NUREG/CR-6573

Investigating Seismotectonics in the Eastern United States Using a Geographic Information System

Manuscript Completed: September 1997
Date Published: February 1998

Prepared by
J. E. Ebel, A. R. Lazarewicz, A. L. Kafka

Boston College
Weston Observatory
381 Concord Road
Weston, MA 02193-1340

E. Zurflueh, NRC Project Manager

Prepared for
Division of Engineering Technology
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555-0001
NRC Job Code W6370

DISCLAIMER

ABSTRACT

A GIS database of earthquake, geological and geophysical data was constructed to study the correlation of the seismicity with the geology and tectonics of the most seismically active areas in the central and eastern U.S. (CEUS), namely: 1. the seismically active area of the Appalachians and east coast, from Maine to Georgia, 2. the broadly active region around the New Madrid seismic zone (Illinois and Indiana to Arkansas and Mississippi), and 3. the broad area of low activity throughout Kentucky and Ohio. The GIS database analyzed consisted of the NCEER earthquake catalog, the magnetic residual field, gravity residual field, topography, crustal stress measurements, regional geology, and hydrography. Multivariate statistical analyses were carried on three different datasets: the epicenters of earthquakes with $M \geq 3.0$, the epicenter of earthquakes with $M \geq 4.5$, or the rate of seismic activity at a given geographic location of a grid cell in the study region.

The multivariate statistical analyses carried out were factor analysis, cluster analysis and discriminant function analysis. The results of the statistical analyses show that the regional geologic structures, and probably the past geologic history, do play a controlling role in determining which areas in the CEUS are seismically active. In particular, those variables that are most important for discriminating seismic and non-seismic cells appear to include: gravity residual; magnetic residual; distance to nearest river, fold, fault, and arch; length of nearest fault; and basement elevation. All of these observables can depend directly or indirectly on the existence of major basement faults, supporting the general idea that preexisting faults or other zones of basement weakness are rupturing due to the modern tectonic stress field. The lack of dependence of stress direction on location of the earthquake activity is evidence that the stress driving the earthquakes is the regional plate tectonic stress field.

CONTENTS

CONTENTS (Cont'd)

APPENDICES

FIGURES

FIGURES (Cont'd)

FIGURES (Cont'd)

TABLES

TABLES (CONT.)

Executive Summary

In the central and eastern U.S. (CEUS) the assessment of seismic hazard is problematic because the active tectonic features are generally not identified. Many ideas have been proposed to explain why earthquakes occur in the CEUS and which geologic structures are associated with the earthquakes. Earthquakes in the CEUS have been attributed to postglacial rebound, the reactivation of preexisting zones of weakness in the continental crust near extensions of oceanic fracture zones, stress concentrations associated with mafic/ultramafic plutonic masses, intersections of major structural features in the crust, reactivation of previously rifted crust, present-day faulting along Iapetan margin faults, and hydroseismicity. It is possible that many, if not all, of these hypotheses proposed to explain the spatial distribution of earthquakes in the CEUS and to identify potentially active geologic features has some merit, and it is possible that many or all of them are operative in some way in the CEUS.

In this study we constructed a GIS database of earthquake, geological and geophysical data, and we used that database to study the correlation of the seismicity with the geology and tectonics of the CEUS. Using earthquake, geological and geophysical parameters derived from this GIS database, we carried out statistical analyses to try to identify seismically active features in the CEUS. We limited our research to the most seismically active areas in the CEUS, namely: 1. the seismically active area of the Appalachians and east coast, from Maine to Georgia, 2. the broadly active region around the New Madrid seismic zone (Illinois and Indiana to Arkansas and Mississippi), and 3. the broad area of low activity throughout Kentucky and Ohio. In our statistical analyses we looked for common geological and geophysical features associated with the seismic activity on a regional basis throughout this study region.

We collected and archived many datasets for this project, and those datasets are now available in a Geographic Information System (GIS) database at Weston Observatory. The seismicity data were from the National Center for Earthquake Engineering Research (NCEER) catalog (Seeber and Armbruster, 1991). The magnetic field, gravity field, topography and gamma ray data were from the Decade of North American Geology (DNAG), Geophysics of North America CD-ROM. The stress data were procured from M.L. Zoback of the USGS (Zoback, 1992). The geology was primarily from the Geology of the Conterminous United States at 1:2,500,000 Scale -- a digital representation of the 1974 P.B. King and H.M. Beikman Map (USGS DDS-11, 1994). Hydrography and political boundaries were from a dataset provided by Environmental Systems Research Institute, Inc. The stratigraphic nomenclature was acquired from a USGS CD-ROM. Of the datasets collected and archived for this project, those datasets selected for our multivariate statistical analyses were: stress azimuth, topographic elevation, gravity residual, magnetic residual, distance to nearest fault, length of nearest fault, and distance to the nearest river. Statistical analyses were carried out on three different earthquake datasets: the epicenters of earthquakes with $M \geq 3.0$, the epicenters of earthquakes with $M \geq 4.5$, and the rate of seismic activity at a given geographic location of a grid cell in the study region.

Factor analysis was carried out on the datasets to investigate interrelationships among the various observables, trying to identify which variables seem to carry similar information and which seem to be correlated with each other in some way. Several relationships were found. Magnetic residual is positively correlated with distance to the nearest fault, negatively correlated with length of nearest fault, and perhaps also positively correlated with distance to nearest river. Topographic elevation is negatively correlated with gravity residual, a relationship expected for isostatically compensated crust. Earthquake magnitude appears to be independent of the other variables, as does the azimuth of the local maximum stress direction.

Cluster analysis takes a set of events and defines clusters of those events based on the similarities of the associated set of observables for each of the events. This study found that the observables for the earthquakes in one subregion do not cluster well with those from another region. One exception to this general result is that the Massena, New York area did cluster in two separate tests with the Charleston, South Carolina area.

In discriminant function analysis, one classifies the events into two or more categories. The method looks for a linear function of the observables that separates by the greatest amount the events in the different categories. A discriminant function was found that separated seismic cells from non-seismic cells. Unfortunately, there is a great deal of overlap in the discriminant function scores between the two groups. Thus, it is not easy to characterize a grid cell as either seismic or non-seismic based on its discriminant function score.

The results of the statistical analyses show that the regional geologic structures, and probably the past geologic history, do play a controlling role in determining which areas in the CEUS are seismically active. In particular, those variables that are most important for discriminating seismic and non-seismic cells appear to include: gravity residual; magnetic residual; distance to nearest river, fold, fault, and arch; length of nearest fault; and basement elevation. All of these observables can depend directly or indirectly on the existence of major basement faults, supporting the general idea that preexisting faults or other zones of basement weakness are rupturing due to the modern tectonic stress field. The lack of dependence of stress direction on location of the earthquake activity is evidence that the stress driving the earthquakes is regional, and is most likely a result of plate tectonic processes.

A GIS is useful in investigating the correlation of seismicity with tectonic features because it allows the machine manipulation of large geographic datasets, such as those of geological and geophysical observations. However, GIS systems are not simply turnkey operations, and they demand highly trained and knowledgeable experts to operate them properly. Critical to the successful application of a GIS is the proper planning and creation of the database. Since a strength of a GIS system is the ability to query the database to extract desired subsets of the data, the database must be planned so that the proper data subsets can be extracted in the query process. The spatial resolution of the datasets must be detailed enough to show the desired information but

low enough resolution to be efficient for GIS program execution and disk storage. The uniformity of the coverage of the data across the area of interest is also of concern, especially when mathematical analyses are to be performed on the data.

Acknowledgments

This project was sponsored and supported by the Nuclear Regulatory Commission, contract NRC-04-94-077. We are grateful to Ernst G. Zurflueh, our program manager, for his support and patience. Through the ups and downs of a project, a supportive partnership with the contract sponsor is invaluable and frequently not visible externally.

We thank Tracy Downing, of Weston Observatory for her careful and outstanding work in converting our work into a presentable form. We also thank Pat Tassia of Weston Observatory for adminstering this contract and smoothing the inevitable paperwork glitches that are ever-present. Effie Lewis, Stephan Smith and David Rose, all of Boston College, provided assistance, ideas and time that proved crucial to getting through technical difficulties.

Finally, our thanks to all the people who have labored endlessly to produce the datasets that we used. Although our primary data sources were from the USGS, we used much needed data from ESRI and from each of the New England states and from New York state. These data organizations have the uneviable job of combining disparate and varying quality data into technically consistent datasets.

1. Introduction

1.1 Seismic Hazard Assessment in the Eastern U.S.

The complete assessment of seismic hazard and corresponding seismic risk to critical facilities such as nuclear power plants requires knowing about the worst earthquakes that can affect these critical facilities. In particular, one needs to know the magnitudes of the potential earthquakes, where they might be located, how strong the ground shaking generated by these earthquakes would be at the critical facilities and how often such events occur. In essence, this means that one must know about events that have a very low probability of occurrence (for example, probabilities of 10^{-4}/year or less). To determine the possible locations of strong earthquakes, one must know which geologic structures are seismically active or are potentially seismically active.

In the central and eastern U.S. (CEUS) the assessment of seismic hazard is problematic because the active tectonic features are generally not identified. Furthermore, we have only a general idea of what kinds of structures might be seismically active and capable of generating strong, damaging earthquakes. There are many reasons for this lack of understanding. First, large earthquakes in the CEUS have estimated return times that are many hundreds to many thousands of years, while the catalog of historic earthquakes is only a few hundred years long at best. Thus, many possible earthquakes are not represented in the historic catalog. Second, the physics of the initiation of earthquake ruptures, especially of large earthquake ruptures, is not well understood. This means that we do not know what geologic characteristics give us the best indications of which structures may be active or under what conditions they may fail in large earthquakes. Third, no common geologic characteristics of seismically active structures have been identified empirically in the areas that have experienced earthquakes, especially larger earthquakes, in the CEUS. It is possible that no single geological characteristic identifies active structures but rather that a perhaps complicated set of geologic or geophysical features together may define which structures are capable of hosting strong earthquakes. Fourth, the relatively high erosion rates in the CEUS can quickly mask or erase the paleoseismological and geomorphological evidence of strong earthquakes during the past few thousands years. This is especially a problem due to the relatively low strain rates in the CEUS, since for most structures the rates of erosion may equal or exceed the rates at which fault rupture geomorphology forms. Finally, there has been relatively little geologic mapping in the CEUS aimed specifically at identifying structures that have had large earthquakes in the recent past.

Many ideas have been proposed to explain why earthquakes occur in the CEUS, and particularly to explain which geologic features may be the hosts of future large events. Investigators earlier this century suspected postglacial rebound as the source of the stress responsible for the earthquakes of the region (e.g., Leet, 1942). However, modern evidence suggests that plate tectonic forces are at work and that the plate driving force is the major contributor of stress for the earthquakes (Zoback and Zoback, 1980). Many investigators have

suggested geologic structures that might be seismically active in the present-day stress field. Sykes (1978) argues that preexisting zones of weakness in the continental crust near extensions of oceanic fracture zones are structures that are being seismically reactivated today. Kane (1977) reported that earthquakes correlate with mafic/ultramafic plutonic masses, while Talwani (1988) argued that large earthquakes are prone to occur at the intersections of major structural features in the crust. Johnston et al. (1994) evaluated a global database of large earthquakes in the stable cratonic crust and showed that most such events take place in crust that had been previously rifted, especially crust that was rifted in Mesozoic time and later. Wheeler (1996) argued that the earthquakes of eastern North America are spatially associated with basement faults at the southeastern edge of the Iapetan margin of cratonic North America and that it is these faults that are the loci of many of the earthquakes in this region. Costain et al. (1987) hypothesized that it is the presence of water in the crust down to depths of 15 km or more that creates the conditions under which ruptures initiate in the CEUS tectonic stress field. Thus, according to Costain et al. (1987), those faults capable of conducting water to significant depths in the earth are those most likely to experience future earthquakes.

It is possible that many, if not all, of these hypotheses proposed to explain the spatial distribution of earthquakes in the CEUS and to identify potentially active geologic features are operative in some way in the CEUS. Nonetheless, the problem of determining the mechanisms that cause CEUS earthquakes has not yet been solved. Testing these various hypotheses (alone and in various combinations) requires the handling and analysis of a large and varied dataset of geological and geophysical observables, such as earthquake locations and magnitudes, fault locations, fault lengths, fault orientations, fault types, fault ages, pluton locations, pluton types, gravity field maps, magnetic field maps, topography, locations of rivers, lengths of rivers, subsurface water conductivity and permeability, and age of crustal rifting. Many other geologic observables may be important as well. Thus, the problem addressed in this study is a multivariate statistical problem of potentially very large dimension.

1.2 Geographic Information Systems

The manipulation and analysis of the many data types listed above is a significant problem in database management. Furthermore, all of the information is map-based, and requires a database system capable of handling geographic data. A Geographic Information System (GIS) is a database system that has been developed specifically to handle these kinds of problems. In a nutshell, a GIS system is a method, implemented on a computer, to store, manipulate and display geographically-based information. All GIS systems consist of a database program (that allows the geographic data to be interrogated, sorted, edited, and combined) and a graphic display program (that can be used to display all or parts of the data). Information describing a particular feature on the graphic display can be called up by GIS systems. For example, if a map of earthquake epicenters is displayed, typical GIS programs allow the user to select a displayed earthquake and call up its source parameters (e.g., latitude, longitude,

magnitude, depth, focal mechanism, stress drop, time of occurrence, etc.) from the earthquake database.

The major advantage of a GIS is that it allows the machine manipulation of large geographic datasets, such as those of geological and geophysical observations. Subsets or combinations of data can be extracted and analyzed to look for hidden patterns and relationships within the data. Many different combinations of subsets of data from the database can be analyzed relatively easily. If additional data are obtained, analyses can be easily redone to evaluate the effects of the new information on the results.

GIS systems require knowledgeable experts and detailed work on the part of those experts. Critical to the successful application of a GIS is the proper planning and creation of the database. This database includes both the objects to be displayed geographically and the information linked to those objects. Since a strength of a GIS system is the ability to query the database to extract desired subsets of the data, one must plan the database so that the proper data subsets can be extracted in the query process. The spatial resolution of the datasets is also an important consideration; it must be detailed enough to show the desired information but low enough resolution to be efficient for GIS program execution and disk storage. Entering input into the database can be a time-consuming process. The uniformity of the coverage of the data across the area of interest is also of concern, especially when mathematical analyses are to be performed on the data. While there is already a great amount of data on computers in GIS systems, those databases are often not constructed with one's particular application in mind. They may not contain all of desired information, or they may contain some information extraneous to the project for which the data are to be used. Thus, it often is necessary for existing GIS databases to be edited, reformatted or otherwise modified to work best in a new project. If needed data are not in digital form, new GIS databases must be created from maps and other sources of information. These data must be either scanned into the computer, copied from non-GIS files, or entered by hand into the computer. Following this, the data must be put into the proper GIS database format. As there are many different GIS programs available on the market, sometimes one faces the task of translating a dataset from the format of one GIS software package to that of another package. By their very nature, GIS systems are not simply turnkey operations, but they demand highly trained and knowledgeable experts to operate them properly.

1.3 Purpose of this Project

In this study we constructed a GIS database of earthquake, geological and geophysical data and we use that database to study the correlation of seismicity with the geology and tectonics of the CEUS. In fact, this is, in essence, two separate studies: (1) the creation of the GIS database and (2) the statistical analysis of information from that database to try to identify seismically active features in the CEUS. We are convinced that a GIS-based approach to this

type of research is the best way to attack this problem. Correlations between seismic, geologic, potential field, and other potentially relevant data can be evaluated most effectively using a GIS system, and once in digital map form the data and analysis can easily be preserved and reused in future analyses as new earthquakes occur or new geologic discoveries are made.

In the following sections we describe more fully the problem of assessing the earthquake hazard in the CEUS, the development of our GIS database, the results of the statistical analyses we carried out, and implications of our work for identifying seismically active structures and evaluating earthquake hazard in the CEUS. GIS databases can be as large and as complex as people, machines and resources allow. To make this project manageable we limited our GIS database to datasets that were regional in extent (resolutions typically of one to a few kilometers). Furthermore, we did not study the seismicity and related features in the entire CEUS, but rather we limited our research to the most seismically active areas in the CEUS, namely: 1. the seismically active area of the Appalachians and east coast, from Maine to Georgia, 2. the broadly active region around the New Madrid seismic zone (Illinois and Indiana to Arkansas and Mississippi), and 3. the broad area of low activity throughout Kentucky and Ohio. In our statistical analyses we looked for common geological and geophysical features of the seismic activity on a regional basis throughout this study region. All of the GIS datasets cover areas much larger than these regions delineated for the statistical analyses, so it is possible for other investigators to take our GIS database and carry out analyses for regions different from the ones we studied.

It is our contention that no one study is going to "resolve" the question of which structures are active in the CEUS and what the potential for strong earthquakes is on the active structures. We view the work we report here as part of an ongoing learning process where new discoveries in the regional geology and geophysics combined with the occurrences of new earthquakes will allow us and others to continually refine our ideas as to which structures in the region are active. Since our GIS database is computer based, it can easily be updated in the future, and new analyses can be done later with the updated information.

2. The Problem of Evaluating Input for Seismic Hazard Analysis in the Central and Eastern United States

Since the tectonic processes that cause earthquakes in the CEUS are poorly understood, it is not yet possible to predict where and when the next large earthquake will occur in this region. To deal with this uncertainty, seismologists have come to rely on probabilistic seismic hazard analysis (PSHA), which involves estimating the probability that a given level of ground motion will be exceeded at a given site during a given period of time. A thorough, up-to-date review of this topic, including a discussion of how PSHA has evolved over the past few decades, was recently published by Budnitz et al. (1997).

During the past few decades, there have been two general approaches to characterizing seismological input for PSHA in the CEUS: 1) evaluating geological and geophysical features to develop "seismic source zones" and to assess their potential for generating large earthquakes, and 2) using the observed record of historical and/or network seismicity as an indicator of areas that have the potential for generating large earthquakes. We refer to the first approach as the "feature-based" approach and to the second as the "seismicity-based" approach. Most actual applications of probabilistic seismic hazard analysis in the CEUS actually involve a combination of both types of approaches, but in general the feature-based approach is exemplified by the Lawrence Livermore National Laboratory studies (LLNL, e.g. Bernreuter et al., 1989) and the Electric Power Research Institute studies (EPRI, e.g. Electric Power Research Institute, 1986). The seismicity-based approach is exemplified by the study of Frankel (1995).

An important aspect of the problem of applying PSHA in the CEUS is the concern that proposed models of seismically active features might be too subjective and capricious. Because of the uncertainty in our understanding of the tectonic processes that cause CEUS earthquakes, seismologists have legitimate differences of opinion regarding input for PSHA and regarding what combination of approaches should be used. There is also a concern that seismologists might overzealously promote their own ideas regarding seismically active features. Thus, there is a need to develop objective and scientific ways of testing hypotheses regarding seismically active features. Barstow et al. (1981) published an early attempt at addressing the problem of how to perform a statistical analysis for evaluating the relationship between seismicity and geological or geophysical features in the CEUS, but the computer technology at the time could not handle the enormous amount and different types of data required for a thorough analysis of this problem. A few examples of other attempts at formally testing hypotheses regarding the relationship between seismicity and features are: Wheeler (1985), Johnston (1989), Ebel and Spotila (1992), and Kafka and Miller (1996).

With recent developments in computer technology and the availability of GIS, it is possible to enter into a new realm of formally testing hypotheses for the feature based approach to characterizing seismological input for PSHA. It is now possible to: 1) bring together the many

different types of datasets that are relevant to the question of whether or not a particular type of feature should be considered to be seismically active, 2) analyze those data interactively, and 3) perform associated hypothesis tests. In earlier studies, gathering the required datasets was so cumbersome that by the time the data were assembled in a particular way, the seismologists were locked into a particular type of analysis. With the computer technology and GIS described in this report, however, if the researcher (or a reviewer) are dissatisfied with the way the problem is set up, it is straight forward to modify the way it is set up and try another approach. In Section 5 of this report, we show some examples of hypothesis testing based on the data we have input here, but it would be quite straightforward to test other hypotheses and/or use other input data. For example, in the analysis described in Section 5, we used the NCEER catalog (Seeber and Armbruster, 1991) for the seismicity data, but it would be quite easy to substitute a different earthquake catalog. Also, if it is later discovered that one or more earthquakes were not reported in a given earthquake catalog, those earthquakes can be added to the database, and the analysis can be run again with the new data included.

Large earthquakes are so infrequent in the CEUS that each new large earthquake has the potential for completely changing our ideas about the mechanisms that cause earthquakes in this region. Using the technology described in this report, when new earthquakes occur (or when new geological and/or geophysical features are discovered and mapped), analyses can be rerun or new analyses can be carried out with the revised input to see if any previous conclusions change.

Given the lack of definitive models to explain where and when large earthquakes will occur in the CEUS, ideas about earthquake processes change rapidly. Models may even become outmoded by the occurrence of a large earthquake between the time a research paper is accepted for publication and when it is published. It is our intention that the GIS system and associated analysis programs described here will be available for other researchers to use in testing their own hypotheses with our or their own data. Other reseachers could re-evaluate our hypothesis tests, and if desired they could use their own data as input. Hopefully, the computer technology will continue to evolve in such a way that the information system put together for this study will be compatible with future information systems, allowing future researchers to rerun our analyses with new information and new knowledge.

One question remains paramount regarding input into PSHA: What level of confidence can we place on the seismological hypotheses that are used as a basis for input to PSHA? The answer to this question demands investigations into ways to: 1) test any proposed hypotheses statistically, and 2) test the sensitivity of a given hypothesis to variations in input data. Our view of this problem is that any hypothesis that claims to explain the potential for large earthquakes must ultimately be demonstrated to explain the occurrence of actual earthquakes (including paleoseismic events). In this study we set up several ways to test whether or not a given hypothesis can be shown to be associated with the actual occurrences of earthquakes. The hypothesis tests that we have assembled in Section 5 of this report represent only one set of possible examples of how our GIS system can be used for such purposes. The computer-based

GIS infrastructure we have created for this project is flexible enough that future researchers will be able to use the system to test their own hypotheses on these datasets and/or on other datasets of their choosing.

3. Geographic Information Systems

GIS is a map-oriented database analysis tool, which can be used to analyze a broad range of geo-scientific data. At the its most basic level GIS is simply a graphical front end to a database reporting system. Although the GIS graphical front end is designed to show a two-dimensional geographical representation of data, it can be used for any two-dimensional representations. GIS integrates data organization, reporting, analysis and visualization supported by tables, charts and databases.

ESRI, producer of Arc/Info, ArcView and related products, has a helpful introduction to GIS on its Internet pages (http://www.esri.com/base/gis/index.html). Portions of this section were extracted and modified from the ESRI web pages. Although we refer to ESRI and other commercial enterprises throughout this report, we do not endorse any particular commercial enterprise or product.

3.1 Definition of GIS

There are several GIS operational systems and technical philosophies on the market, but there are some general descriptions that apply to all GIS methods. A GIS stores information as a collection of thematic layers that can be linked together by geographic information. A theme is a clearly specified class of information, e.g., rivers, topography, counties. Geographic information contains either an explicit geographic reference, such as a latitude and longitude or some other set of coordinates, or an implicit reference such as an address or road name. Geocoding is used to create explicit geographic references (multiple locations) from implicit references (descriptions such as addresses).

In general there are two independent types of graphical data: vector and raster.

Vector data. These data can be points (e.g., earthquake epicenters), lines (e.g., roads, rivers) or polygons (e.g., lakes, areal extents of cities). These *entities* are stored by geographic location, e.g., longitude and latitude. Also, there are supporting data that allow one map to register (overlay accurately) another map and annotations. These data are drawn on a screen, and GIS recognizes them as drawn. For example, a line is nothing more that a set of adjoining pixels on a screen of the same color. GIS has a feature identification mechanism that recognize these adjoining pixels as being part of the same line. A graphical representation of vector data is shown in Figure 3.1.

