message for careless users */
if(argc != 3){
  fprintf(stderr, "\nUSAGE:  %s infile outfile\n", argv[0]);
  exit(1);}

/*  open the data file READ ONLY */
if((datafp = fopen(argv[1], "r")) == NULL){
  fprintf(stderr, "\nunable to open data file:  %s.\n",
    argv[1]);
  exit(1);}

/*  open the output file */
if((outfp = fopen(argv[2], "w")) == NULL){
  fprintf(stderr, "\nunable to open output file:  %s.\n",
    argv[2]);
  fclose(datafp);
  exit(1);}

/* create the data vector */
datavfp = vector(1,MAX_VECTOR_LENGTH);

/*  read the first line from the input file */
if(read_line(datafp, inst, MAXCHAR) == EOF){
  printf("\nEmpty input file.  Terminating %s.\n", argv[0]);
  fclose(datafp);
  fclose(outfp);
  exit(1);}

/* extract the id and value from the line, extracting new lines
   from the data file if the line is bad  */
while(sscanf(inst, "%[^,],%f", tempst, &dataf) != 2)
  if(read_line(datafp, inst, MAXCHAR) == EOF){
    printf("\nBad input file.  Terminating %s.\n", argv[0]);
    fclose(datafp);
    fclose(outfp);
    exit(1);}
/*  put the first value in the data vector and set the first
    value of the id string.  */
datavfp[cnti] = dataf;
strcpy(idst, tempst);

while(read_line(datafp, inst, MAXCHAR) != EOF){
  if(sscanf(inst, "%[^,],%f", tempst, &dataf) == 2){

    if(strcmp(idst, tempst) != 0){
      /* when a new id is encountered,
      (1) perform the lognormal fitting for the last set of
          data, */
```

```
        fit_logn(datavfp, cnti, &medf, &meanlogf, &stdlogf,
                 &R2, &t, &F, &Se, &sigF);

        /* (2)print the results to the output file */

fprintf(outfp,"%s,%d,%.2f,%.3f,%.3f,%.4f,%.2f,%.2f,%.3f,%f,",
        idst, cnti, medf, meanlogf, stdlogf, R2, t, F, Se,
        sigF);

        /* (3)calculate probabilities for nitrate at 0.1, 1, 5, &
          10 mg levels */
        Pd = 1. - normp((-1.0 - meanlogf)/stdlogf);
        P1 = 1. - normp((0.0 - meanlogf)/stdlogf);
        P5 = 1. - normp((0.699 - meanlogf)/stdlogf);
        P10 = 1. - normp((1.0 - meanlogf)/stdlogf);

        /* (4)print the results to the output file */
        fprintf(outfp,"%.2f,%.2f,%.2f,%.2f\n", Pd, P1, P5, P10);

        /* (5)reinitialize the data vector. */
        cnti = 0;
        strcpy(idst, tempst);} /* end new id block */

      /* always add new data to the current vector */
      cnti++;
      datavfp[cnti] = dataf;}
    }

  /* do calculations and prints for the last data set */
  fit_logn(datavfp, cnti, &medf, &meanlogf, &stdlogf, &R2, &t, &F,
           &Se, &sigF);
  fprintf(outfp,"%s,%d,%.2f,%.3f,%.3f,%.4f,%.2f,%.2f,%.3f,%f,",
    idst, cnti, medf, meanlogf, stdlogf, R2, t, F, Se, sigF);
  Pd = 1. - normp((-1.0 - meanlogf)/stdlogf);
  P1 = 1. - normp((0.0 - meanlogf)/stdlogf);
  P5 = 1. - normp((0.699 - meanlogf)/stdlogf);
  P10 = 1. - normp((1.0 - meanlogf)/stdlogf);
  fprintf(outfp,"%.2f,%.2f,%.2f,%.2f\n", Pd, P1, P5, P10);

  /* close the files and go home */
  fclose(datafp);
  fclose(outfp);
}  /* end main program */

