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.

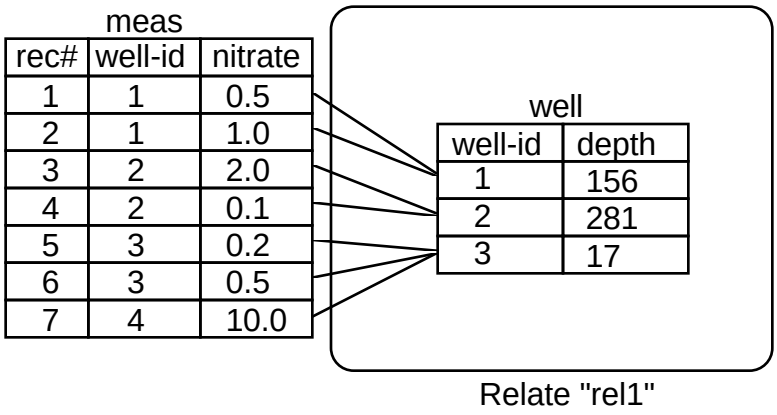

Figure 5.1 Example of a RELATE

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."

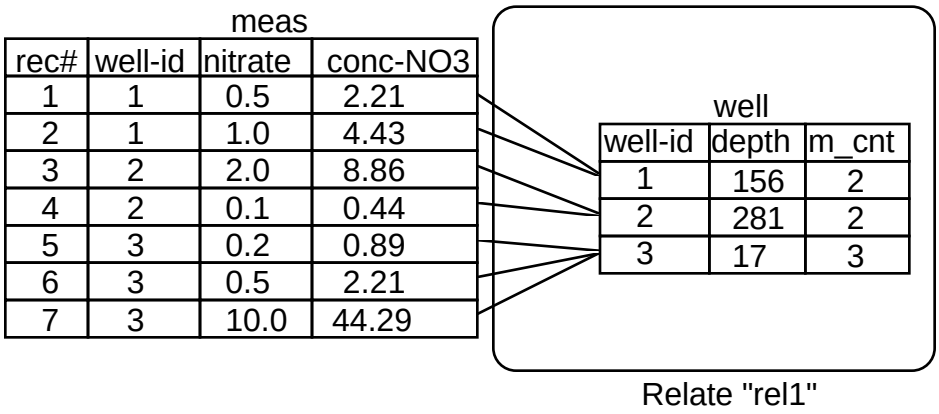

Figure 5.2 Using a RELATE to Count Measurements in Wells

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.

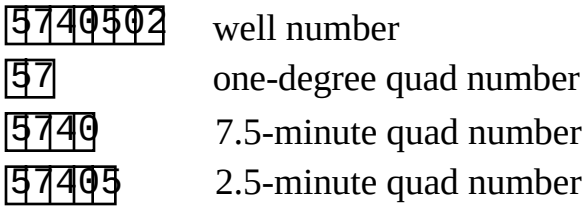

Figure 5.3 One INFO Item Redefined into Four Items

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.)

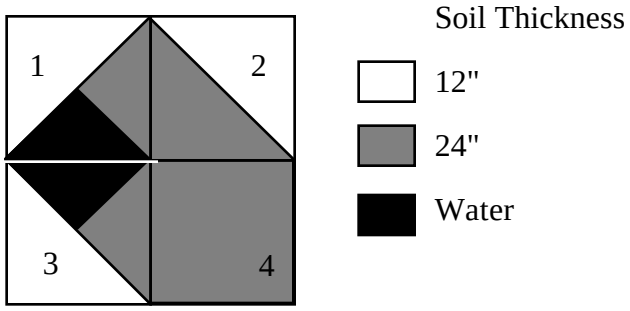

Figure 5.4 Simplified Quadrangle/Map Unit Intersection

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.