Raster data. These data are *images* that are recognizable to the eye and represent data that are continuous, e.g., bedrock type or satellite images. Entities are not defined in raster data. For example, GIS does not know that adjoining pixels of the same color in a raster image belong to the same entity. GIS can, however, identify pixel color and gray scale and thereby differentiate among different points and areas on the screen.

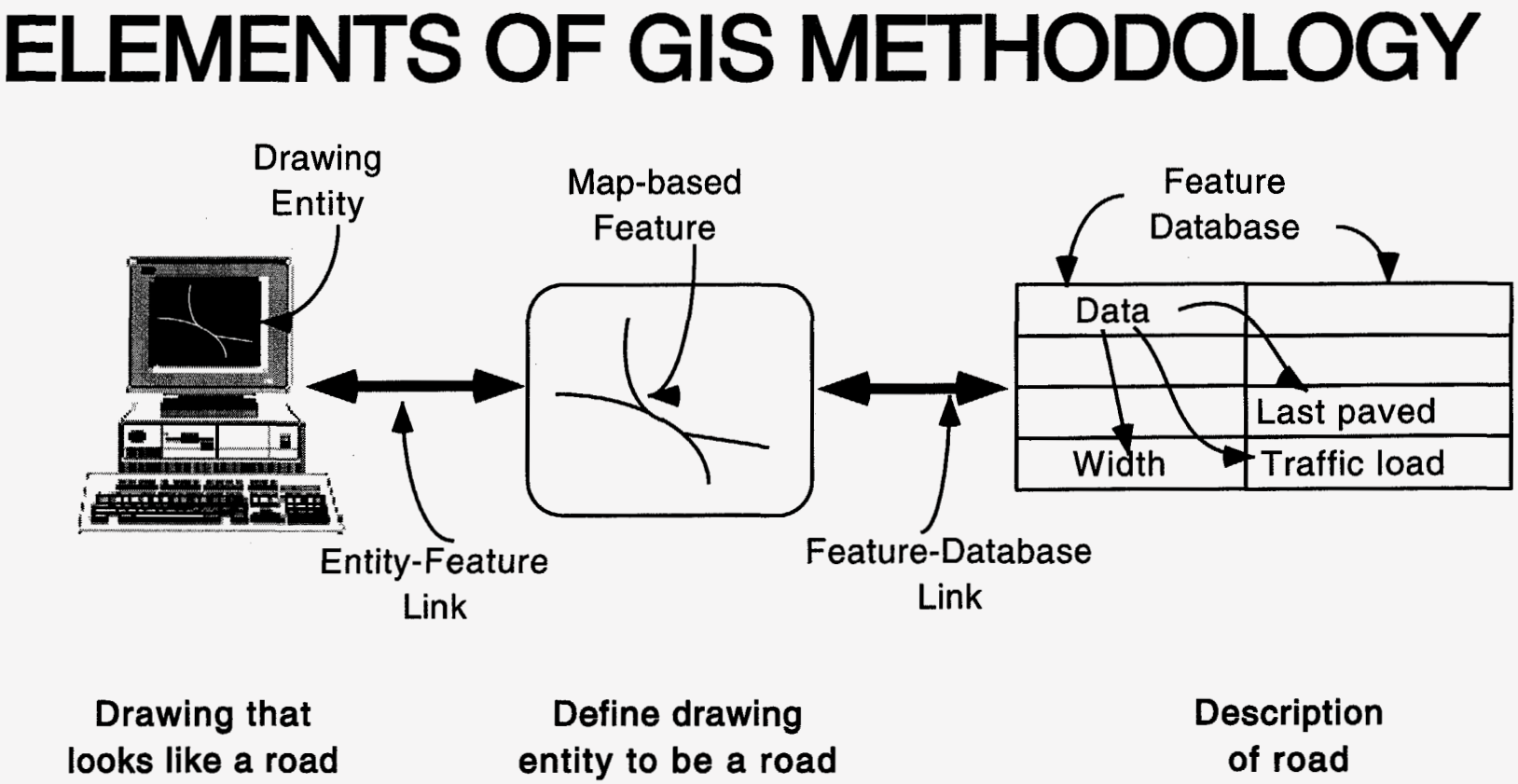

Figure 3.1. Three essential elements (computer drawing, map-based feature and feature database) of GIS. The connections between the computer image of a map elements and the database describing those map elements is the fundamental essence of the GIS methodology.

A major technical challenge to GIS designers is combining and manipulating vector and raster data. These data types are dissimilar enough that different methodologies are required to handle each type of data. However, methods for converting one to the other are crucial to good GIS design. For example, the creation of the gradient field datasets (e.g., topographic or gravity gradient) required such a data conversion. Our versions of Arc/Info and Acrview most conveniently created the gradient fields from raster data. Thus, the point fields needed to be converted to raster images before the gradients could be calculated. Also, some operations can only be made with certain data type (e.g., contours can only be computed for raster or point data -- it makes no sense to compute contours of line or polygon objects. However, the contour itself is vector line coverage data.) In addition, much data used for GIS are originally in data streams or tables that need to be converted to GIS formats. After the data are manipulated in GIS, they can be left in the GIS domain for presentation, or they can be converted back to tables. These tables can then be used in non-GIS analyses.

3.2 GIS Processes

A GIS project consists of five basic data-related processes: input, manipulation, data management, reporting or visualization, and query and analysis.

Data Input. Data useful to any project often comes in a variety of formats and representations. These data can be sequential ASCII numbers, x-y pairs, databases, paper maps, digital images, etc. This first step is to convert all the required data into a digital format that is readable by a GIS — but it need not be in GIS format at this stage. Data may be obtained from tapes, CDs, digitized from paper, etc. Many data are already available in GIS formats as well — although some GIS formats, e.g., ESRI's exchange format with file names in the form of <name>. e00 are proprietary and difficult or impossible to transfer among different vendors' software.

Data Manipulation. Once the data are digitized and available, they must be converted into formats compatible with the GIS software to be used and must be consistent with the project objectives. Typical data manipulation tasks include coordinate transformations and conversion of x-y-z data to GIS formats.

Data Management. GIS data and easily grows into very large datasets. Managing the data, (e.g., version control, data quality, and state of progress) is crucial to GIS work. Management of all these data types and of any analysis methods used is key to successful GIS design. The GIS press often quotes that 60% of all GIS projects fail due to poor design. In our work, we frequently needed to change course as we learned more about our data, as new technologies became available, and as our own experience grew. Continual change is a hallmark of GIS work — midcourse corrections are common and necessary.

Reporting and Visualization. The key objective of GIS is visualization of data in geographic terms. As GIS is comprised of a graphical front end to a database reporting system, geographic visualization and database reporting allows formal viewing and representation of the data. Since most GIS projects use complex databases, reporting results in a form that a non-GIS expert can understand is crucial to a project's success.

Query and Analysis. Once the GIS is set up and organized, database queries and technical analyses can proceed. The databases should be readable, and a method of viewing the data and analysis should be available.

3.3 Operation of a GIS

GIS is a powerful mapping and database tool that must be carefully adapted to the user's specific needs. Just as there are many ways to organize data and information, there are many ways to plan and execute a GIS analysis. Furthermore, the tools that GIS provides are geared for a wide range, but well-defined set, of tasks. Any project that requires advancing GIS beyond these well-defined tasks inevitably requires sophisticated and specialized programming resources and capabilities, in addition to the focused data and information management.

Advancing GIS for a particular geoscience project often requires GIS functions beyond what is normally available in the off-the-shelf product. These functions must be designed, written in GIS programming languages and installed into the GIS. For example, ESRI has based their products on specific languages: AML for Arc/Info and Avenue for ArcView. Proficiency in these languages requires capable programming skills. Unfortunately, even within a single vendor's product line, there are multiple languages, many of them proprietary. Consequently, several programming languages may be required to develop a project-oriented GIS.

4. Implementation of GIS for this Project

An overall objective of this project is to develop a means to evaluate the correlations between earthquakes and a broad range of geo-scientific data. In order to accomplish this objective, we needed to fuse the disparate geo-scientific datasets using GIS technology. We therefore required a system of hardware, software, and data management tools for this project. Finally, geographically organized data sets needed to be exported in tabular format for statistical analysis.

This project could be done entirely without the use of GIS, or entirely in the GIS environment, or with a combination of both. Project planning involves understanding what functions and capabilities exist and where they are. For example, formal statistical analysis is easier done outside the GIS environment, but the preparation of the data for the statistical analyses, such as computing the distance between an earthquake epicenter and the shortest distance to the nearest fault, is much easier in a GIS environment. We worked with a combination of GIS and non-GIS approaches to assemble our final datasets for the statistical analyses. The following subsections are a guide for other users who plan to implement a GIS for a project like this.

4.1 Selection of Software, Hardware and Data

In developing a computing intensive environment, such as is required for GIS, it is crucial to carefully plan and balance the design and use of computing hardware, operating system, GIS software, training and data management. We initially planned to carry out the project on the fastest PC, with enhanced memory and disk capacity, available at that time. However, delivery delays for PCs combined with uncertainties about the robustness of GIS software on a PC caused us instead to acquire a DEC Alpha 4/233 workstation with a UNIX operating system. Initially, the Alpha computer had 64 Mb of memory and 3 Gb of disk space, but we later needed to add another 4 Gb disk to handle the data. Even with the added disk capacity, we sometimes were forced to use disk space on other machines in the Department of Geology and Geophysics to handle some of the GIS datasets we acquired.

Just as we were forced to adapt hardware through the course of this work, we also required significant GIS software changes. Initially, we planned to use the CARIS GIS software package, which is designed to be used in geological applications. In the course of the study we discovered that many of the datasets that we wanted to include in our GIS could not be read by the CARIS package, so we decided to change to the ESRI Arc/Info and Arcview GIS packages. In fact, much of our real progress was the result of having received the Arcview 3 program with Spatial Analyst (released in 1996). Spatial Analyst provides the two-dimensional representations that we use, such as contours and gradient maps. This combination of programs proved to have most of the capabilities that we needed to assemble our databases.

The practicality of amassing the data for the GIS databases we wanted to use in our statistical analyses forced us into compromises and alterations of our initial plans. Initially, we thought we could collect very high resolution datasets and then edit these into the databases for our analyses. However, this proved impractical. For example, the geological and geophysical datasets available at the state level vary greatly from state to state. Often, information from the individual states does not match at state boundaries. Some of the data we wanted (such as locations of major rivers) are part of a much larger database (i.e., all water bodies), and the particular data we wanted could not be extracted from the database without an inordinant amount of effort. Some large datasets completely overwhelmed our computing resources. We collected a significant amount of data, but not all the data that we originally planned to collect. Locating, accessing and formatting "all of the data" became too difficult, so we evolved the project to one that focused on regional or national datasets that were in, or could be readily converted to, our GIS format, rather than attempting to acquire and use "all of the available data."

The management of computing resources (hardware and software) proved to be a significant effort in this project, and larger than we originally had anticipated. During the three project years, there were significant changes in the applicable hardware and software that was available on the market. Computing capability improves at a staggering rate. GIS is rapidly evolving technology, requiring significant computing resources and training. Simultaneously, most data sources are in unique data formats, which may or may not be easily readable by specific GIS software. The practical implications were that we had to alter course, several times, in our choice of hardware platform and software packages. Although these changes were ultimately beneficial for the project, they exacted a penalty by way of the project resources (mostly time) that they required.

4.2 GIS Requirements

There are several well defined areas in GIS technology that must be taken into consideration before any major development process begins. These areas are potential "show-stoppers" that are generally not included in GIS planning recommendations. They are:

1. User-friendliness. A complete GIS system designed to carry out a particular project may require hundreds of commands, many of which depend on the order that they are used. In some GIS systems, these commands are line commands, (i.e., they are to be typed in by hand) often contained in macro-language modules created by the user. In other systems, they are implemented in window menus and interfaces created by the user to execute the commands. For a particular GIS system to be useful, the project functions must be designed to be easy for users to understand and implement. They should not be so complicated that the user is overwhelmed with the number of possible options to execute.

2. Database system. Most GIS software uses databases of information already in hand. Some GIS software include rudimentary database systems. However, some GIS vendors take the viewpoint that there are many expert database systems available and that the GIS vendor's expertise is in the geographic display systems. Consequently, many do not provide a database system as part of their software. For PC-based GIS, the lack of a built-in database system is not critical since a Windows, OS/2 or Macintosh-based database system only costs a few hundred dollars. Furthermore, the PC database developers have agreed on an ODBC standard for linking programs to database programs. However, this approach does not help the UNIX users. The ODBC standard is not used in UNIX systems, and database systems for UNIX platforms are of the ORACLE and INGRES class, which are much more expensive than PC versions (several thousand dollars plus annual maintenance fees).

3. Data exchange files. ESRI, the developers of Arc/Info, have been using a specific GIS interchange format. Many GIS data files available in the United States are available as Arc/Info interchange files. Competitive vendors have been having difficulty in building a translator for these files, since the ESRI format has changed several times. Furthermore, ESRI considers that interchange file format to be proprietary and does not release it to competitors. Although several organizations have built limited translators to move files between the Arc/Info format and other formats, these vendors often do not have a formal file translator.

4.3 Project Data Types

In carrying out this project, our work involved a number of steps that needed to be carried out in sequential order. The technical plan that we pursued followed these steps:

- Identify the necessary datasets .
- Determine which GIS formats are required for which datasets.
- Determine which GIS formats may be useful for which datasets.
- Convert all relevant data sets to GIS formats.
- Create all secondary GIS datasets (e.g., gravity gradients from gravity maps).
- Perform map-oriented GIS queries.
- Convert any raster data into vector data (required for building tables).
- Build tables that can be exported for use in statistical analysis packages.
- Analyze data using formal statistical analysis techniques.
- Report results.
- Feedback to previous steps as required.

4. Implementation of GIS for this Project

We constructed a number of data tables using our GIS, and these data tables were then used in the statistical analyses reported in Section 5. The data tables that we constructed were:

Independent variables (available):

- Significant earthquakes - location, magnitude, intensity.
- Locations of grid cells over the project area.

Dependent variables (available):

- Significant earthquakes - location, magnitude, intensity (used in the grid cell matrix).
- Seismicity (number earthquakes per grid cell).
- Elevation (average).
- Elevation gradient (average).
- Magnetic field residual.
- Magnetic residual gradient (raster format).
- Gravity field residual.
- Gravity residual gradient (raster format).
- Distance to nearest fault.
- Age of local crust.
- Distance to nearest river (classified by water flow).
- Azimuth of the maximum compressive stress of the nearest stress measurement.
- Bedrock geology.

In addition, we list here additional data that we acquired that may be useful for future analyses:

Dependent variables (could be derived from available data):

- Distance to nearest fault of specified age
- Strike of nearest fault
- Total past energy release per grid cell

Finally, we list here some data that are not included in our current GIS but which would be useful for future analyses.

Dependent variables (not currently available):

- Type of crust.
- Distance to nearest pluton.
- Metamorphic value of crust.
- Type and age of nearest intrusive.
- Type of terrain.
- Amount of local crustal folding.

- Crustal thickness.
- Type of fault displacement.
- Age of last displacement on nearest fault.
- P and T axes of local focal mechanisms.

The datasets that are listed as the dependent variables "not currently available" are not available in GIS format. It is possible, however, that some of these variables (e.g., distance to nearest pluton) can be extracted from the USGS geology database. We have, for example, solved the related problem of determining the distance from a given point to the nearest fault and of extracting the name of the fault and the coordinates of the location on the fault closest to our given point. The reason that we consider as "possible" the determination of the distance to the nearest pluton is that the geology dataset is really a map of the bedrock at the 1:2,500,000 scale and has over 250 rock type entries in its database attributes. Each rock type would to be codified in some way and entered in a GIS lookup table. Then the distance to the nearest occurrence of rock types typical of plutons could then be determined. Some hand editing would probably be necessary to ensure that the rock types selected were indeed the desired plutons.

4.4 Data Processing

The volume of data that was managed in this project grew at a phenomenal rate through the course of the work. During the project we learned that GIS datasets cannot be easily provided in a "generic" form that can be immediately applied to help solve any scientific problem. In more traditional research, converting data formats for specific applications is a usually a relatively minor component in the analysis process. Analysis using GIS requires intensive efforts in data formatting and management — the data must be specifically configured for the analysis efforts at hand. Furthermore, we found that many GIS and related datasets are available as "here are all the data available." Such datasets sometimes contained massive quantities of information extraneous to our project. Consequently, project-specific datasets needed to be culled from huge datasets. This effort usually taxed and sometimes crashed our computer resources. Finding, or acquiring, more limited and therefore manageable datasets was often difficult and sometimes impossible. Even when data are obtained from a single source (e.g., USGS) they are in various formats, for different computing systems, and usually not in easily downloadable ASCII formats. Instead, CD-based data often have their own programs for a specific computer type (e.g., DOS-based PC) that will extract the desired data. These data then need to be converted into a form that can be imported into a GIS system. Our experience also suggests that many of these CD-based programs do not work properly on new computers. For example, to speed up processing they may write directly to the video hardware, whereas "windows-based" operating systems require that graphics information be written to the operating system, not the hardware. The result was that data extraction was more involved than we had planned. A final snag we encountered was that our ordered data from the USGS was delayed by the USGS for several months due to their own backlog. In summary, we encountered several

problems that appear inevitable in any project of this type. First, GIS datasets are very detailed and require much disk space. Second, project related datasets are often not in GIS format and need to be converted. Finally, existing data extraction software does not always work as expected.

Our statistical analysis tool (SPSS) required tabular ASCII data. To complete our project, we needed to create table entries both from our vector data and from our raster-based datasets. In general, we wanted values, such as distance to the nearest vector feature or identity of that feature, relative to some point location, such as an earthquake epicenter or the center of a grid cell. These values were easily calculated from the vector data, with the output stored in a table. These tables were then exported in ASCII format to SPSS.

The raster data, especially that originally generated from raster data, were more difficult to deal with. Some of our raster data were derived from vector data (i.e., the contoured magnetic field residual gradient map). Our solution to dealing with raster data was to convert it to vector data, add any attributes or other important information and then to develop and out in tabular ASCII form the required information.

5. Using GIS for Evaluation of Earthquake Hazards at Nuclear Power Plant Sites in the CEUS

5.1 Research Goal: Identification of Seismotectonically Active Structures in the CEUS

The goal of the research presented here is to help better characterize the seismic hazard to nuclear power plants in the CEUS by attempting to identify seismotectonically active structures on a regional basis. We have chosen to follow the lead of Barstow et al. (1981) and use multivariate statistical analyses of the datasets we have accumulated in our GIS database to look for evidence of active structures. As part of this work, we also are learning about the practical and inherent advantages and limitations of using GIS for multivariate statistical analysis. Thus, we intend this research not only to generate results in its own right but also to provide guidance on how to better handle and analyze such extensive datasets in future studies.

This study is timely in several ways. It makes use of a relatively new technology, GIS, that was not available when Barstow et al. (1981) carried out their study. GIS systems have become ubiquitous during the past decade, and they have grown greatly in terms of their sophisitication. Even during the course of this research project several new and improved GIS packages reached the market. The rapid evolution in the speed of computer hardware and the capacity of disk drives has further enhanced GIS capabilities. As described in the previous section, many hardware and software problems that we faced earlier in this work were later solved when new or upgraded products were acquired.

More recent datasets were incorporated into this study than were available to Barstow et al. (1981). The earthquake catalog used here, that of Seeber and Armbruster (1991), includes historical data as well as instrumental data from regional seismic network monitoring in the CEUS through 1984. Barstow et al. (1981) created a composite earthquake list from regional earthquake catalogs with events through the mid-1970s. Thus, while the Barstow et al. (1981) study was based predominantly on historic epicenters and magnitudes inferred from felt reports, this study relied to a great extent on epicenters and magnitudes determined from instrumental readings. Likewise, the geologic and potential field datasets used in this study are a later generation than those used by Barstow et al. (1981). Barstow et al. (1981) relied to a great extent on the geology published in the 1962 Tectonic Map of the U.S., while our regional geology data came from a revised digital version (Schruben et al., 1994) of the U.S. tectonic map of King and Beikman (1974). Our study used potential field and topographic data published in the mid-1980s as part of the Decade of North American Geology (DNAG) effort, whereas the potential field data of Barstow et al. (1981) came from studies published in the 1960s and early 1970s. These more recent datasets incorporate both newer and more accurate data than the older maps.

5.2 Summary of Barstow et al. (1981)

Barstow et al. (1981) used multivariate statistical analyses and spatial correlations between geologic structures and the locations of moderate to large earthquakes in an attempt to identify the relationship between geologic structures and seismicity in the CEUS. They divided the CEUS into an overlapping grid of circular cells, each of 61 km radius. For each cell they compiled a list of 60 different geological and geophysical observables, including such parameters as average fault trend and age, number of faults, topographic elevation, gravity residual, magnetic residual, intrusive type, intrusive density, number and average angle of fault-fault intersections, and number and angle of fault-arch intersections. They defined an earthquake activity rate above which they considered a cell to be seismically active and below which they considered a cell not active. They then selected 24 seismic cells and 24 non-seismic cells for their statistical analyses.

Barstow et al. (1981) carried out both univariate and multivariate analyses on their 48 selected cells. They reported that the results of their univariate analyses were not particularly useful. For the most part, they learned that faults, major rivers, and gravity anomalies ≥10 mgal were virtually ubiquitous among their selected cells.

They did, however, report more encouraging results from their multivariate analyses. They conducted discriminant function analyses, principal component analyses, factor analyses and cluster analyses on various subsets of their data. The discriminant function analyses correctly classified more than 70% of the cells as seismic or nonseismic based primarily on proximity to pre-Triassic rifts, density of faults, and density of fault/intrusive intersections, along with earthquake frequency and cumulative strain release. Other variables that they found were able to discriminate seismic from non-seismic cells were: basement elevation; distance to glacial front; distance to nearest fault; distance to nearest fold; distance to, shape of, and trend of a positive magnetic anomaly; trend of a negative magnetic anomaly; angle of a fault/fault intersection; cumulative length of faults; and average trend of faults.

Their other three analyses indicated that the geological and geophysical observables are regionally dependent, with only weak statistical similarities between different regions (i.e., the northeastern U.S., the southeastern U.S. and the central U.S.). For instance, in their cluster analysis most events clustered within a region, and there were relatively few cases where cells in one region clustered with cells in another region. Furthermore, the quantitative similarity of the clustered cells, even those from within the same region, was not high. In general, though, seismic cells tended to cluster with seismic cells, and the same was true for nonseismic cells.

While we used the same multivariate statistical procedures in this study as employed by Barstow et al. (1981), there are several important differences between our analyses and theirs. First, the variables in our datasets are not the same as in their datasets. We limited our geological and geophysical variables to those that we were able to obtain in digital form, either as a GIS

dataset directly or as a digital file that we could convert to a GIS dataset. Thus, we were not able to include many variables (e.g., fault age, seismic delay time, distance to nearest fold, density of intrusive, and shape of intrusive, among others) used by Barstow et al. (1981). On the other hand, because we could query our GIS and derive large tables of data with relative ease, we were not limited to just a few tens of datapoints (i.e., 48 cells) as were Barstow et al. (1981), who were forced to accumulate all of their statistical data by hand from paper maps. Thus, we could do statistical tests on either actual earthquake locations (even thousands of locations if desired) or on any number of spatial cells from the region. Furthermore, the GIS infrastructure that we have built for this project would allow future researchers to include many of the variables used by Barstow et al. (1981), provided that the information could be converted to digital form (either by scanning or hand digitizing).

5.3 Description of Datasets Used in This Study

We collected and archived many datasets for this project, and those datasets are now available at Weston Observatory. Of those datasets, only a subset are directly relevant to the task at hand, i.e. to investigate patterns in the data that could be diagnostic of where strong and damaging earthquake could occur in the CEUS.

The following datasets were compiled as part of this project and are now available at Weston Observatory. Detailed information about these datasets is given in Appendix A.