/*****************************************************************
*  function read_line
*****************************************************************/
/*  Reads a line from the input file pointed to by ifp and places
it in the string str.  Returns 0 if nothing unexpected happens.
Returns 1 if more than maxci charaters appear in the line.
Returns EOF if EOF encountered. */

int read_line(ifp, str, maxci)
  FILE *ifp;
  char *str;
  int  maxci;
{
```

```
  char c;
  int  cnti = 0, reti = 0;

  while((c = getc(ifp)) != '\n' && c != EOF){
    if(cnti <= maxci)
      str[cnti] = c;
    cnti++;}
  if(cnti > (maxci+1)){
    reti = 1;
    cnti = maxci + 1;}
  if(c == EOF) reti = EOF;
  str[cnti] = '\0';

  return reti;
}  /* end function count_file_lines */

/*****************************************************************
*  function fit_logn
*****************************************************************/
void fit_logn(vectorf, ni, medfp, meanlogfp, stdlogfp, R2, t, F,
              Se, sigF)
  float  *vectorf, *medfp, *meanlogfp, *stdlogfp, *R2, *t, *F,
         *Se, *sigF;
  int ni;
{
  float *z, *logx, *sig, probx, a, b, siga, sigb, chi2, q, df;
  int ii, ji;  /* counter, number of unique values in vector */

  /* special case for 1 value in data vector */
  if(ni == 1){
    *medfp = vectorf[1];
    *meanlogfp = log10(vectorf[1]);
    *stdlogfp = *t = *F = *Se = *R2 = *sigF = 0;
    return;}

  /* allocate vectors for values and normal variates */
  z = vector(1, ni);
  logx = vector(1,ni);

  /* sort the data vector and find its median value (sorted from
     low to high values). */
  mdian1(vectorf, ni, medfp);

  /* for each unique value in the sorted vector, caculate an
     excedance probability from Blom's formula, the corresponding
     normal variate, and the log of the value */
  ji = 1;
  for(ii = 1; ii < ni; ii++){
    if(vectorf[ii] != vectorf[ii + 1]){
      probx = ((ni-ii+1) - 0.375)/(ni + 0.25);
      z[ji] = normz(probx);
      logx[ji] = log10(vectorf[ii]);
      ji++;}
    }
  /* calculate same values for last value in the data vector */
  probx = ((ni-ii+1) - 0.375)/(ni + 0.25);
  z[ji] = normz(probx);
```

```
  logx[ji] = log10(vectorf[ii]);

  /* special case for 1 value */
  if(ji == 1){
    *meanlogfp = log10(vectorf[1]);
    *stdlogfp = *t = *F = *Se = *R2 = *sigF = 0;
    return;}

  fit(logx, z, ji, sig, 0, &a, &b, &siga, &sigb, &chi2, &q);

  *meanlogfp = -a/b;
  *stdlogfp = 1./b;

  df = ji - 2;
  if(df == 0.){
      *t = *F = *Se = *sigF = 0.;
      *R2 = 1.0;
      return;}

  *t = b / sigb;
  *F = *t * *t;
  *Se = sqrt(chi2/(ji-2));
  *R2 = 0;
  for(ii = 1; ii <= ji; ii++)*R2 += z[ii]*z[ii];
  *R2 = 1 - chi2 / *R2;
  *sigF = betai( df/2., 0.5, df/(df+*F) );

  return;
}

/*****************************************************************
*  function normz
*****************************************************************/
float normz(p)
  float p;
{
  float z, w;

  if(p >= 1.0 || p <= 0.0){
    fprintf(stderr, "range error on function normz.\n");
    z = -9999;
    return z;}

  if(p == 0.5) z = 0;

  else{
  if(p < 0.5) w = sqrt(-log(p*p));
  if(p > 0.5) w = sqrt(-log((1-p)*(1-p)));
  z = w - (2.515517+0.802853*w+0.010328*w*w)/
          (1.+1.432788*w+0.189269*w*w+0.001308*w*w*w);}

  if(p <= 0.5) return z;
  else return -z;
}

/*****************************************************************
*  function normp
```