(1) Seismicity - We obtained a number of earthquake catalogs that cover various parts of the study area. For the statistical analysis, we used the National Center for Earthquake Engineering Research (NCEER) catalog (Seeber and Armbruster, 1991), which we downloaded from the internet.

(2) Magnetic Field - We collected two datasets of the magnetic field: the National Geophysical Data Grids for the Conterminous United States (USGS DDS-9, 1993), and the Decade of North American Geology (DNAG), Geophysics of North America CD-ROM. For the statistical analysis, we used the DNAG magnetic data.

(3) Gravity Field - The gravity data for this study are from the same datasets as the magnetic field data (USGS DDS-9, 1993 and DNAG). As in the case of the magnetic field, we used the DNAG data for the statistical analysis.

(4) Topography - The topography data for this study are from the same datasets as the magnetic field and gravity data. As in the case of the magnetic and gravity fields, we used the DNAG data set for the topography observable in the statistical analysis.

(5) Stress - The two data sets we obtained for crustal stress measurements are from the DNAG Geophysics of North America CD-ROM and via e-mail from M.L. Zoback of the USGS (Zoback, 1992). For the statistical analysis, we used the data provided by M.L. Zoback.

(6) Geology - The geological data for this project are primarily from the Geology of the Conterminous United States at 1:2,500,000 Scale — A digital representation of the 1974 P.B. King and H.M. Beikman Map (USGS DDS-11, 1994). This dataset includes faults at a regional scale. Additional data on regional faults are available in Weston Observatory archives. The fault data that we used for the statistical analyses are derived from the Geology of the Conterminous United States digital map. No other geological data were used for the statistical analysis described in the later parts of this report.

(7) Hydrography - Hydrography data archived at Weston Observatory for this project are from the 1:100,000 Scale Digital Line Graph (DLG) Data, Hydrography and Transportation (USGS Geo Data, 1993) and from a hydrography data set provided by ESRI. The USGS data are stored on a CD-ROM and include states in the Mississippi Valley, Northern Great Lakes and East Coast. For the hydrography observables used for the statistical analyses described in this report (distance to nearest river and distance to nearest drainage), we used the ESRI data.

(8) Political Boundaries - We acquired the USGS dataset of political boundaries for the Conterminous United States. These data are stored on a CD-ROM, and they provide (at 1:2,000,000 scale) political boundaries, administrative boundaries, streams, water boundaries, hypsography, roads and trails, railroads, and cultural features. These data were not used for the statistical analyses. The political boundaries shown on the maps for this report are from a dataset provided by ESRI.

(8) Gamma-Ray Data - The gamma-ray data we acquired for this project are from the same datasets as the gravity, magnetic, and topography data (USGS DDS-9, 1993 and DNAG). The gamma-ray data were not used in the statistical analysis.

(9) Stratigraphic Nomenclature - We acquired the Stratigraphic Nomenclature Databases for the United States, its possessions and territories (USGS DDS-6, 1996). The data are stored on a USGS CD-ROM. These data were not used in the statistical analysis.

(10) Massachusetts GIS Data: We obtained the complete GIS data set for Massachusetts as developed by the state's Mass GIS office. The Mass GIS data are stored on a CD-ROM. None of the Mass GIS data were used in the statistical analyses.

Of the datasets collected and archived for this project, those that contain information relevant to the problem we are addressing are the seismicity, magnetic field, gravity field, topography, stress, geology and hydrography datasets. Of those datasets, we selected the following observables for our multivariate statistical analyses: stress azimuth, topographic

elevation, gravity residual, magnetic residual, distance to nearest fault, length of nearest fault, and distance to the nearest river, as described in Section 5.4. Maps of each of these datasets are shown in Figures 5.1- 5.8. Included on those maps are a delineation of the area within which the statistical analyses were conducted.

Table 5.1

Datasets Used for Statistical Analyses

Dataset	Source	Comments
Seismicity	NCEER (from Internet)	1924-1984, m≥3
Gravity	DNAG CD ROM	Units: Mgals Grid Spacing (lat): 2.5 min Grid Spacing (lon): 2.5 min
Magnetics	DNAG CD ROM	Units: Gammas Grid Spacing (lat): 2.5 min Grid Spacing (lon): 2.5 min
Topography	DNAG CD ROM	Units: Meters Grid Spacing (lat): 5 min Grid Spacing (lon): 5 min
Stress	Zoback (1992)	Used stress azimuth
Geology	USGS DDS-11 (1994)	Used distance to nearest fault and length of nearest fault
Hydrography	ESRI	Used distance to nearest river

The seismicity, gravity, magnetics, topography and stress data are for the geographic region: 25°N ≤lat≤50°N, 110°W≤lon≤65°W.

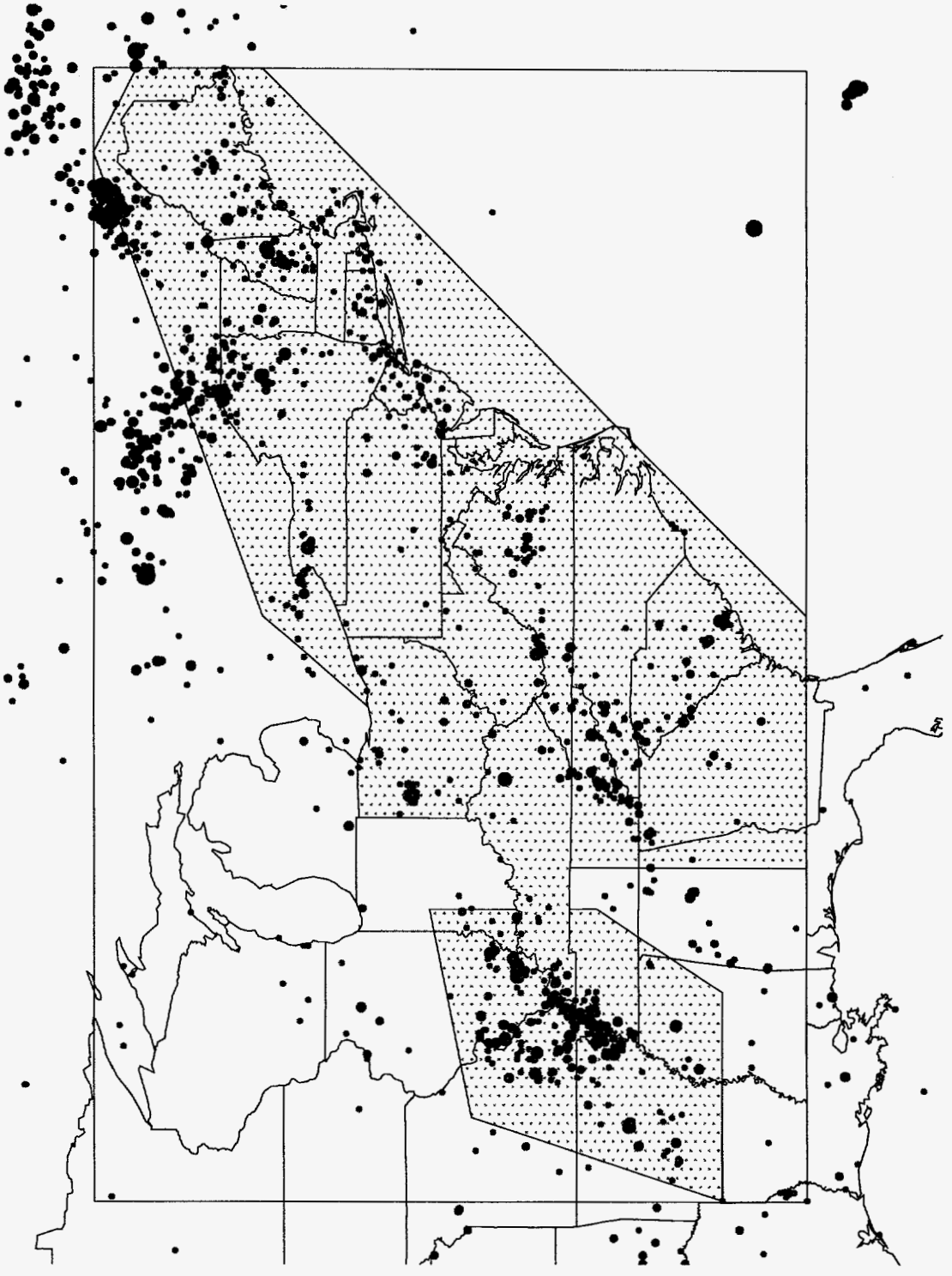

Figure 5.1. GIS representation of seismicity. Dot size proportional to magnitude (7 steps, magnitude 3.0 to 6.6). Dotted area indicates the region where the statistical analysis was performed.

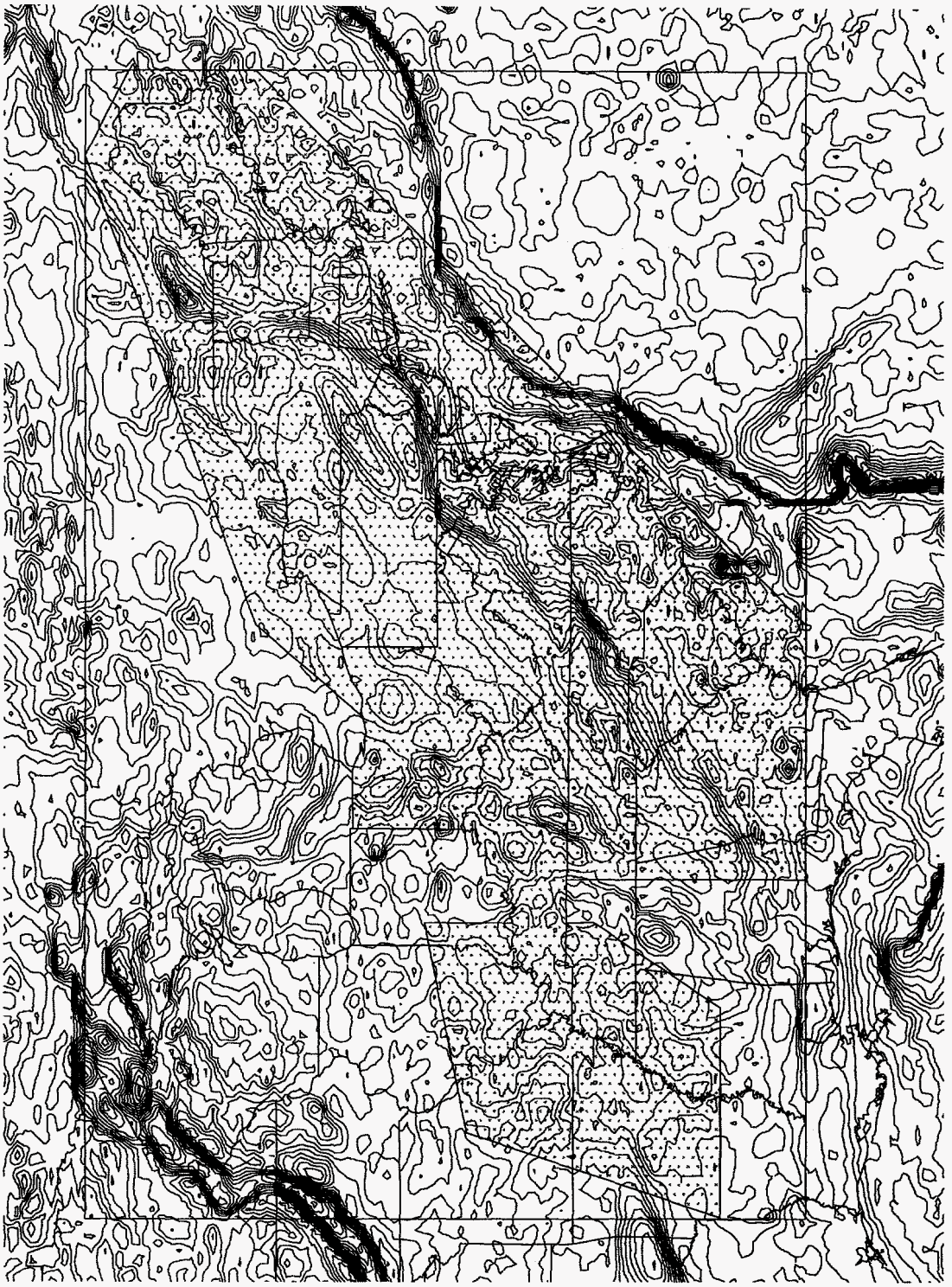

Figure 5.2. GIS representation of the gravity field residual. Dotted area indicates the region where the statistical analysis was performed.

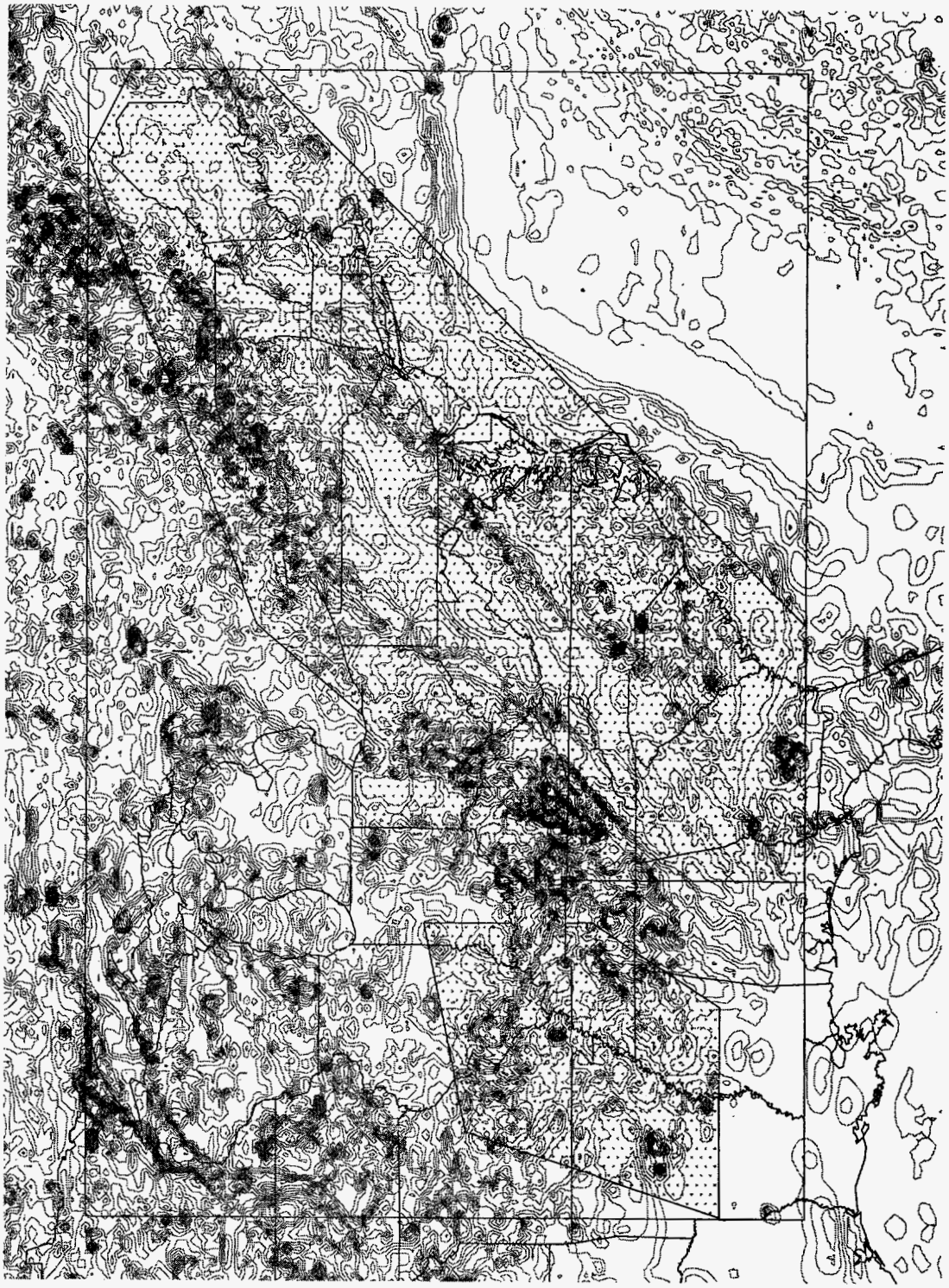

Figure 5.3. GIS representation of the magnetic field residual. Dotted area indicates the region where the statistical analysis was performed.

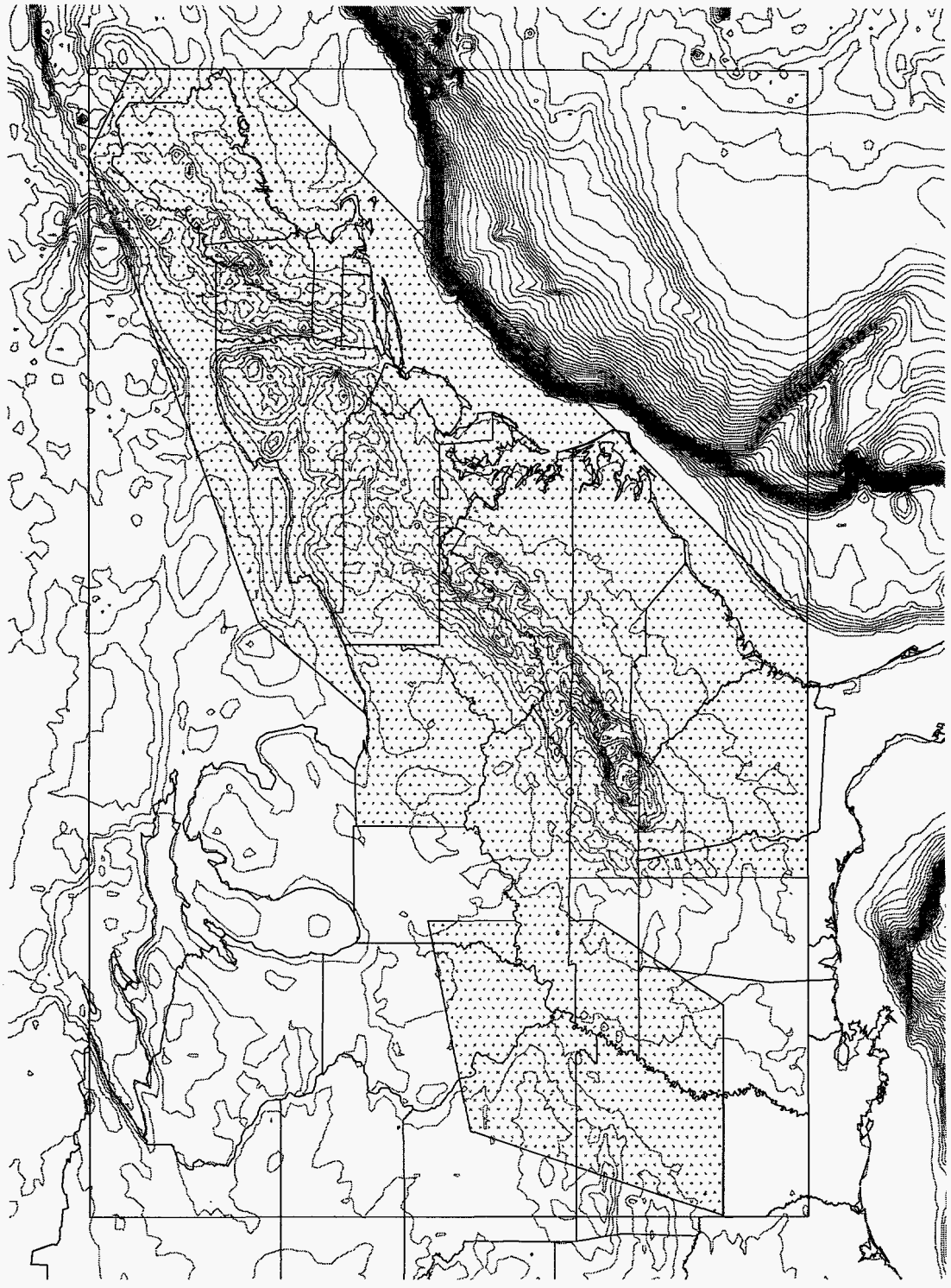

Figure 5.4. GIS representation of topography. Dotted area indicates the region where the statistical analysis was performed.

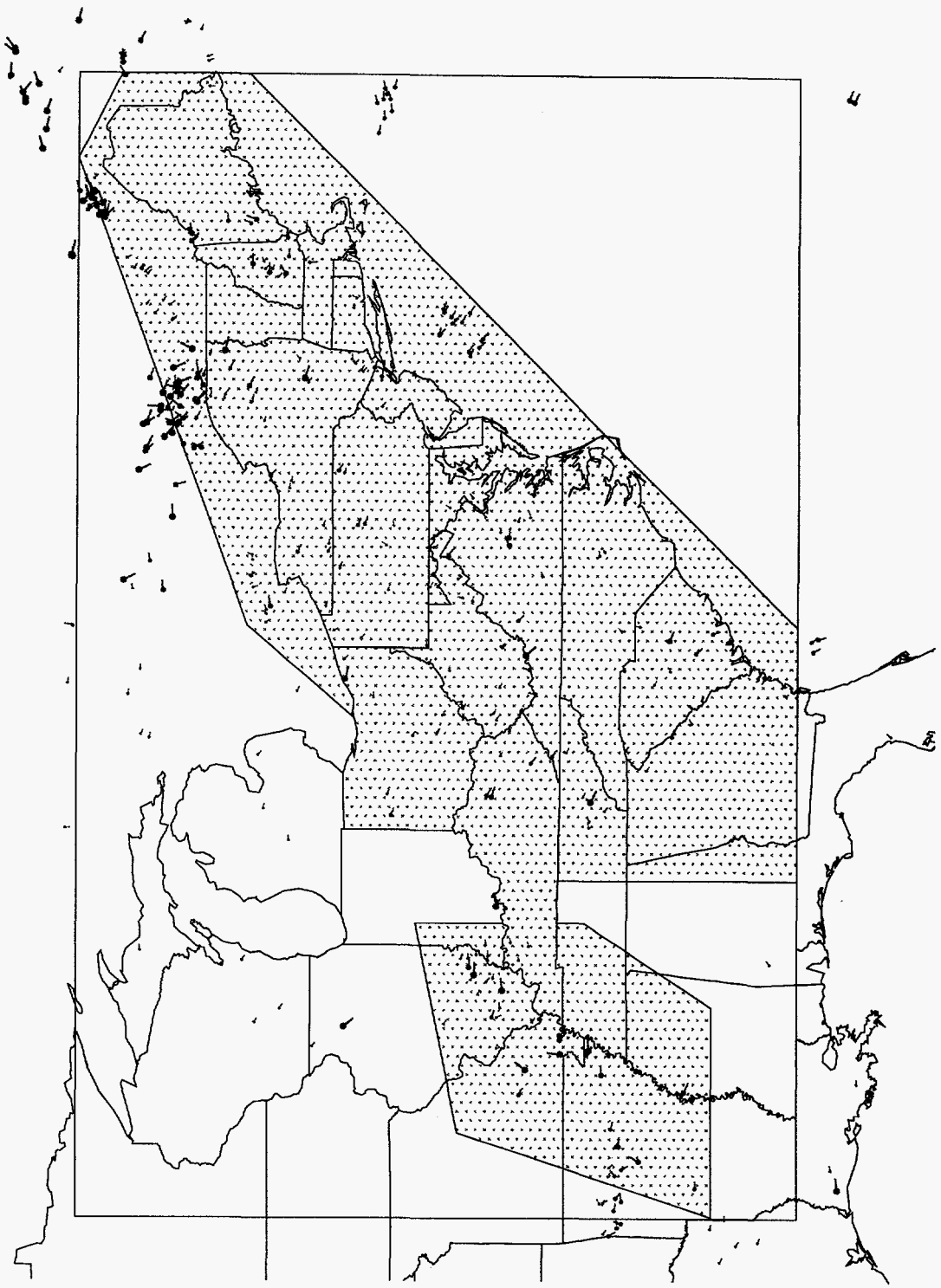

Figure 5.5. GIS representation of the maximum principal stress data. Dotted area indicates the region where the statistical analysis was performed.

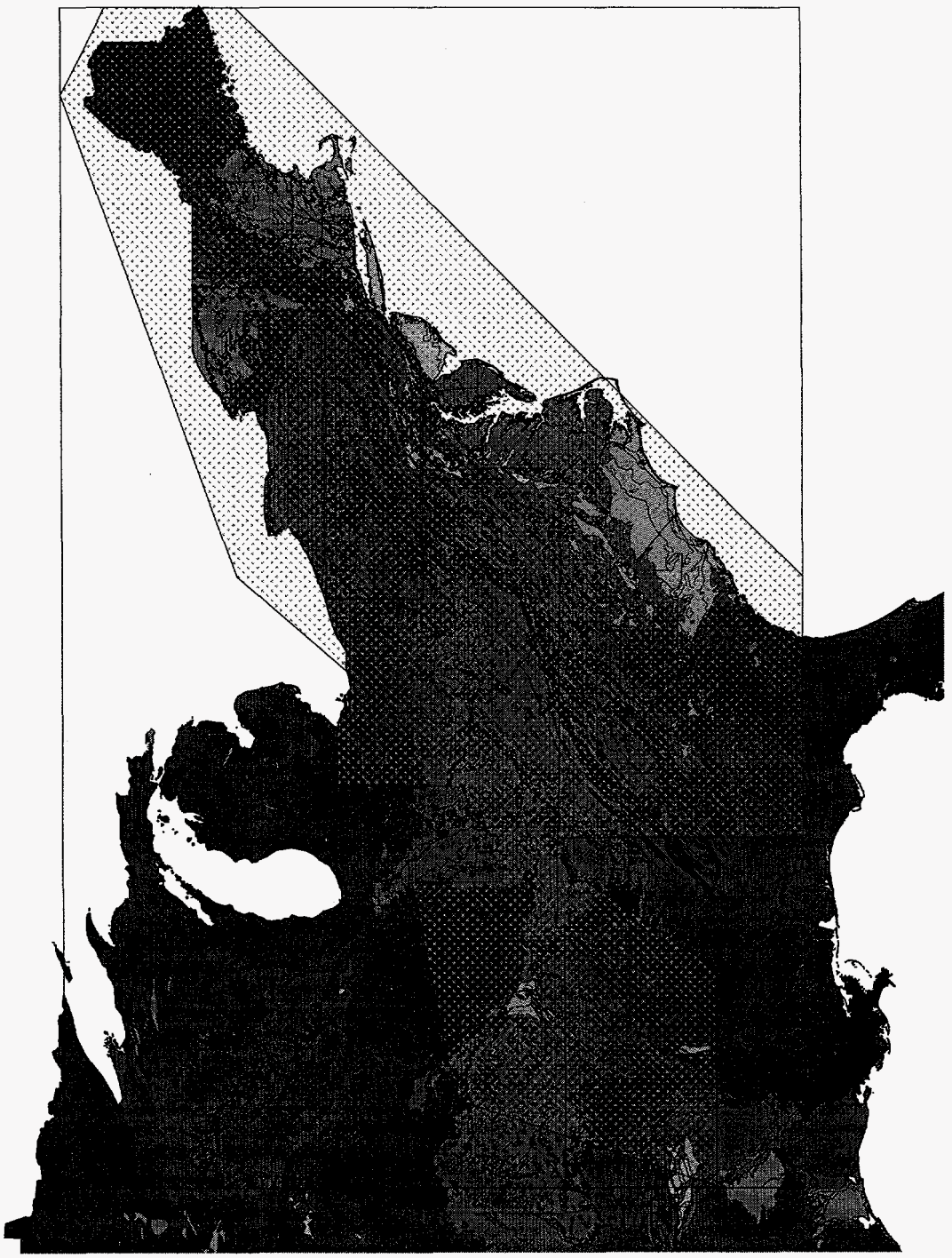

Figure 5.6. GIS representation of the USGS regional bedrock geology. Dotted area indicates the region where the statistical analysis was performed.

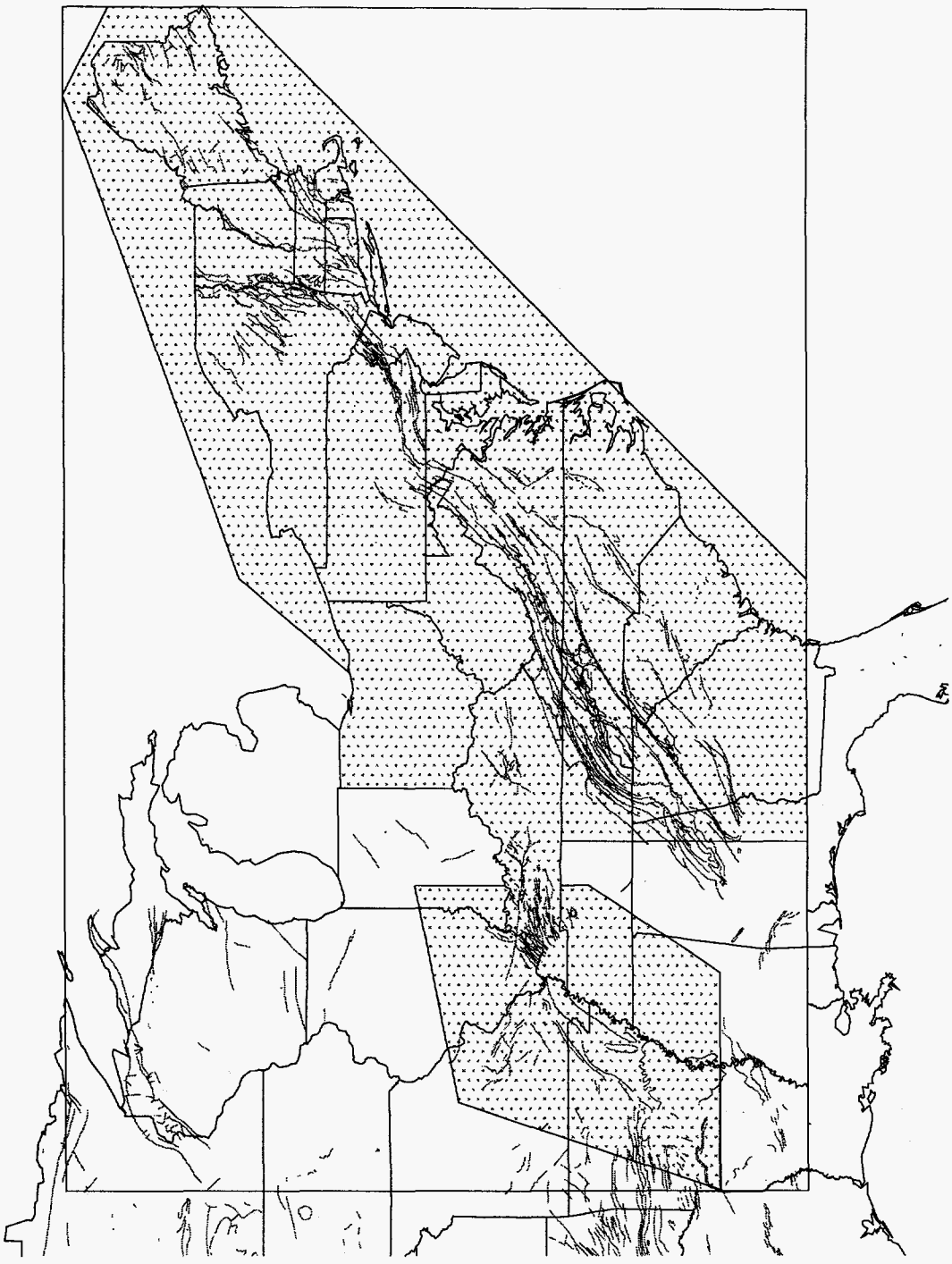

Figure 5.7. GIS representation of the USGS regional fault data. Dotted area indicates the region where the statistical analysis was performed.

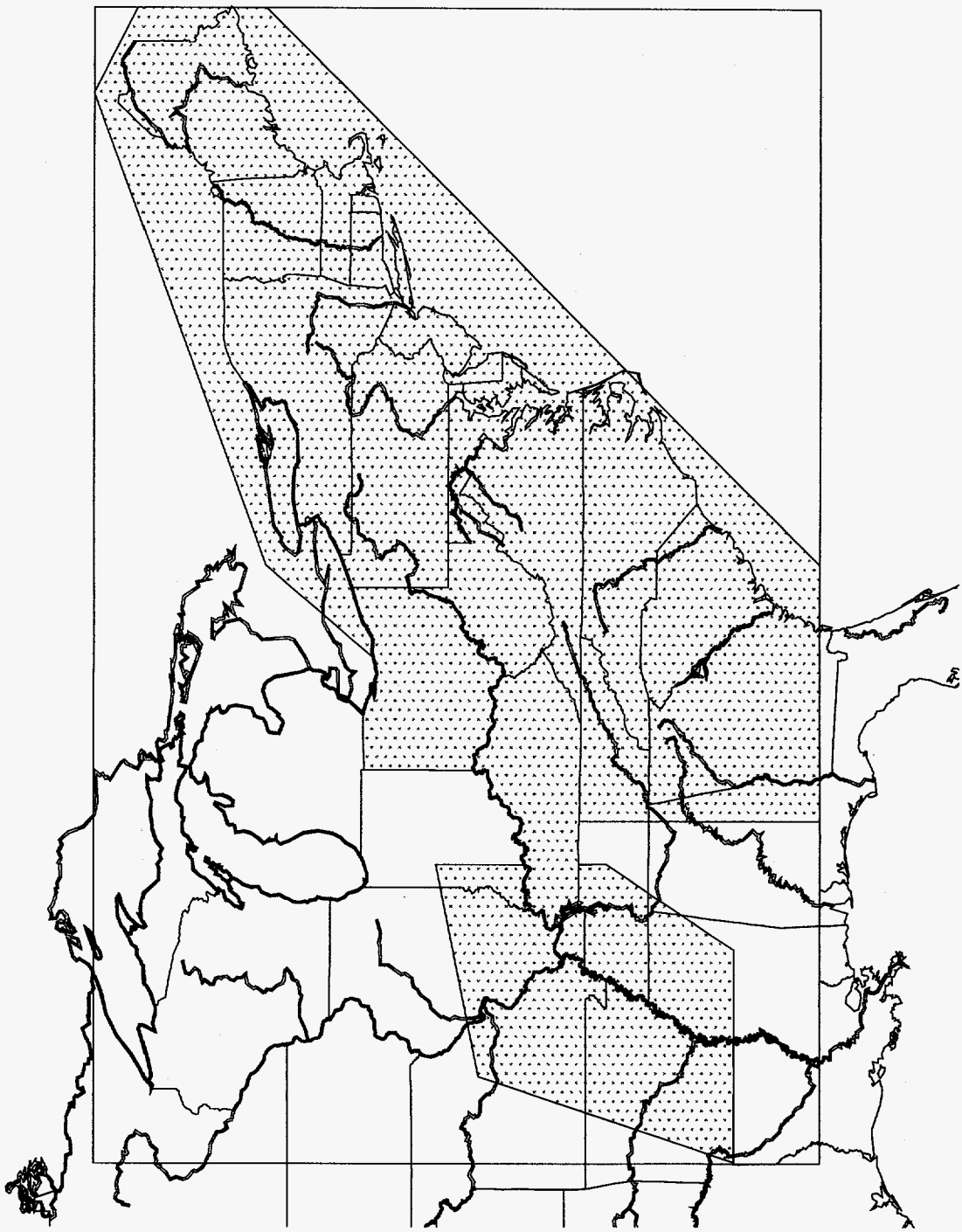

Figure 5.8. GIS representation of the major rivers. Dotted area indicates the region where the statistical analysis was performed.

5.4 Statistical Analyses

The purpose of the statistical analyses is to identify patterns in the data that could be diagnostic of localities where strong and damaging earthquakes could occur. Such patterns may be based on a single observable (e.g., intersecting faults or proximity to mafic/untramafic plutons), or they might be based on a perhaps complicated set of observables (for example, proximity to a major river combined with low elevation, negative gravity anomaly and a nearby fault of Triassic age or younger). The former are often inferred from simple charts or regressions of the data, while the latter demand that some type of multivariate analysis be carried out. Since the GIS allows us to accumulate any combination of geological and geophysical observables from our assembled datasets, it is natural that we use the GIS technology to attempt to find multivariate indicators of those localities that may be prone to strong earthquakes.

We have chosen to carry out three different multivariate statistical analyses of the data for this study. These are factor analysis, cluster analysis and discriminant function analysis (e.g., Davis, 1986). Each method assumes that one has a set of events that have associated with them a set of observed variables. In our study an event is the epicenter of an earthquake with $M \geq 3.0$, the epicenter of an earthquake with $M \geq 4.5$, or the geographic location of a grid cell in the study region. The observed variables, which we also call the observables or observations in this study, is a set of geological and geophysical observations (or calculations based on observations) associated with an event location. For instance, the topographic elevation, magnetic field residual, and the distance to the nearest major river are three of the observed variables, while the magnetic field gradient is an observable calculated from the magnetic field residual observations; the full set of geological and geophysical observables that we considered for our analyses are listed in Table 5.2. Included in that table are comments describing in which statistical analyses, if any, a particular observable was used. Some observables were not used in any statistical analysis, either because they could not be easily parameterized, because they were similar to other variables, or because their resolution was too low to give meaningful results.

Each of the three multivariate analyses used in this study looks for a different pattern in the data. In factor analysis, one seeks interrelationships among the various observables for a set of events, trying to identify which variables seem to carry similar information or which seem to be correlated in some way. Redundant information among the observed variables can be identified, as can linear relationships among combinations of the variables. In this study, we used factor analysis to look for correlations among the observables as well as to seek any pattern that may depend on event size. The approach we used first found the eigenvectors, or factor loadings, for the most significant eigenvalues from the covariance matrix of the data. The eigenvectors were then rotated using the varimax method (Davis, 1986) to either maximize or minimize the contribution of each variable to each eigenvector. The varimax method seeks out the smallest combination of variables that contribute significantly to each eigenvector (or factor).

Table 5.2
Observed Variables Considered for Statistical Analyses

Variable	Analysis Where Used
Event magnitude	Factor analysis of event datasets
Number of events in grid cell	Determined cell type in grid cell analysis
Maximum event magnitude in grid cell	Not used in any statistical analysis
Topographic elevation	All analyses
Topographic horizontal gradient	Not used in any statistical analysis
Greatest principal stress azimuth of nearest stress measurement	All analyses
Gravity field residual	All analyses
Gravity field horizontal gradient	Not used in any statistical analysis
Magnetic field residual	All analyses
Magnetic field horizontal gradient	Not used in any statistical analysis
Distance to nearest fault	All analyses
Length of nearest fault	All analyses
Description of nearest fault	Not used in any statistical analysis
Distance to nearest river	All analyses
Distance to nearest drainage	Not used in any statistical analysis
Area of nearest lake	Not used in any statistical analysis
USGS geologic unit number	Not used in any statistical analysis
USGS geologic rock type	Not used in any statistical analysis

Cluster analysis takes a set of events and defines clusters of those events based on the similarities of the associated set of observables for each of the events. Thus, groups of events with similar observables can be identified, as can events with observables that are dissimilar from those of the other events. We used cluster analysis in this study to look for two possible relationships among the data: those event locations that are similar to places where events with $M \geq 4.5$ have occurred and those grid cell locations that are similar to cell locations that show higher levels of seismic activity for all magnitudes. Our cluster analysis employed a hierarchical approach in which a step-by-step process is followed to progressively build up clusters. Initially, only the most closely related events or clusters of events are clustered, while later in the analysis more distantly-related clusters are created until all of the events have been included in a final cluster. A squared Euclidean norm of the variables is used to calculate the similarity of each pair of events or clusters. Each time a cluster is created or modified, it is represented by the mean values of the variables of the events in the cluster.

In discriminant function analysis, one classifies the events into two or more categories. The method looks for a linear function of the observables that separates by the greatest amount the events in the different categories. Once the discriminant function is defined, it can be applied

to events (either ones in the existing dataset or to new events) to assess to which category they belong. In this study, we applied discriminant function analysis to the epicenter dataset and to the grid cell dataset. In the epicenter dataset we sought a function that discriminates between epicenters with M≥4.5 and those with M<4.5. In the the grid cell dataset, we looked for a discrimant function that discriminates between cells with one or more earthquakes of M≥3.0 and cells with no earthquakes of M≥3.0. In both of these tests we attempted to find a discriminant function that would tell us which localities have observables that suggest they could be the sites of seismic activity and perhaps larger earthquakes in the future.

The factor and discriminant function analyses assume that the data are distributed in a multivariate gaussian manner. Furthermore, for all of the analyses the data should have similar means and standard deviations; otherwise, if one or a few variables have much larger numerical values than the other variables, the few large variables will heavily bias the final results. In our study we determined univariate histograms for each variable from among all of the events. Those variables which deviated significantly from a univariate gaussian distribution were reparameterized by their base 10 logarithm in an attempt to bring them closer to a gaussian shape. The was done in an attempt to bring the entire dataset closer to a multivariate gaussian distribution. We also created a parallel dataset where each variable was normalized with zero mean and unit standard deviation. The normalized data were intended to eliminate the bias in the multivariate analyses resulting from variables with large absolute values. We ran analyses on both the original unnormalized dataset and on the normalized data.

5.5 Data Histograms

Our first step in the data analysis was to see how closely the observables in our dataset appeared to fit the assumption of a multivariate gaussian distribution. We did this by creating histograms showing the univariate distribution of each variable in the analysis for all of the events. Strictly speaking, a univariate gaussian distribution for every variable does not in itself ensure that the total dataset is multivariate gaussian, but the existence of a multivariate gaussian dataset does mean that each variable will have a univariate gaussian distribution (Cooley and Lohnes, 1971).

The univariate histograms for each variable used in the analysis are shown in Figures 5.9-5.40 for both the event dataset and the grid cell dataset. It is clear from visual inspection of the histograms that none are exactly gaussian, although a number are roughly gaussian in appearance. Some of the observables only have positive values (e.g., topographic elevation or distance to the nearest fault), and most of these are peaked near zero on the abscissa. In these cases, we also generated histograms of the base 10 logarithms of these variables to see if the transformed values better resembled gaussian shapes. From these histograms, we decided to use in our multivariate analyses the original values for stress azimuth, topographic elevation, gravity residual and magnetic residual, while we selected the logarithmic values for distance to nearest

fault, length of nearest fault, and distance to nearest river for our analyses. Since for each variable the histograms for the event dataset and the grid cell dataset were similar, we used the same combination of original and logarithmically transformed variables for both datasets.

The univariate histograms also revealed some problems in the data. We had planned to include the gradients of the topography, gravity residual field and magnetic residual field as additional variables in the multivariate statistical analyses, but the histograms (Figures 5.12, 5.14, 5.16, 5.28, 5.30, and 5.32) showed that these are heavily peaked near the smallest gradient values with a relatively few larger gradient values. This is due to the relatively smoothed regional fields of our original data, so the computed gradient values are necessarily small and show relatively little variation across the region. We felt that the actual local gradient values (resolution of 1-2 km) of these fields are not well-represented by the gradient values we computed, so we decided not to include the gradients in the statistical analyses.

We also found that the fault length histogram for the event dataset (Figure 5.19) as well as the log10 (fault length) histogram for that dataset (Figure 5.20) have two peaks, violating the assumption of a simple gaussian distribution for this variable. In fact, the log10 (fault length) histogram appears fairly gaussian with a peak around a value of 4.0, but then there is a second, isolated peak at 5.6 (Figure 5.20). The latter peak is likely due to very long faults in the more seismically active parts of the CEUS. Just by chance long faults have a higher likelihood of having earthquakes near them if the earthquakes are randomly scattered throughout the region, and this second peak probably reflects the undue weighting of a relatively few faults in the data. However, we chose to keep this parameter in the statistical analyses (using the log10 value) because we felt that this was an important geologic feature that could not be eliminated. Curiously, these peaks do not show up in the histograms for the grid cell dataset (Figures 5.35 and 5.36).

Stress azimuth for the event data (Figure 5.10) was the other variable that showed two peaks in its histogram. The highest peak is near 70° azimuth, reflecting the average regional stress field in the CEUS (Zoback and Zoback, 1980). A second, smaller peak shows up at an azimuth of 0°, which may be a default value corresponding to uncertain data. Again, we felt that this was an important variable to include in the analysis, so we ignored the bimodal distribution in this histogram.

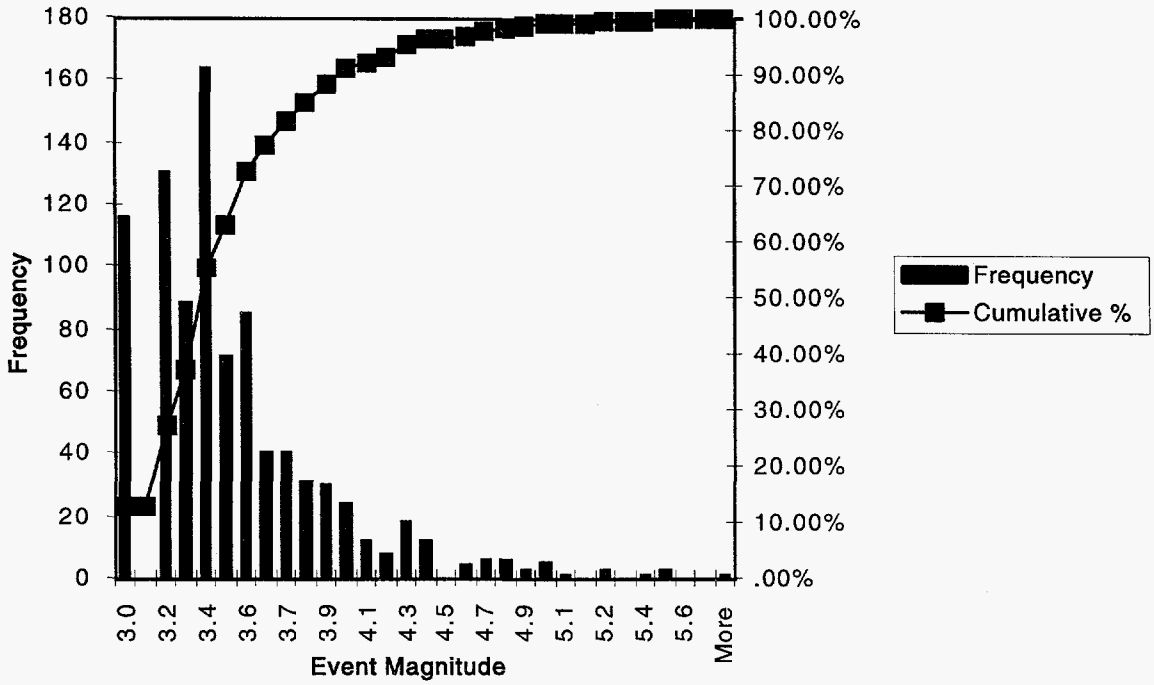

Figure 5.9. Interval histogram and cumulative plot of the event magnitudes for the observations from the full epicenter dataset.