```
*****************************************************************/
float normp(z)
  float z;
{
  float abz, prob;

  abz = fabs(z);
  prob = 1.+0.196854*abz+0.115194*abz*abz
          +0.000344*abz*abz*abz+0.019527*abz*abz*abz*abz;
  prob = 1./(2.*prob*prob*prob*prob);

  if(z<0)return prob;
  else return 1-prob;
}
```

```
/*****************************************************************
/*
/*  fit_quad.aml -- fits nitrate detections for 7.5' quadrangles
/*  to lognormal distributions.
/*
/*    USAGE:  &r fit_quad
/*
/*****************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    delstat -- status of delete operation
/*
/*  GLOBAL
/*    none
/*
/*****************************************************************

tables

  relate add
    quad
    counts.quad
    info
    quad_7.5m
    quad_7.5m
    ordered
    rw
  ~

  sel counts.quad
  sort quad_7.5m

  sel include.nit
  sort quad_7.5m
  /* don't fit quads with single measurements or no detects
  res quad//meas_cnt gt 1 and quad//dtct_cnt gt 0
  unload fit.in quad_7.5m nit_adj

  system logfit fit.in fit.out

  sel logfit.quad
  add from fit.out

  &sv delstat := [DELETE fit.in]
  &sv delstat := [DELETE fit.out]

q stop

&return
/*****************************************************************
/*
/*  gt1_plot.aml -- plots 7.5' quds shaded according to
/*  probability of detecting nitrate gt 1 mg/l.
/*
/*    CALLED BY:  user
```

```
/*
/*    CALLS:  none
/*
/*    USAGE:  &r gt1_plot <screen|ill> {illustrator file name}
/*
/*****************************************************************
/*
/*  VARIABLE LIST
/*
/*  LOCAL
/*    output -- (argument, string) "screen" directs output to
/*          display;
/*           "ill" directs output to illustrator file.
/*    filename -- (argument, string) name of illustrator file
/*
/*  GLOBAL
/*    none
/*
/*****************************************************************
&args output filename
/*  set up program environment
lineset carto.lin
markerset plotter.mrk
shadeset colornames.shd
&if %output% = 'screen' &then ~
  display 9999 size canvas 500 650
&else &if %output% = 'ill' &then &do /* illustrator block
  &if [NULL %filename%] &then &do
    &ty \using default output filename\
    &sv filename := bw.out
  &end
  &if [exists %filename%] &then &system rm %filename%
  display 1040 3
  %filename%
  &ty \sending output to ILLUSTRATOR file: %filename%\
&end /* end illustrator block
&else &do /* error block
  &type Invalid output option
  &return
&end /* end error block

/*  Begin map composition.
/*  Page size set for Apple LaserWriter page writable page area
/*  loses 0.35" left 0.47" right 0.42" top 0.42" bottom from 8.5
/*  by 11 detrmined by experiment 7/22/93 TAE
/*****************************************************************
/*
/*  set appropriate map limits for composition
/*
/*****************************************************************
maplimits 1.32 3.01 6.73 8.42 /* shifted up 1.0" from original
units page
pagesize 7.68 10.16
/* draw boundaries of writable area of dissertation page on screen
/*&if %output% = 'screen' &then &do
  linesym 101
  box 0.05 0.05 7.58 10.06 /* page boundaries
```

```
  box 1.20 0.89 6.85 8.98  /* dissertation page area, slopped
  line  1.2 1.89 6.85 1.89 /*
/*&end
/*****************************************************************
/*
/*  LL and UR corners of writable zone for dissertation pages are:
/*  (1.15, 0.84) (6.90, 9.08)
/*
/*  Allowing one inch for a title and figure number and about 0.05
/*  inches (based on experiment) all around for slop, the limits
/*  of the mapable region are:
/*    1.20 1.89 6.85 8.98
/*
/*  map contents follow.
/*
/*****************************************************************
mapextent quads_7.5
shadetype color
linesymbol 101