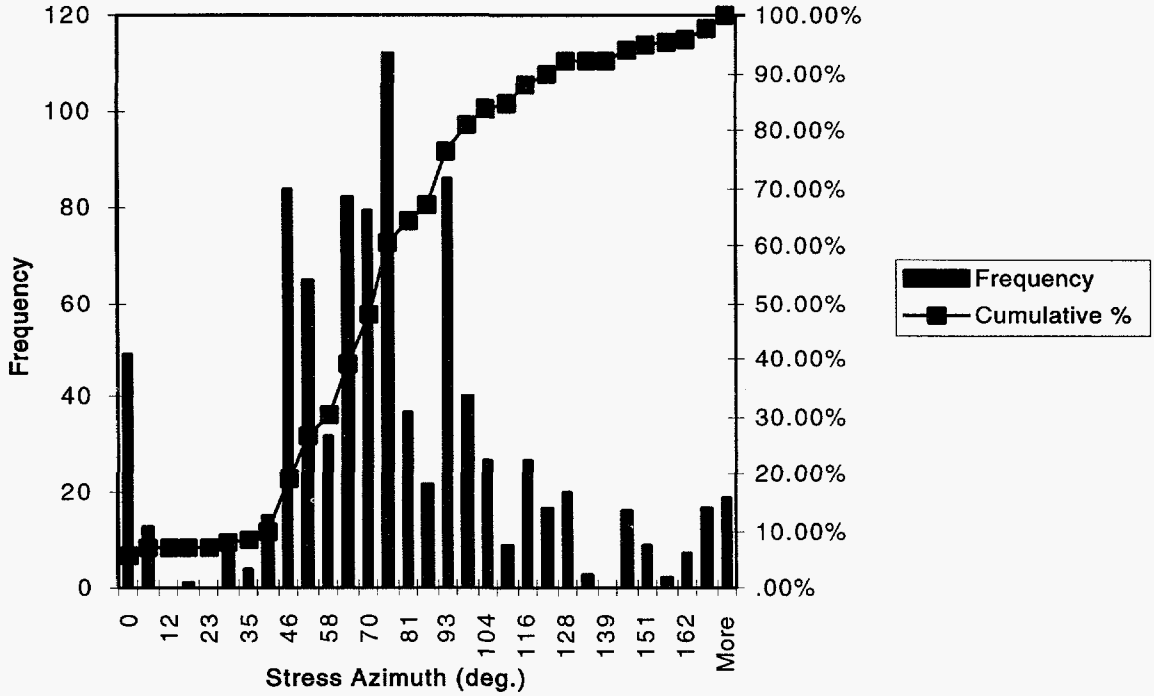

Figure 5.10. Interval histogram and cumulative plot of the azimuths of the greatest principal stresses for the observations from the full epicenter dataset.

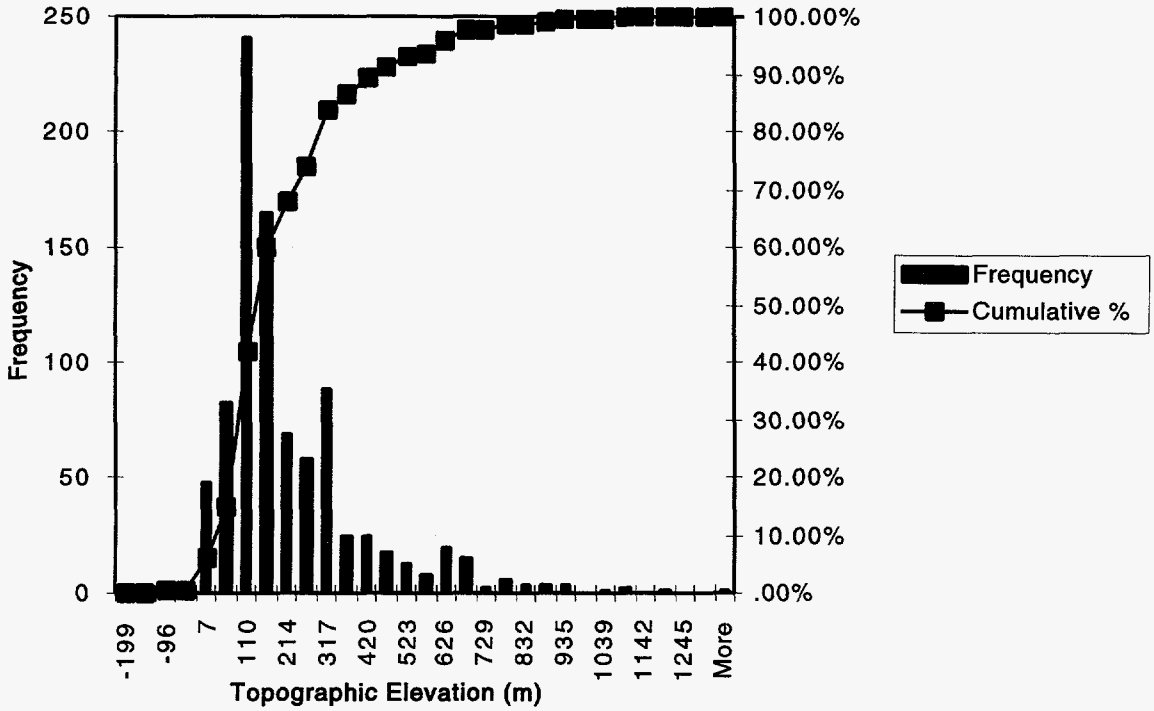

Figure 5.11. Interval histogram and cumulative plot of the topographic elevations for the observations from the full epicenter dataset.

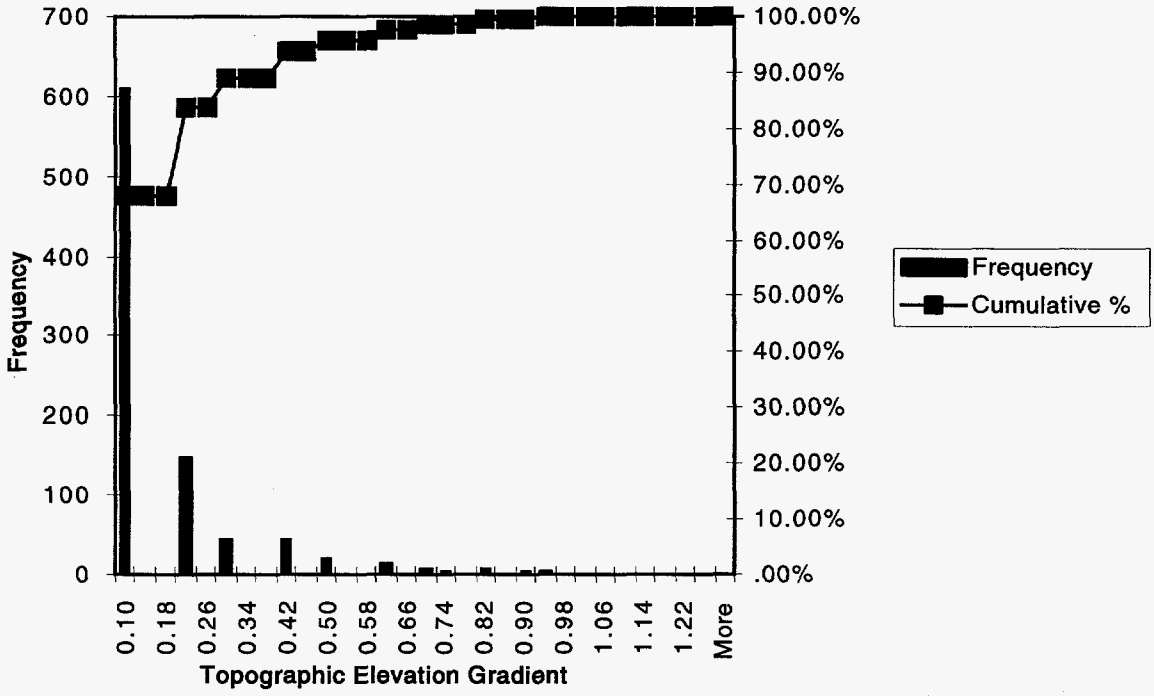

Figure 5.12. Interval histogram and cumulative plot of the nearest topographic elevation gradient contour for the observations from the full epicenter dataset.

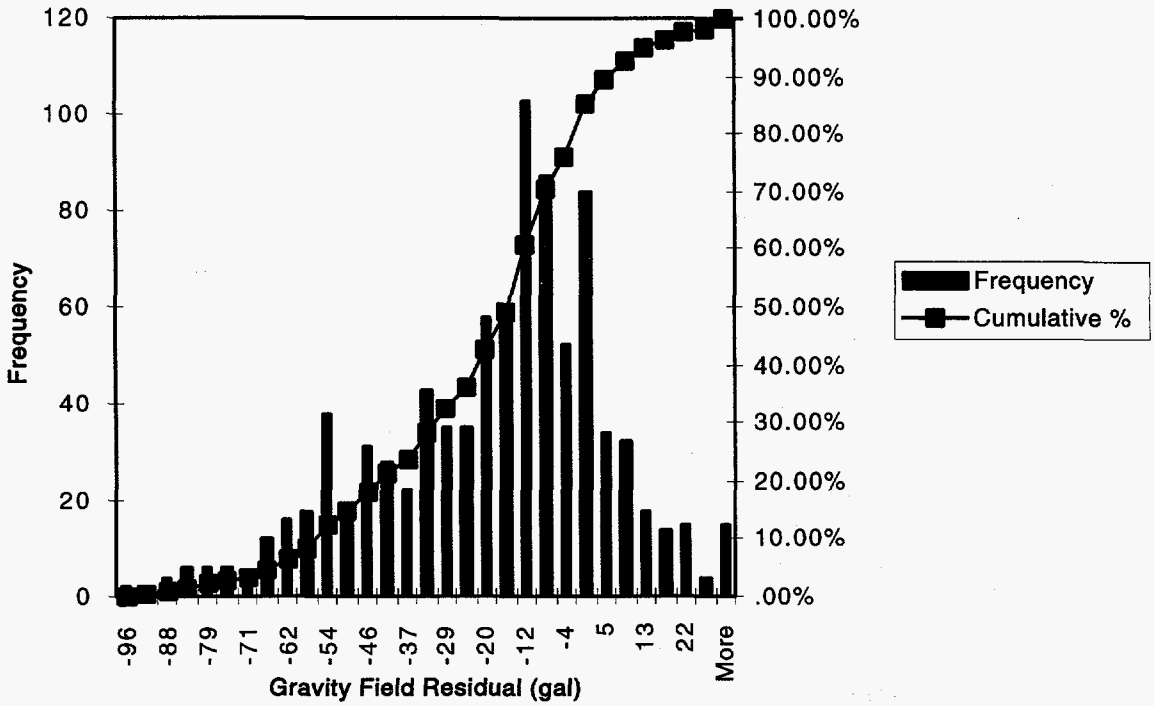

Figure 5.13. Interval histogram and cumulative plot of the gravity field residual for the observations from the full epicenter dataset.

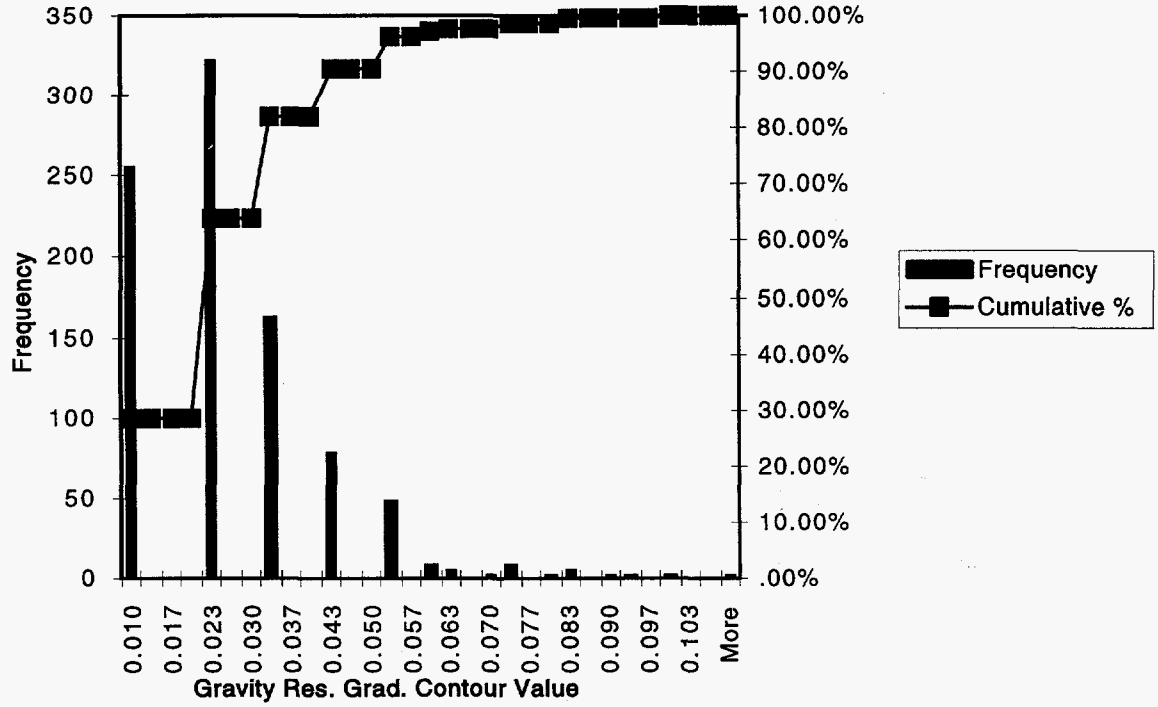

Figure 5.14. Interval histogram and cumulative plot of the nearest gravity field residual gradient contour for the observations from the full epicenter dataset.

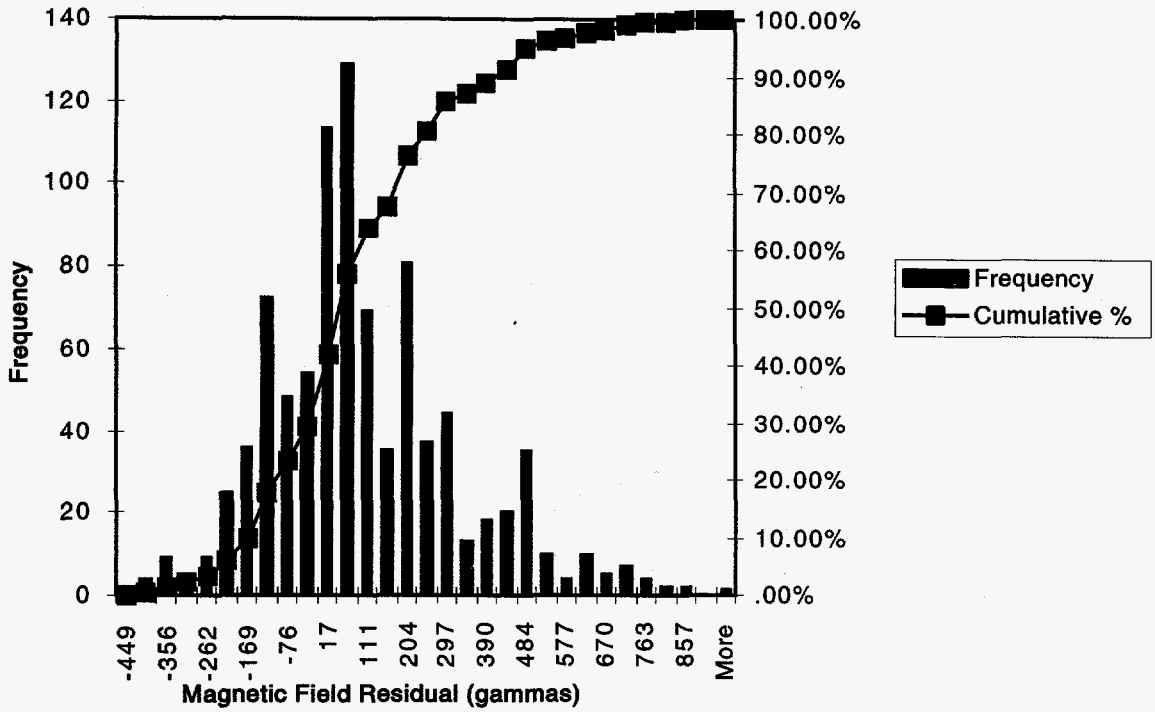

Figure 5.15. Interval histogram and cumulative plot of the magnetic field residual for the observations from the full epicenter dataset.

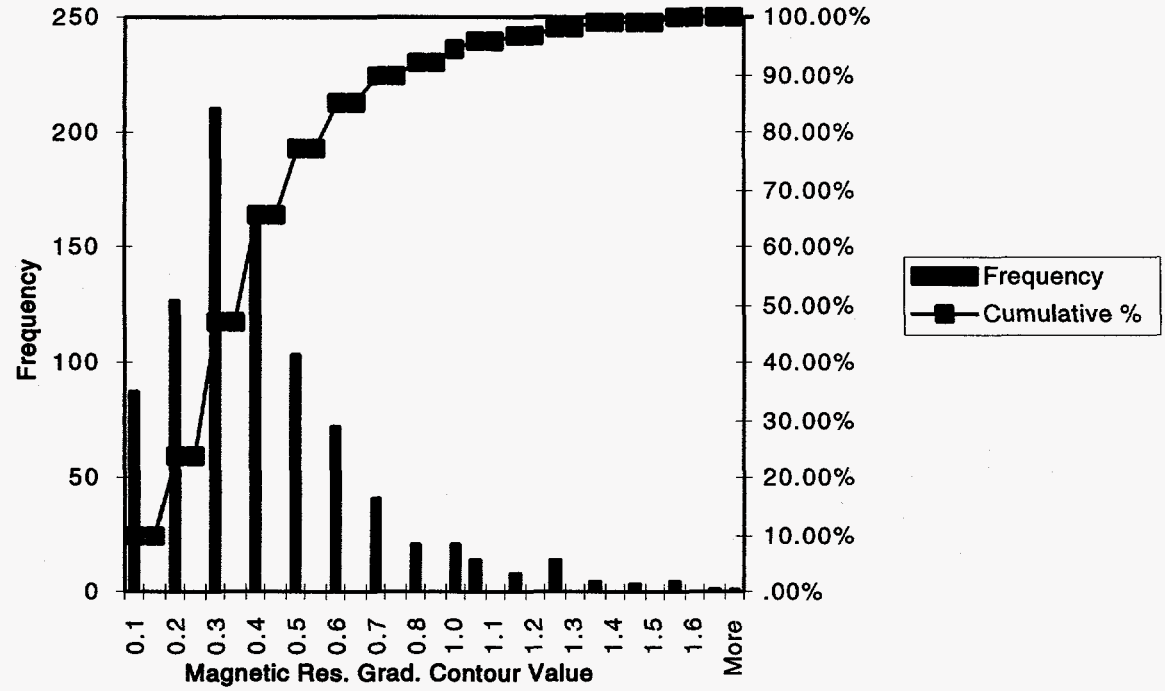

Figure 5.16. Interval histogram and cumulative plot of the nearest magnetic field residual gradient contour for the observations from the full epicenter dataset.

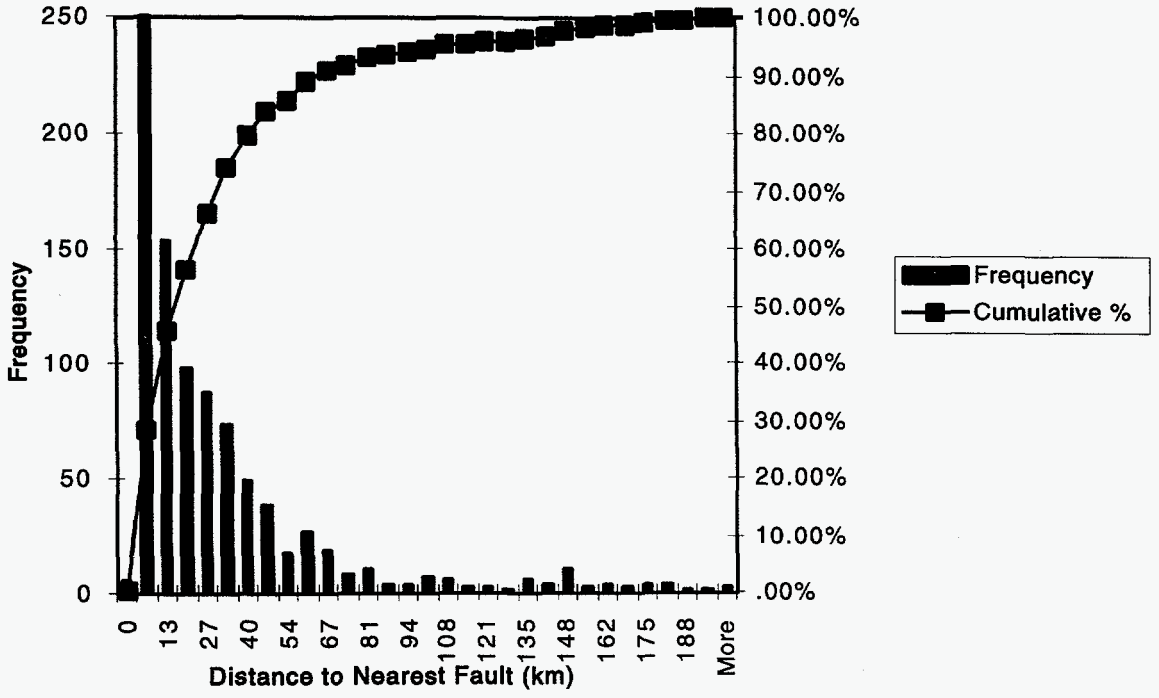

Figure 5.17. Interval histogram and cumulative plot of the distance to the nearest fault for the observations from the full epicenter dataset.

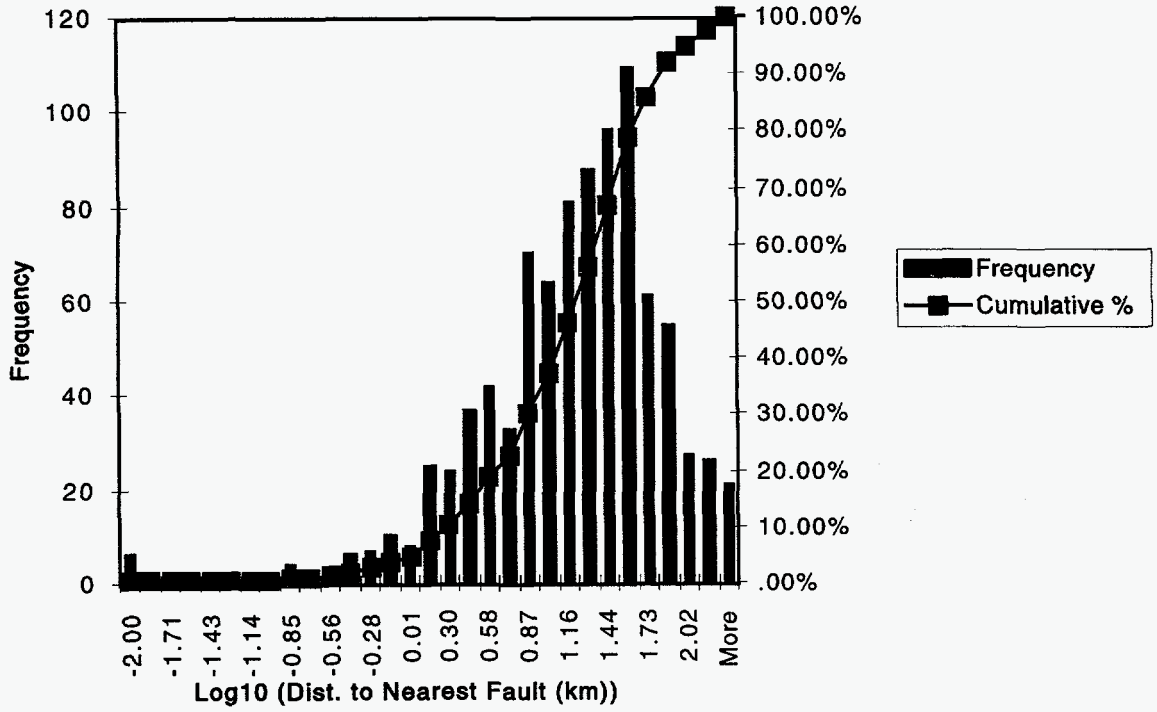

Figure 5.18. Interval histogram and cumulative plot of the log10 (distance to the nearest fault) for the observations from the full epicenter dataset.

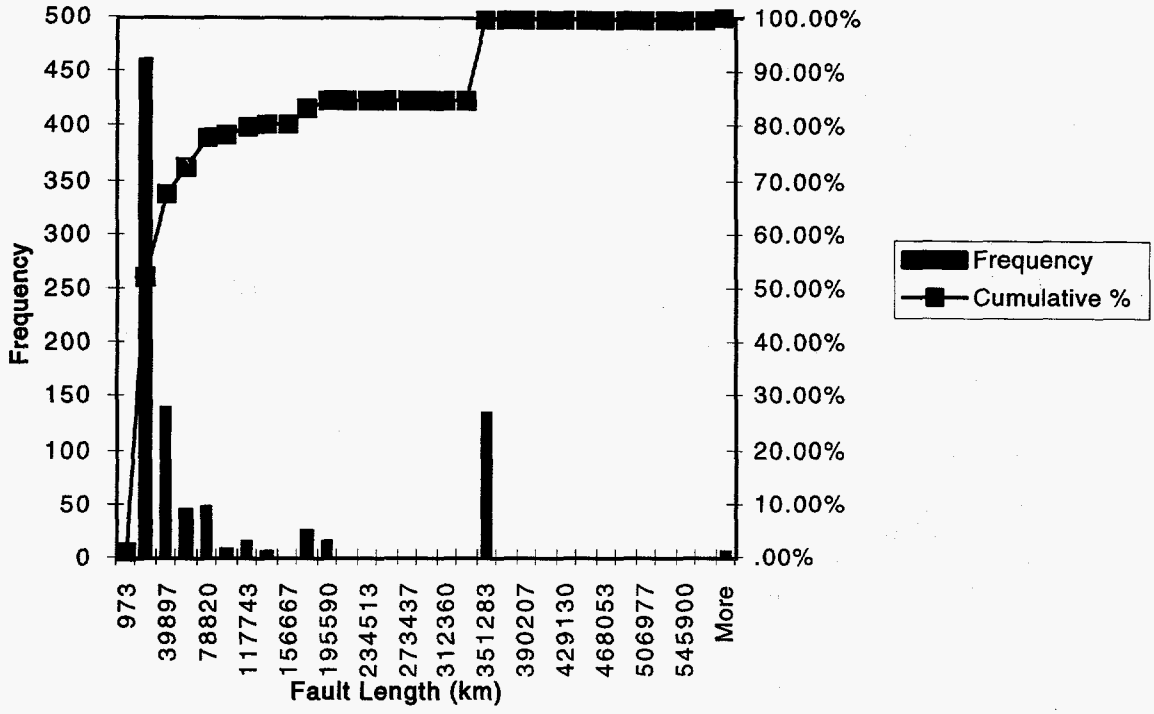

Figure 5.19. Interval histogram and cumulative plot of the length of the nearest fault for the observations from the full epicenter dataset.

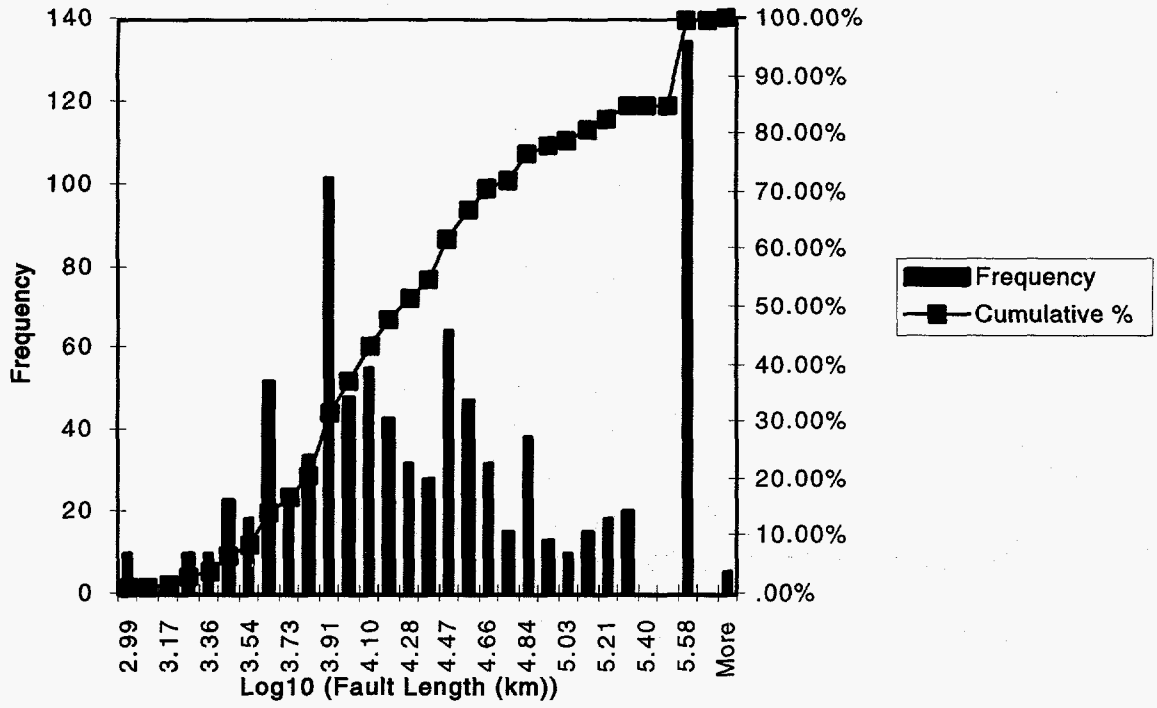

Figure 5.20. Interval histogram and cumulative plot of the log10 (length of the nearest fault) for the observations from the full epicenter dataset.

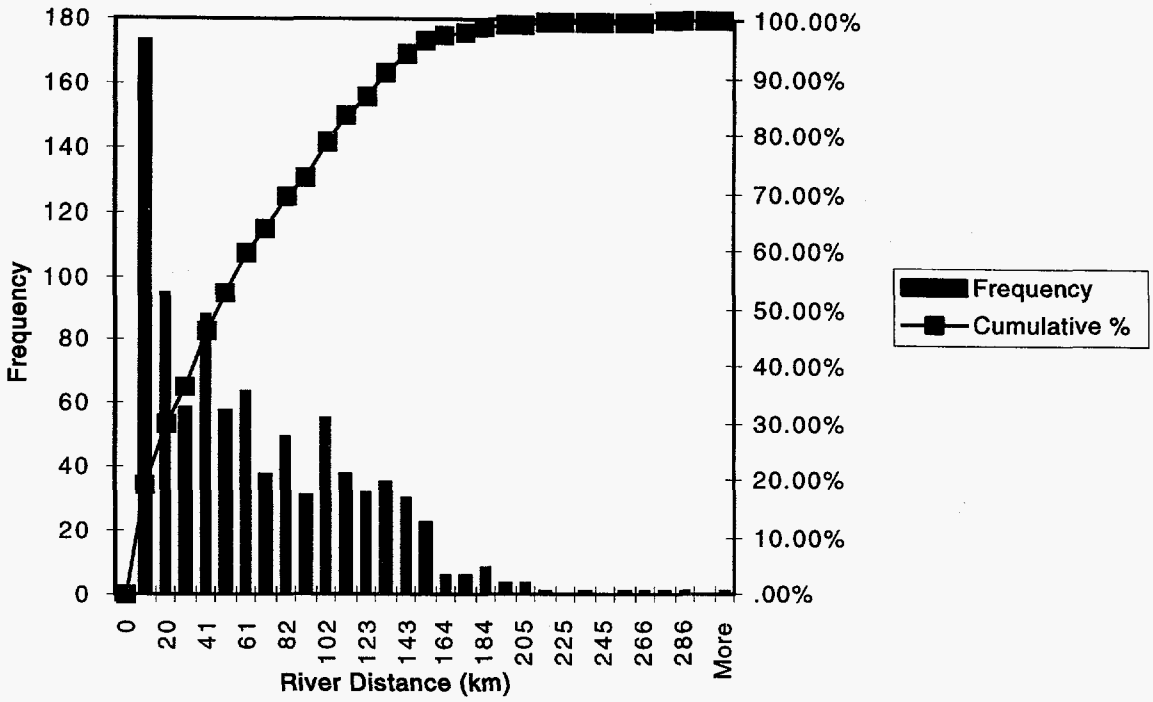

Figure 5.21. Interval histogram and cumulative plot of the distance to the nearest river for the observations from the full epicenter dataset.

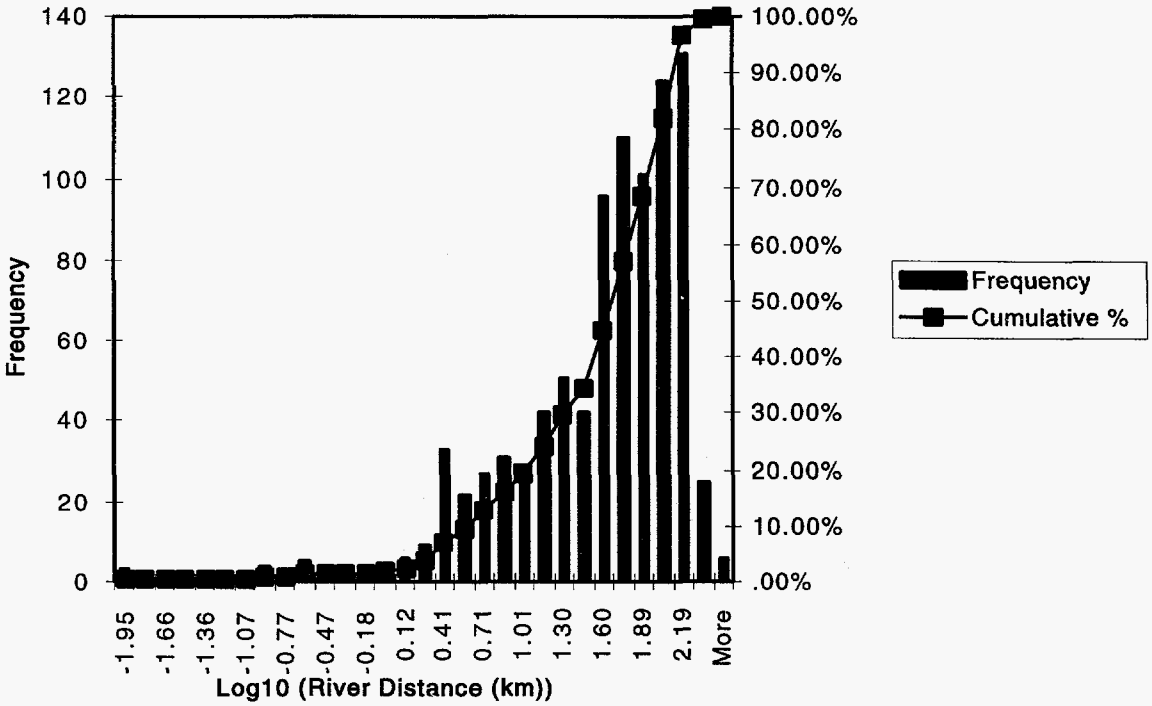

Figure 5.22. Interval histogram and cumulative plot of the log10 (distance to the nearest river) for the observations from the full epicenter dataset.

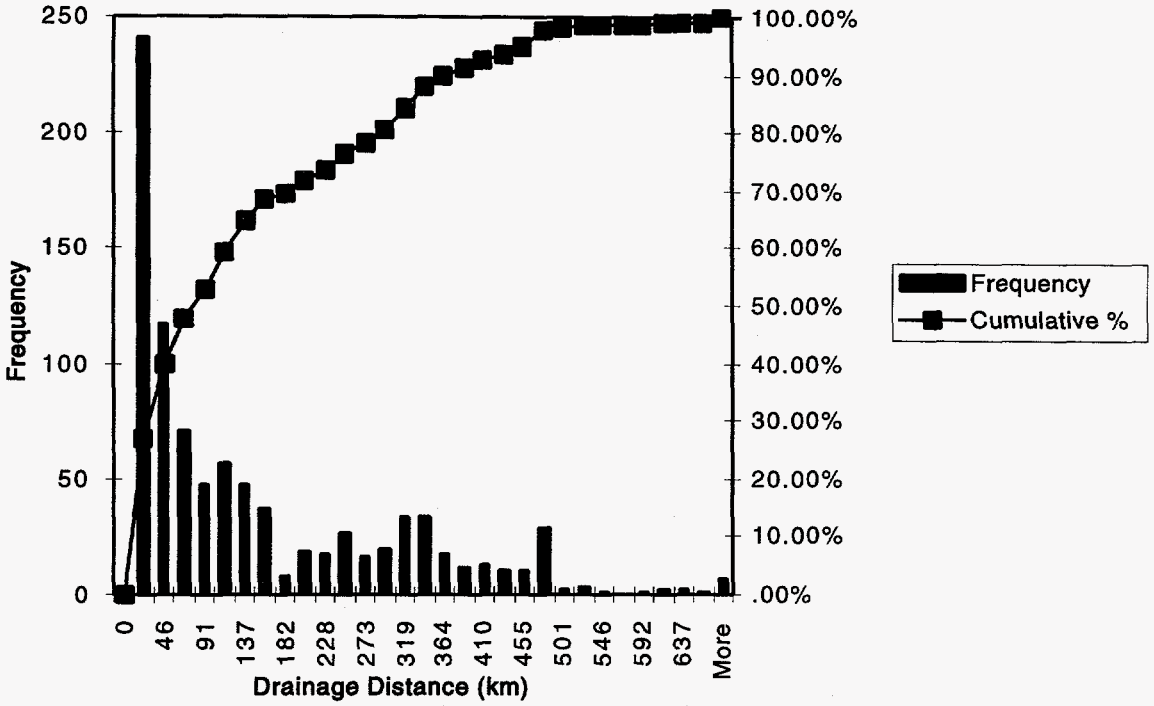

Figure 5.23. Interval histogram and cumulative plot of the distance to the nearest drainage for the observations from the full epicenter dataset.

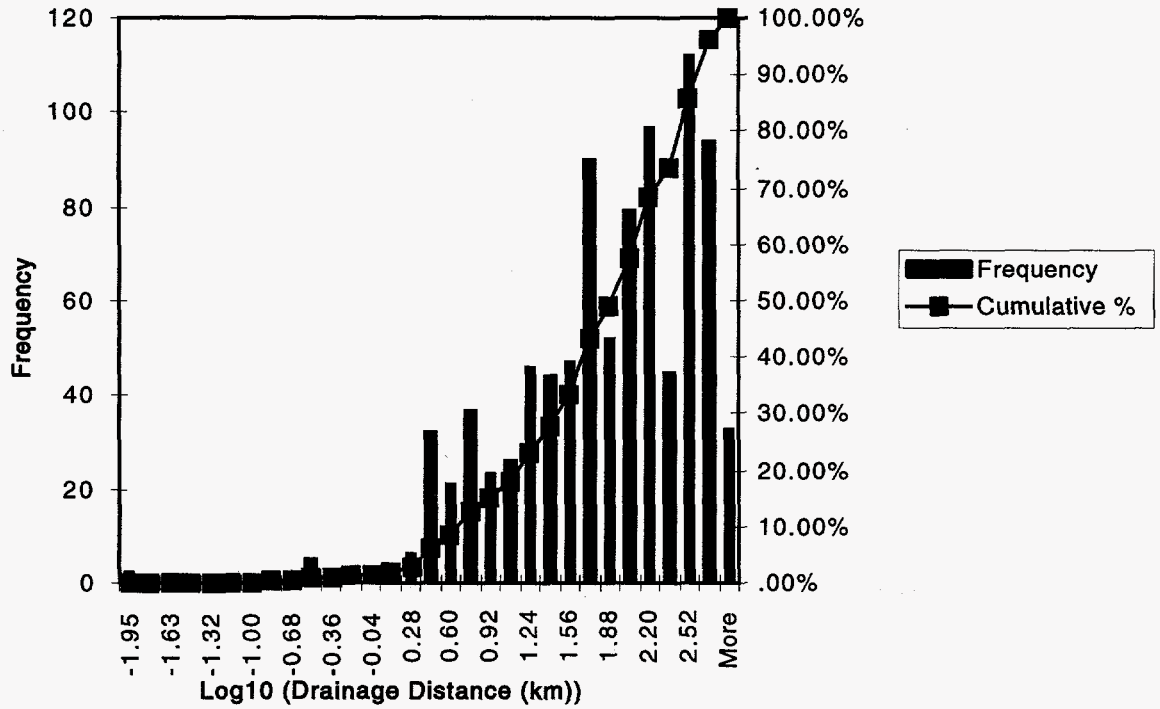

Figure 5.24. Interval histogram and cumulative plot of the log10 (distance to the nearest drainage) for the observations from the full epicenter dataset.

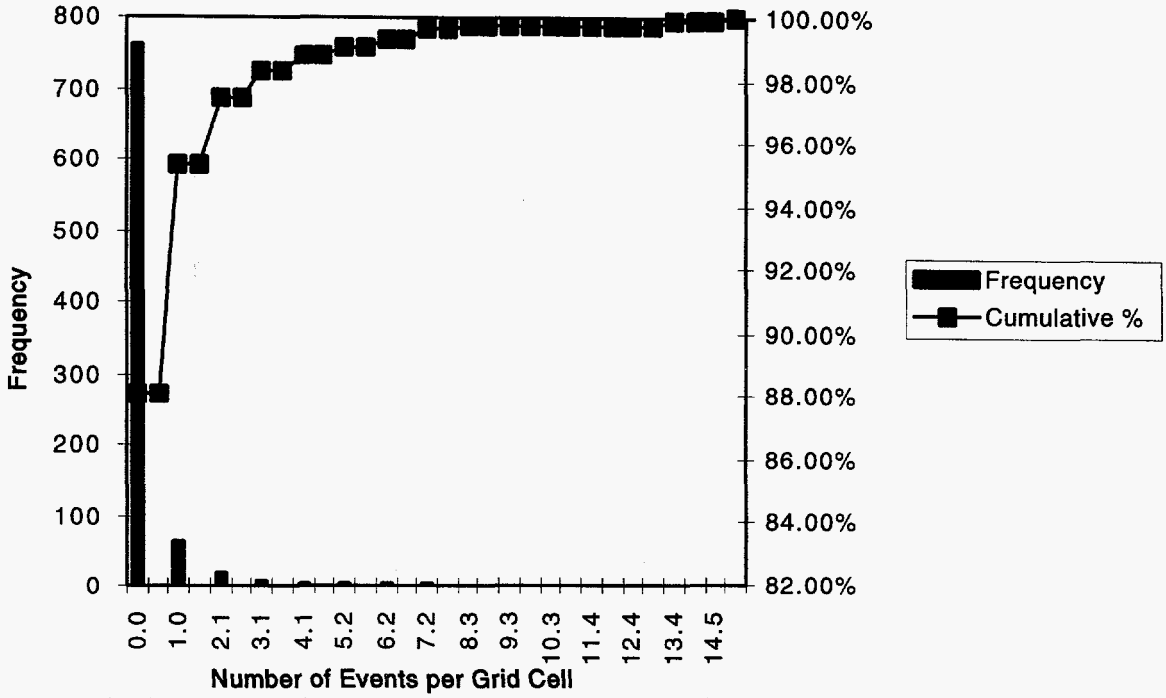

Figure 5.25. Interval histogram and cumulative plot of the number of events per grid cell for the observations from the grid cell dataset.

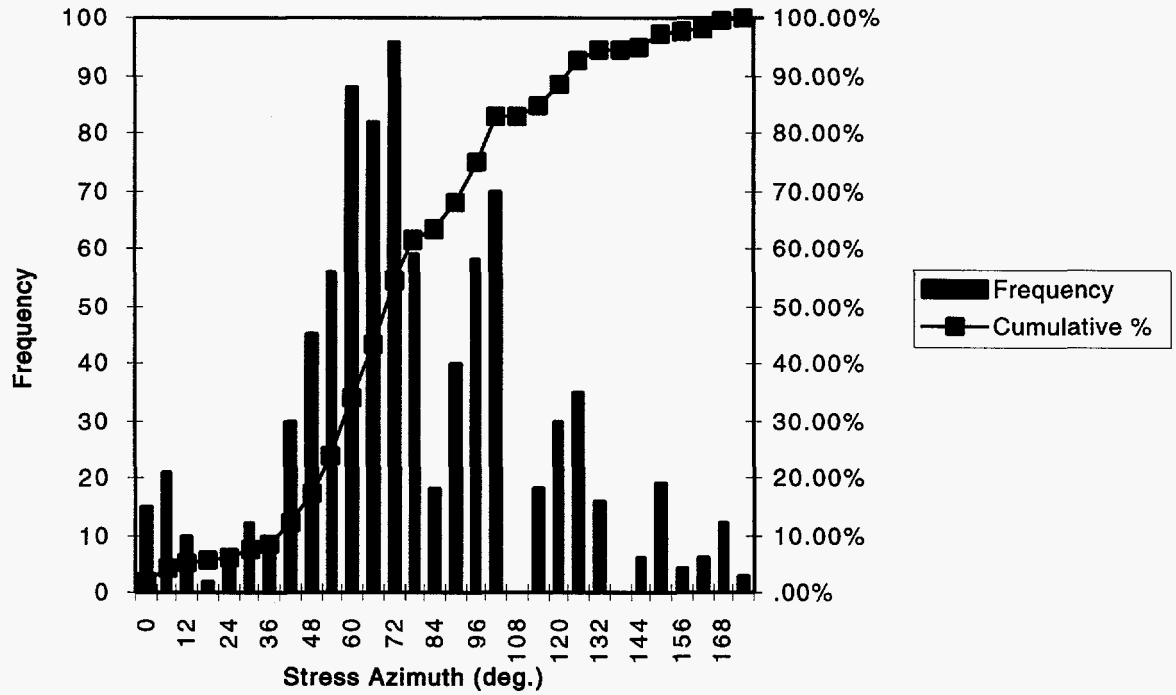

Figure 5.26. Interval histogram and cumulative plot of the azimuths of the greatest principal stresses for the observations from the grid cell dataset.

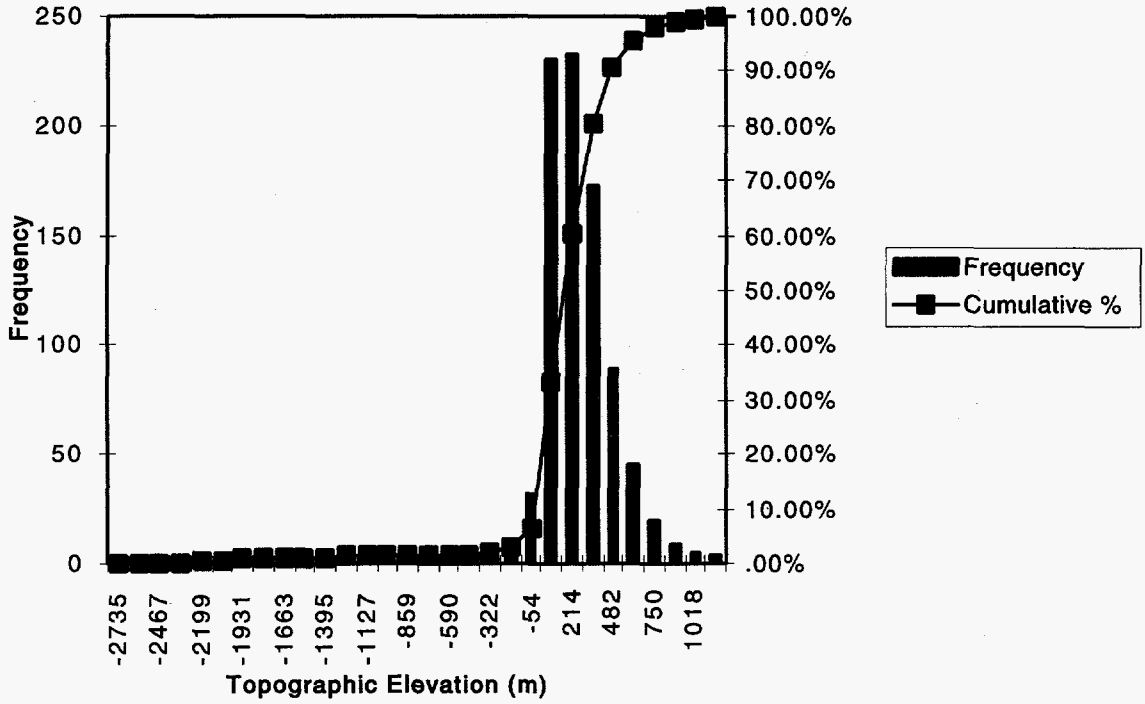

Figure 5.27. Interval histogram and cumulative plot of the topographic elevations for the observations from the grid cell dataset.