RELATE add
  quad
  bino.quad
  INFO
  quad_7.5m
  quad_7.5m
  ordered
  ro
~

asel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt lt 12
asel quads_7.5 poly quad//quad_7.5m ne quad//quad_7.5m
shadecolor cmy 0 0 0
polygonshades quads_7.5 1000

nsel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt ge 12
res quads_7.5 poly quad//gt1_prob ge 0.0 and quad//gt1_prob lt 0.2
shadecolor cmy 100 0 100
polygonshades quads_7.5 1000

nsel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt ge 12
res quads_7.5 poly quad//gt1_prob ge 0.2 and quad//gt1_prob lt 0.4
shadecolor cmy 39.6 19.6 80.4
polygonshades quads_7.5 1000

nsel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt ge 12
res quads_7.5 poly quad//gt1_prob ge 0.4 and quad//gt1_prob lt 0.6
shadecolor cmy 0 0 100
polygonshades quads_7.5 1000

nsel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt ge 12
res quads_7.5 poly quad//gt1_prob ge 0.6 and quad//gt1_prob lt 0.8
```

```
shadecolor cmy 0 35.3 100
polygonshades quads_7.5 1000

nsel quads_7.5 poly
res quads_7.5 poly quad//meas_cnt ge 12
res quads_7.5 poly quad//gt1_prob ge 0.8
shadecolor cmy 0 100 100
polygonshades quads_7.5 1000

asel quads_7.5 poly
polys quads_7.5

relate drop
  quad
~
/*****************************************************************
/*
/*  end map contents
/*
/*****************************************************************
&if %output% = 'ill' &then ~
  display 9999
&return


relate drop
  quad
~
&return
```

```
/*  aw_avg.aml -- calculates area weighted average values of
/*  soil parameters from STATSGO coverage onto 7.5' quadrangles.

/*  Program runs in TABLES subsystem.

/*  Establish relate environment to access mapunit and
/*  quadrangle tables.

relate add
  quad
  params.quad
  INFO
  quad_7.5m
  quad_7.5m
  ordered
  rw
  mapu
  study.mapu
  INFO
  muid
  muid
  ordered
  ro
~
/*  In quadrangle data file, clear old soil parameter values
sel params.quad
  calculate soilarea = 0
  calculate ommar = 0
  calculate thkar = 0

/*  SOILCELL is polygon intersection of quads and STATSGO mapunits
/*  select polygon table and remove non-soil polygons from
/*  calculations.
sel soilcell.pat
  reselect muid ne 'TXW'
  reselect muid cn 'TX'

  /*  sum soil area, parameter-area products
  /*  in related quadrangle data table.
  calculate quad//soilarea = quad//soilarea + area
  calculate quad//ommar = quad//ommar + area * mapu//av-av-org
  calculate quad//thkar = quad//thkar + area * mapu//avthk

/*  In quadrangle table, calculate area-weighted parameter
/*  averages by dividing summed area-parameter products by
/*  soil areas.
sel params.quad
  reselect soilarea gt 0
  calculate avsoilomm = ommar / soilarea
  calculate avsoilthk = thkar / soilarea

/*  clean up and quit
relate drop
  quad
  mapu
~
&return
```

Vita

Thomas Anders Evans, the son of Grant Evans and the late Alice H. Evans, was born on October 19, 1959 in Hot Springs, Arkansas, where he attended the Lakeside schools. Mr. Evans holds Bachelor's and Master's degrees in Mechanical Engineering from Rice University and the University of Texas at Austin, respectively.

Before returning to the University of Texas to seek a doctorate, Mr. Evans worked for the Astro/Space Division of the General Electric Company. Since January, 1995, he has been employed by the United States Army Corps of Engineers at the Hydrologic Engineering Center in Davis, California.

Permanent address: 1231 F Street, Davis, California 95616

This dissertation was typed by the author.