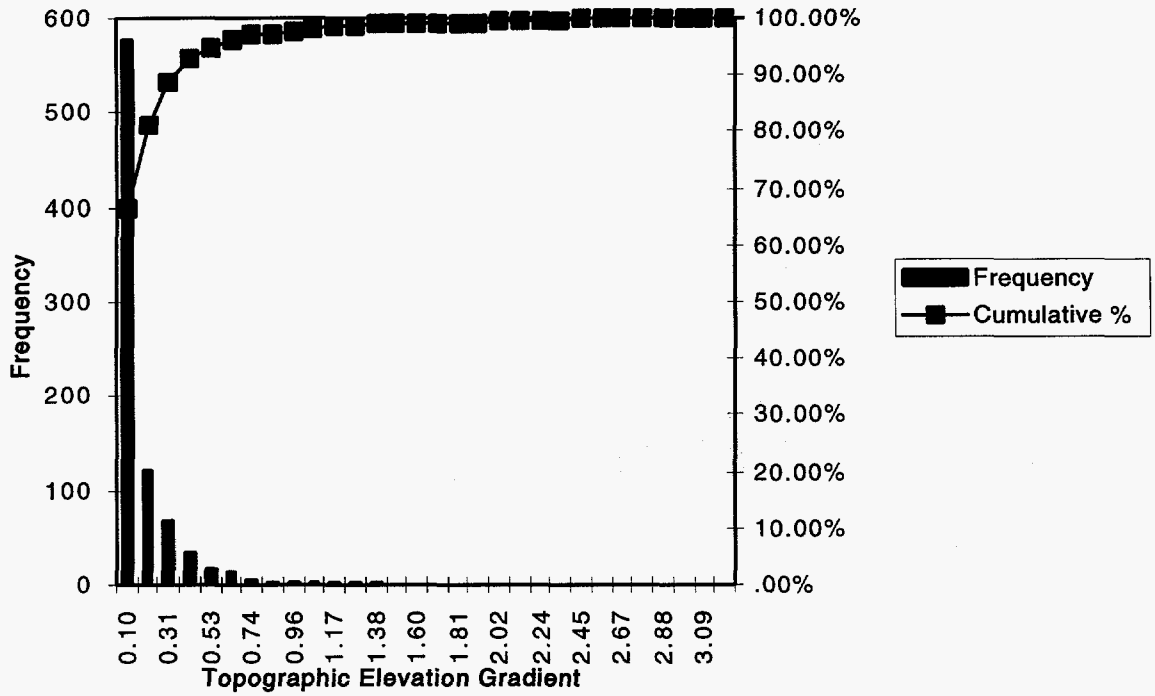

Figure 5.28. Interval histogram and cumulative plot of the nearest topographic elevation gradient contour for the observations from the grid cell dataset.

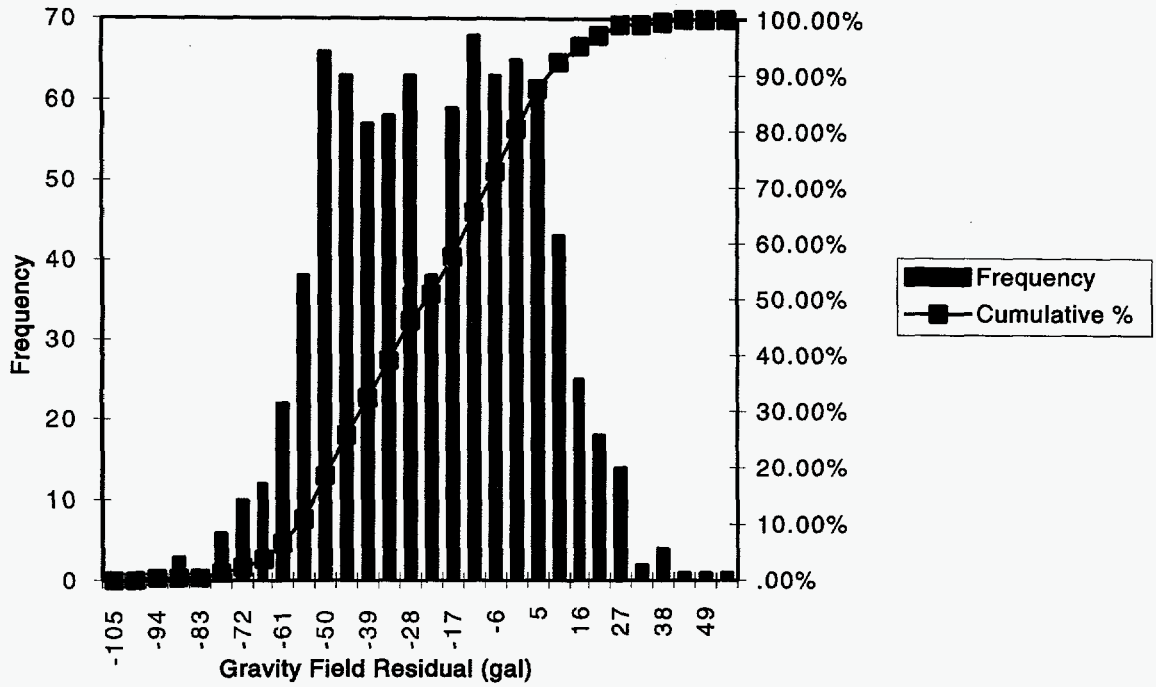

Figure 5.29. Interval histogram and cumulative plot of the gravity field residual for the observations from the grid cell dataset.

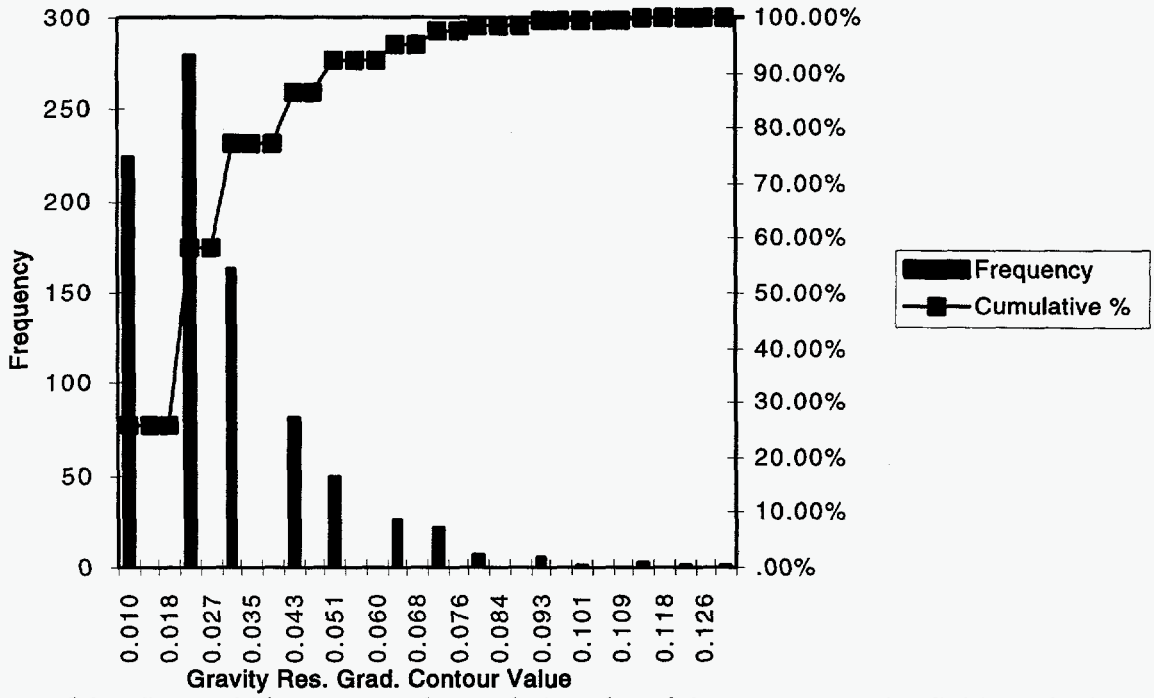

Figure 5.30. Interval histogram and cumulative plot of the nearest gravity field residual gradient contour for the observations from the grid cell dataset.

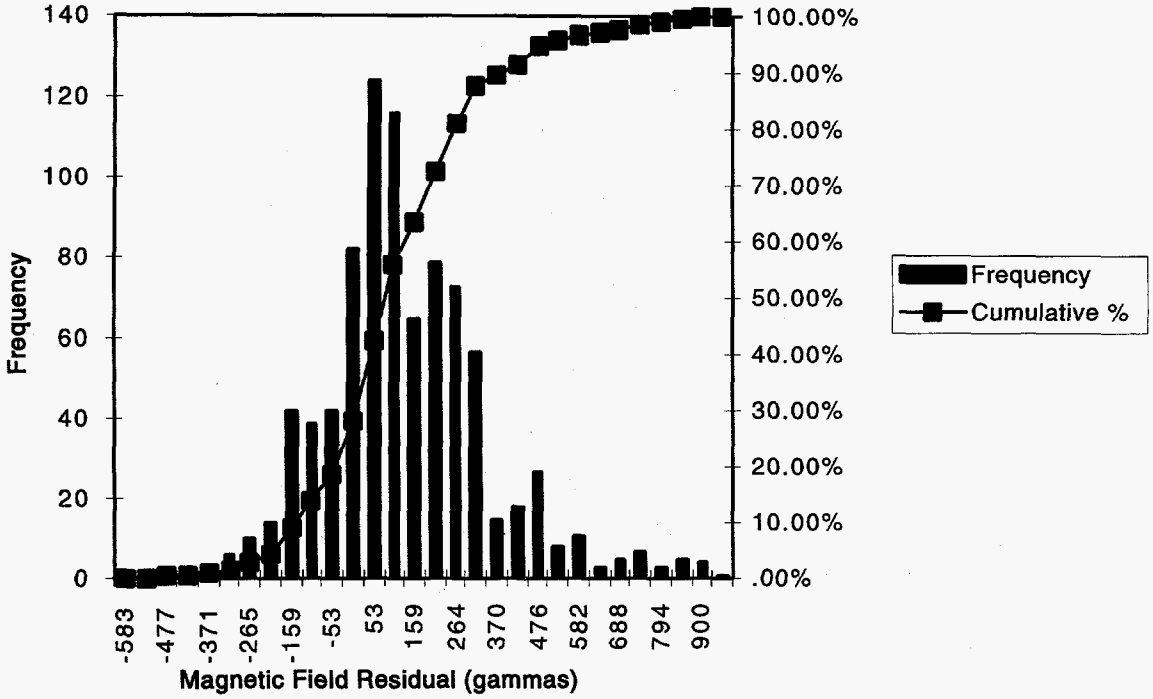

Figure 5.31. Interval histogram and cumulative plot of the magnetic field residual for the observations from the grid cell dataset.

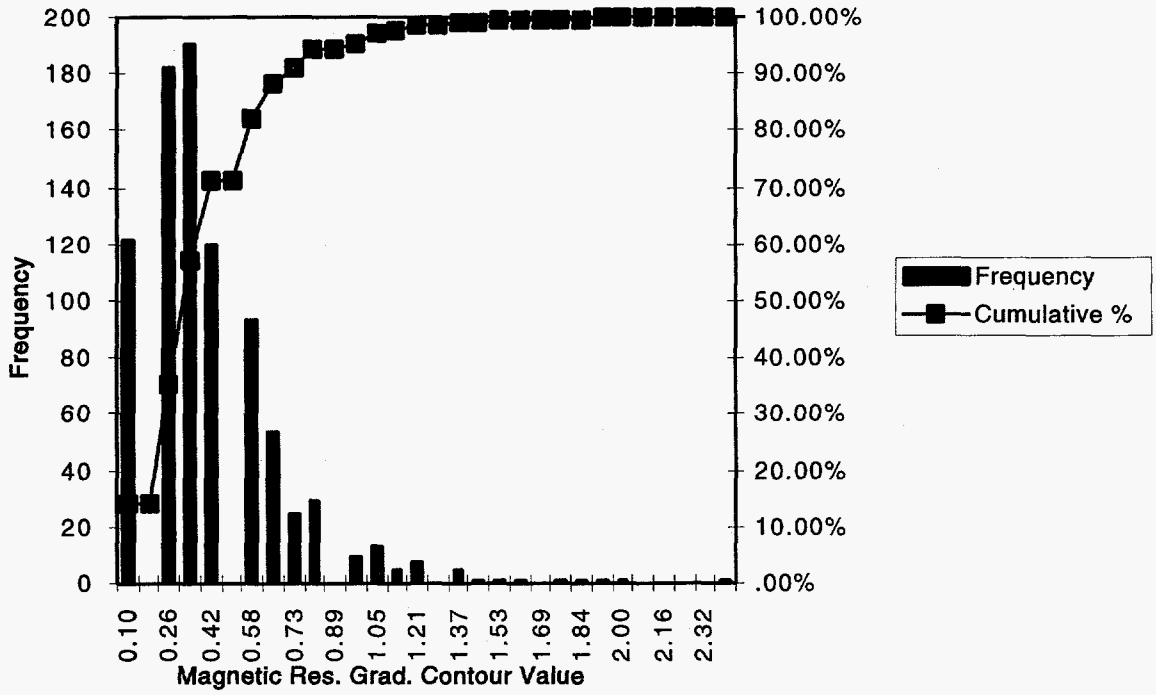

Figure 5.32. Interval histogram and cumulative plot of the nearest magnetic field residual gradient contour for the observations from the grid cell dataset.

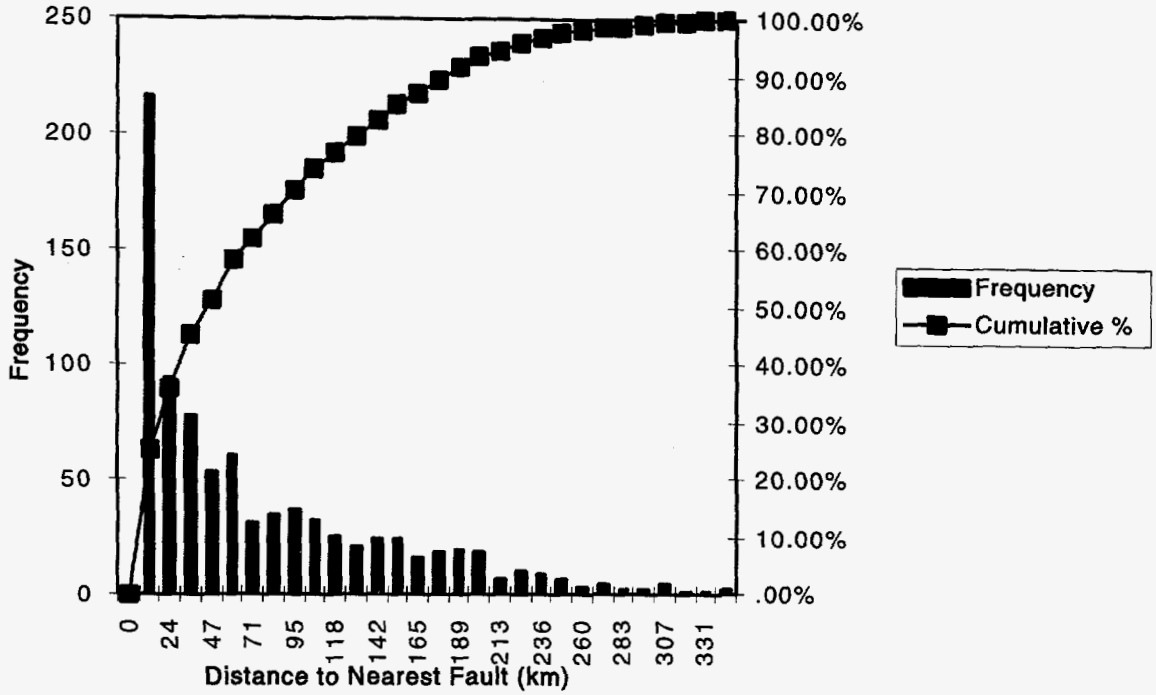

Figure 5.33. Interval histogram and cumulative plot of the distance to the nearest fault for the observations from the grid cell dataset.

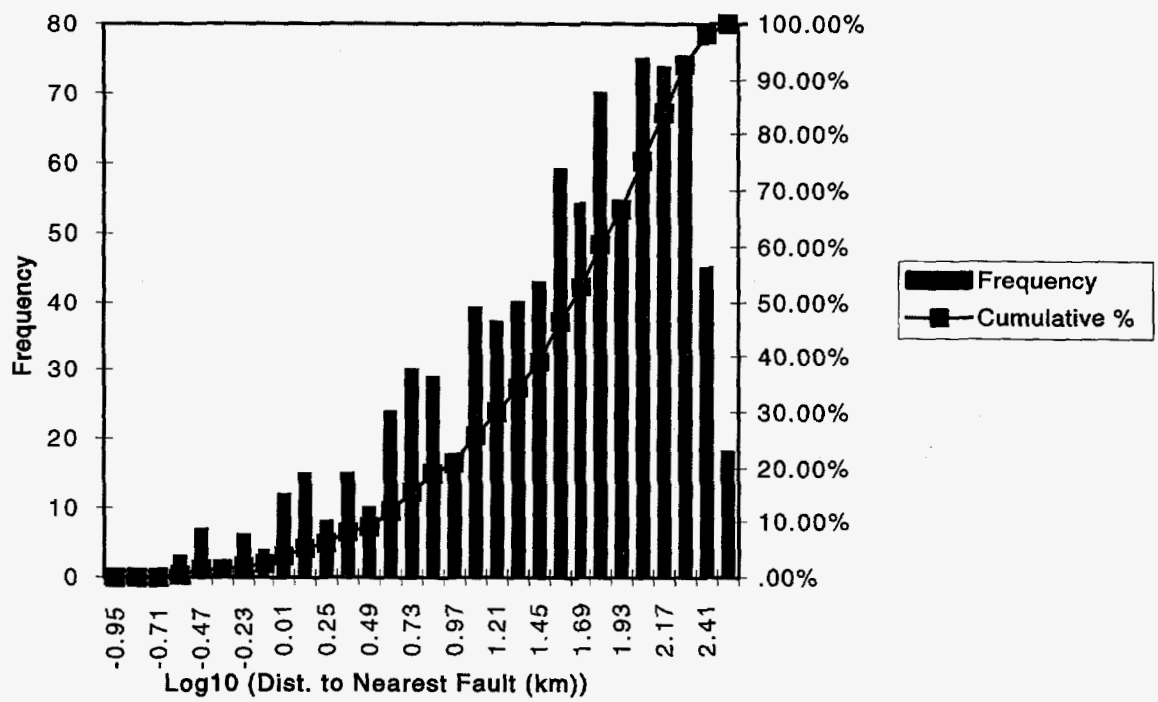

Figure 5.34. Interval histogram and cumulative plot of the log10 (distance to the nearest fault) for the observations from the grid cell dataset.

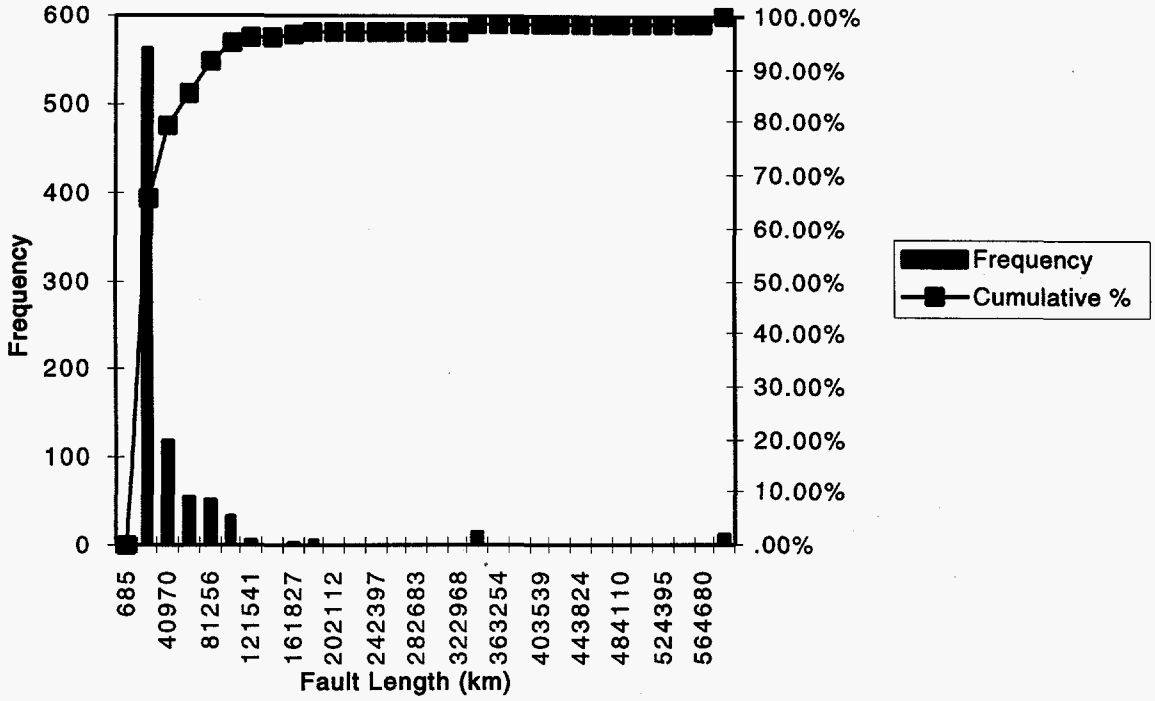

Figure 5.35. Interval histogram and cumulative plot of the length of the nearest fault for the observations from the grid cell dataset.

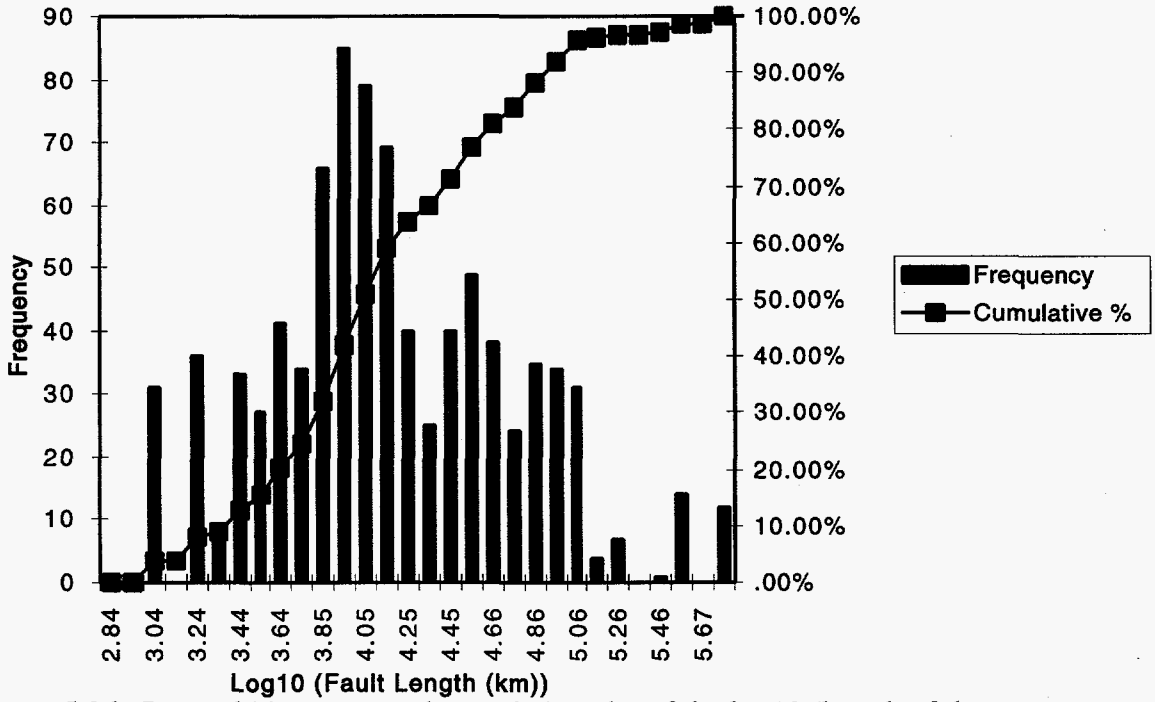

Figure 5.36. Interval histogram and cumulative plot of the log10 (length of the nearest fault) for the observations from the grid cell dataset.

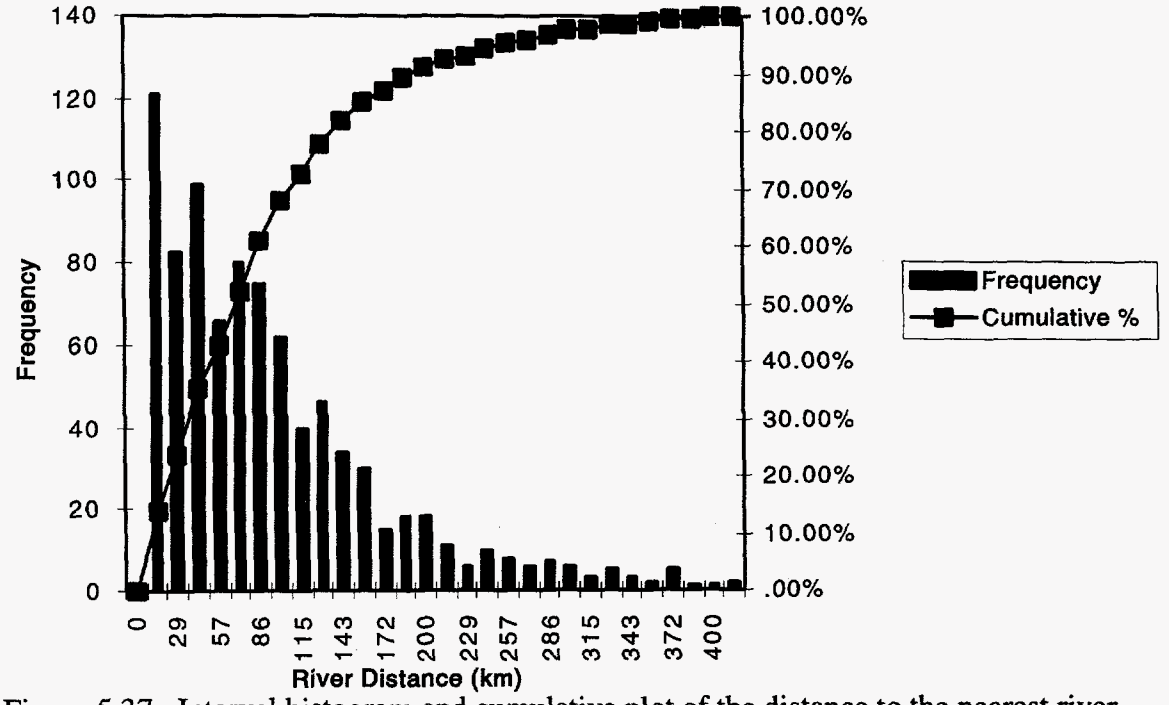

Figure 5.37. Interval histogram and cumulative plot of the distance to the nearest river for the observations from the grid cell dataset.

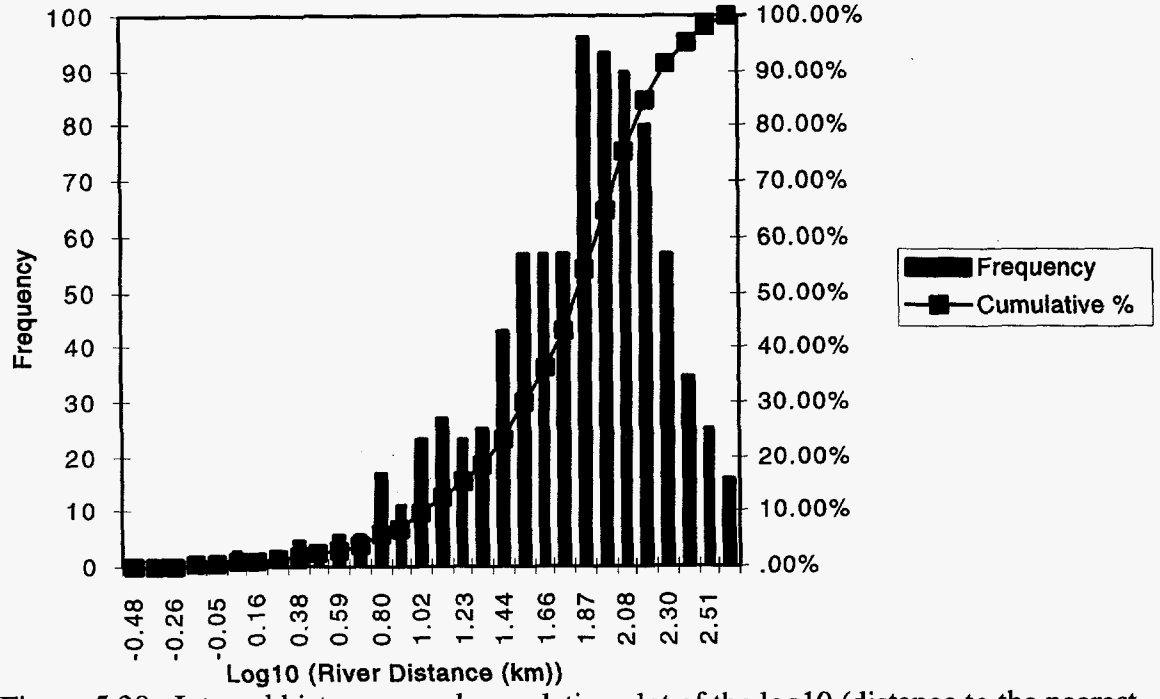

Figure 5.38. Interval histogram and cumulative plot of the log10 (distance to the nearest river) for the observations from the grid cell dataset.

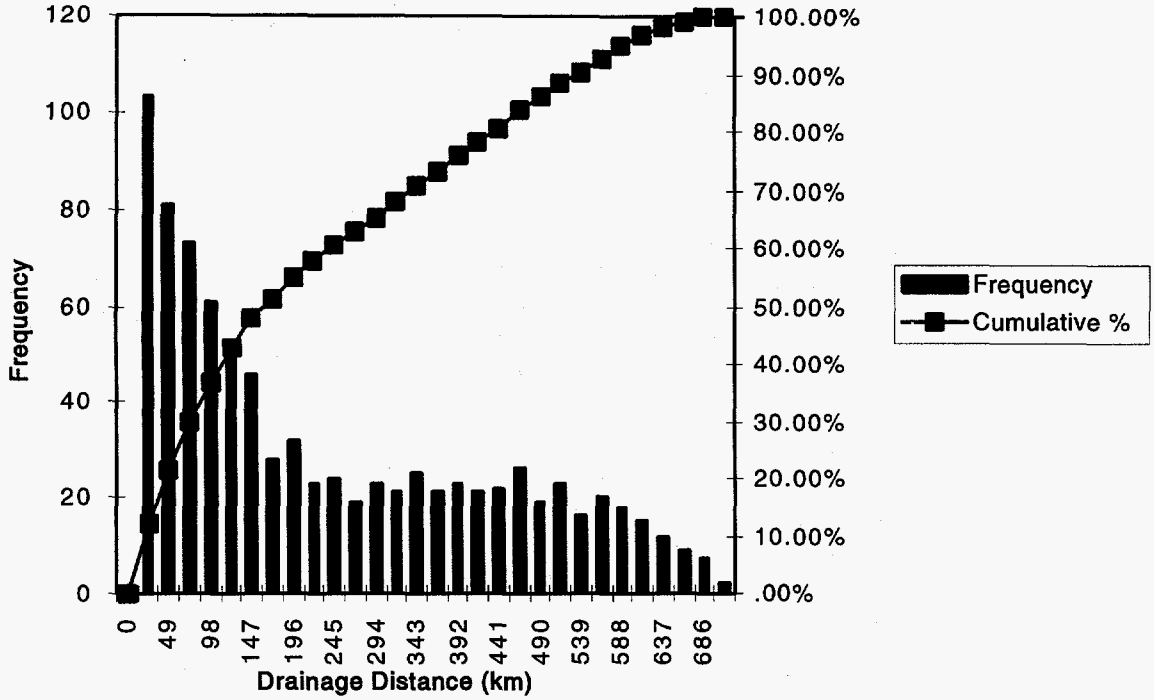

Figure 5.39. Interval histogram and cumulative plot of the distance to the nearest drainage for the observations from the grid cell dataset.

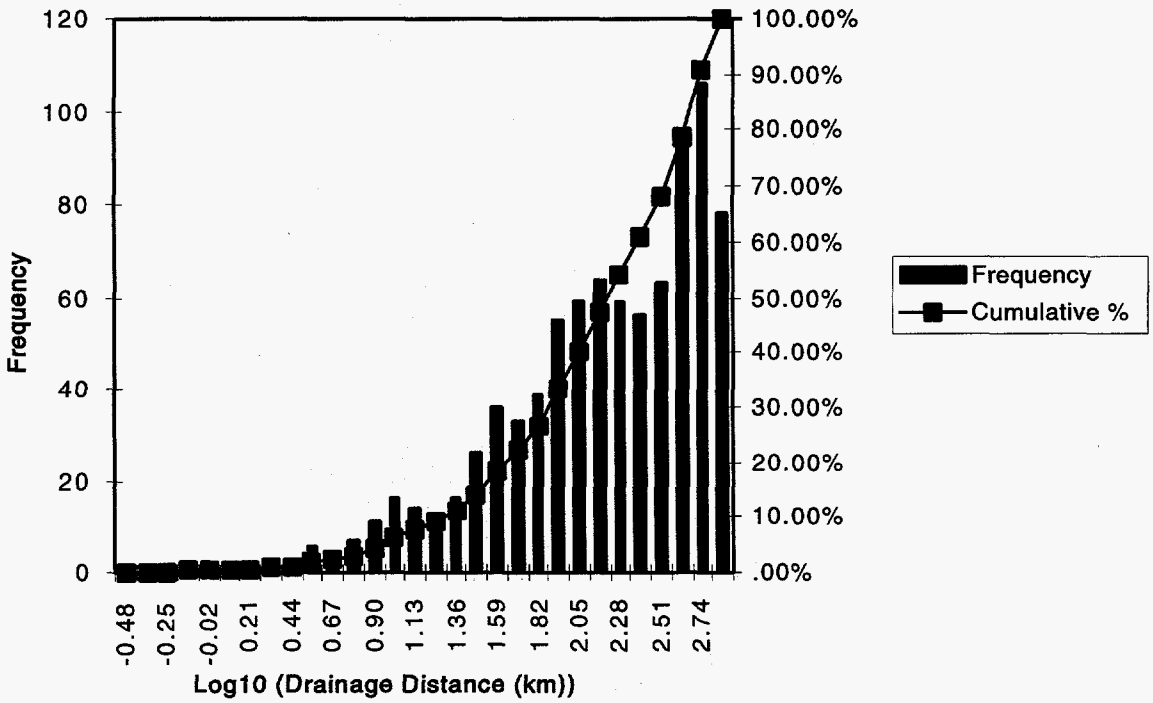

Figure 5.40. Interval histogram and cumulative plot of the log10 (distance to the nearest drainage) for the observations from the grid cell dataset.

Finally, our examination of the raw data for distance to nearest river and distance to nearest drainage showed that in most cases, these two variables have exactly the same value. This is not surprising, since the nearest drainage in many areas is the nearest major river. Thus, we decided that these two variables were redundant, and we opted to use only the log10 (distance to nearest river) in the multivariate statistical analyses.

5.6 Analysis of M≥4.5 Events

Our first set of statistical analyses was carried out on a subset of our dataset, those 33 events from our catalog with magnitudes of 4.5 or greater. These events are listed in Table 5.3. We carried out two sets of tests. The first was a factor analysis to look for commonalities among the event observables. The second was a cluster analysis to see which of the larger events have common sets of observables.

5.6.1 Factor Analysis

The factor analysis was carried out on both the unnormalized data as well as on the normalized data with mean and standard deviation removed from each variable. For each dataset, two cases were run. In one case the variables topography, gravity residual, magnetic residual, log10 (length of nearest fault), log10 (distance to nearest fault), log10 (distance to nearest river) and azimuth of the nearest stress measurement were used. In the second case, event magnitude was included as another observable. Only those eigenvectors (or factors) with eigenvalues of 0.7 or greater are reported, as those account for better than 80% of the variance of the data in all cases. The eigenvectors from the varimax analyses along with their corresponding eigenvalues are listed in Table 5.4 for both the normalized and the unnormalized data. Differences between the results using the normalized and unnormalized data are very small, typically showing up only in the third or fourth significant figure.

The runs without magnitude were intended to seek relationships among the variables, while the runs with magnitude included were intended to investigate whether magnitude is related to any of the other variables. For both the normalized and the unnormalized data, two sets of variables were found to be related. Magnetic residual is positively correlated with log10 (distance to nearest fault) and negatively correlated with log10 (length of nearest fault) according to one eigenvector, while topographic elevation is negatively correlated with gravity residual according to another eigenvector. In all cases, these eigenvectors have the two largest eigenvalues, indicating they are the most pronounced relationships in the data. Since many faults are marked by strong anomalies in the local magnetic field, the former relationship implies that the regional magnetic field residual contains information about nearby major faults. Magnetic field anomalies have long been used to look for faults in geophysical exploration (Telford et al., 1976), so this statistical result is consistent with longstanding practice. The inverse relationship between topography and gravity residual is well known (e.g., Press and Siever, 1978), and in fact

it arises out of the gravity field corrections used to compute the Bouger gravity field for isostatically compensated crust.

Table 5.3
Study Events with M≥4.5

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
32	-88.20	38	4/27/1925	0	4.9	VII	ILLINOIS_U.S.A.
45	-87.20	37.9	9/2/1925	0	4.5	VI	KENTUCKY_U.S.A.
90	-90.20	36	5/7/1927	0	4.7	VI	ARKANSAS_U.S.A.
127	-82.83	36.11	11/3/1928	0	4.6	VI	TENNESSEE_U.S.A.
146	-78.40	42.91	8/12/1929	9	5.2	III	NEW_YORK_U.S.A.
190	-73.78	43.47	4/20/1931	5	4.8	VII	NEW_YORK_U.S.A.
199	-84.27	40.43	9/20/1931	5	4.7	VII	OHIO_U.S.A.
204	-89.80	34.1	12/17/1931	0	4.7	VII	MISSISSIPPI_U.S.A.
294	-84.27	40.49	3/2/1937	2	4.9	VII	OHIO_U.S.A.
296	-84.28	40.47	3/9/1937	3	4.8	III	OHIO_U.S.A.
365	-90.14	38.18	11/23/1939	0	4.9	V	ILLINOIS_U.S.A.
385	-90.10	38.2	11/23/1940	0	5	VI	ILLINOIS_U.S.A.
387	-71.37	43.87	12/20/1940	10	5.5	VII	NEW HAMPSHIRE_U.S.A.
388	-71.28	43.91	12/24/1940	8	5.5	VII	NEW HAMPSHIRE_U.S.A.
461	-74.72	44.96	9/5/1944	12	5.8	III	NEW YORK_U.S.A.
463	-74.65	45	9/5/1944	1	4.6		NEW YORK_U.S.A.
562	-70.58	44.84	10/5/1949	20	4.7	V	MAINE_U.S.A.
825	-80.13	33.05	8/3/1959	1	4.6	VI	SOUTH CAROLINA_U.S.A
873	-88.64	37.9	6/27/1962	0	5.4	VI	ILLINOIS_U.S.A.
890	-90.05	36.64	3/3/1963	15	4.8	VI	MISSOURI_U.S.A.
1009	-90.94	37.48	10/21/1965	5	4.9	VI	MISSOURI_U.S.A.
1082	-90.44	37.44	7/21/1967	15	4.6	VI	MISSOURI_U.S.A.
1096	-89.85	38.02	3/31/1968	1	4.5	V	ILLINOIS_U.S.A.
1114	-88.37	37.91	11/9/1968	21	5.5	VII	ILLINOIS_U.S.A.
1146	-80.93	37.45	11/20/1969	5	4.6	VI	VIRGINIA_U.S.A.
1222	-80.58	33.31	2/3/1972	2	4.5	V	S.CAROLINA_U.S.A.
1233	-89.08	37	6/19/1972	13	4.5	IV	ILLINOIS_U.S.A.
1268	-83.99	35.89	11/30/1973	12	4.6	VI	TENNESSEE_U.S.A.
1279	-88.07	38.55	4/3/1974	15	4.7	VI	ILLINOIS_U.S.A.
1354	-90.48	35.58	3/25/1976	17	4.9	VI	ARKANSAS_U.S.A.
1496	-83.89	38.19	7/27/1980	16	5.2	VII	KENTUCKY_U.S.A.
1584	-71.62	43.51	1/19/1982	7	4.7	V	NEW_HAMPSHIRE_U.S.A.
1706	-74.31	44.03	10/7/1983	7	5.2	VI	NEW_YORK_U.S.A.

Plots of magnetic residual versus log10 (distance to nearest fault) and versus log10 (length of nearest fault)) are shown in Figures 5.41 and 5.42. The correlation found in the factor analysis is evident in these plots. Plots of topography versus gravity residual and of event magnitude versus azimuth of nearest stress measurement are given in Figures 5.43 and 5.44. In the former case, the negative correlation of the data is seen clearly. In the latter case, the correlation appears to be spurious, being caused by stress azimuth values of 0° for the two largest events in the dataset. These zero values might not be real measurements, and could possibly represent sites where the orientation of the maximum compressive stress was not reported in the original data.

Some major conclusions can be drawn from these results. First, the magnetic field residual contains the most information about the regional faults, while the gravity field residual does not. Second, the regional Bouguer gravity field contains much of the same information as contained in the topography data, reflecting the isostatic compensation of the crust. Third, there is no strong correlation between event magnitude and any other of the geological or geophysical variables analyzed.

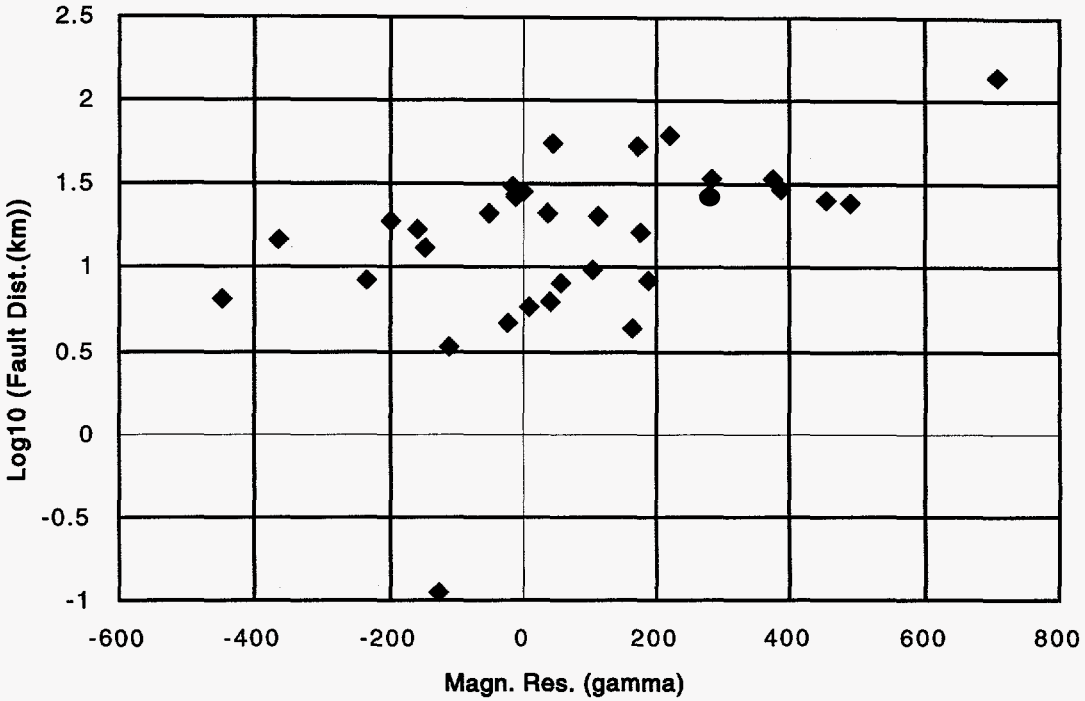

Figure 5.41. Plot of log10 (distance to the nearest fault) versus the magnetic field residual value for the dataset of M≥4.5 events.

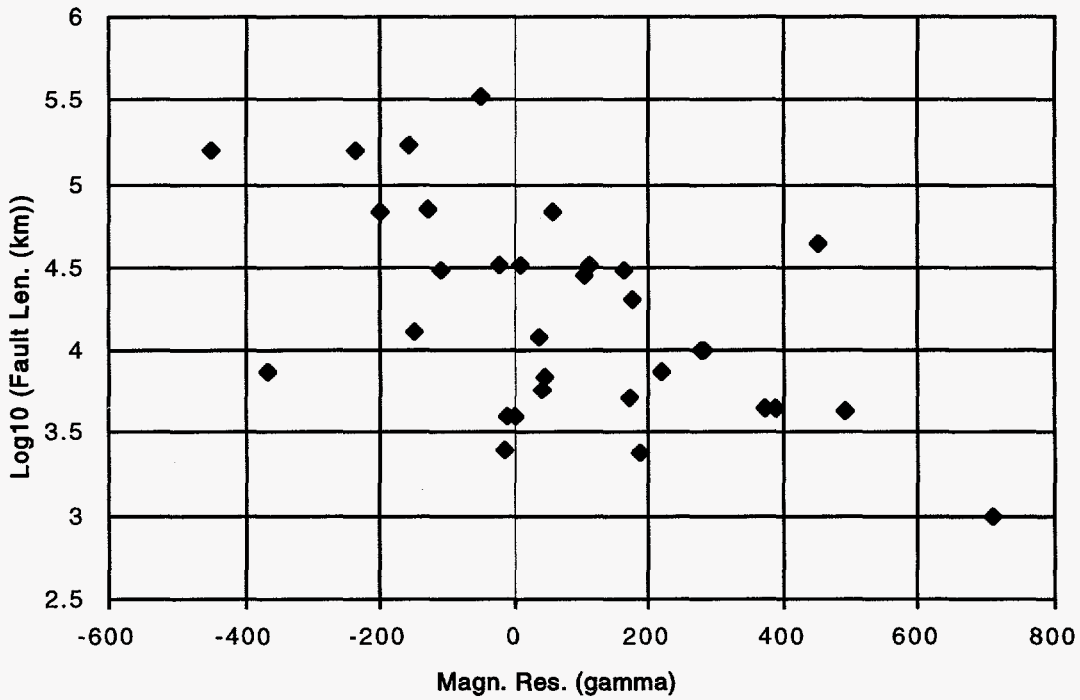

Figure 5.42. Plot of log10 (length of the nearest fault) versus the magnetic field residual value for the dataset of M≥4.5 events.

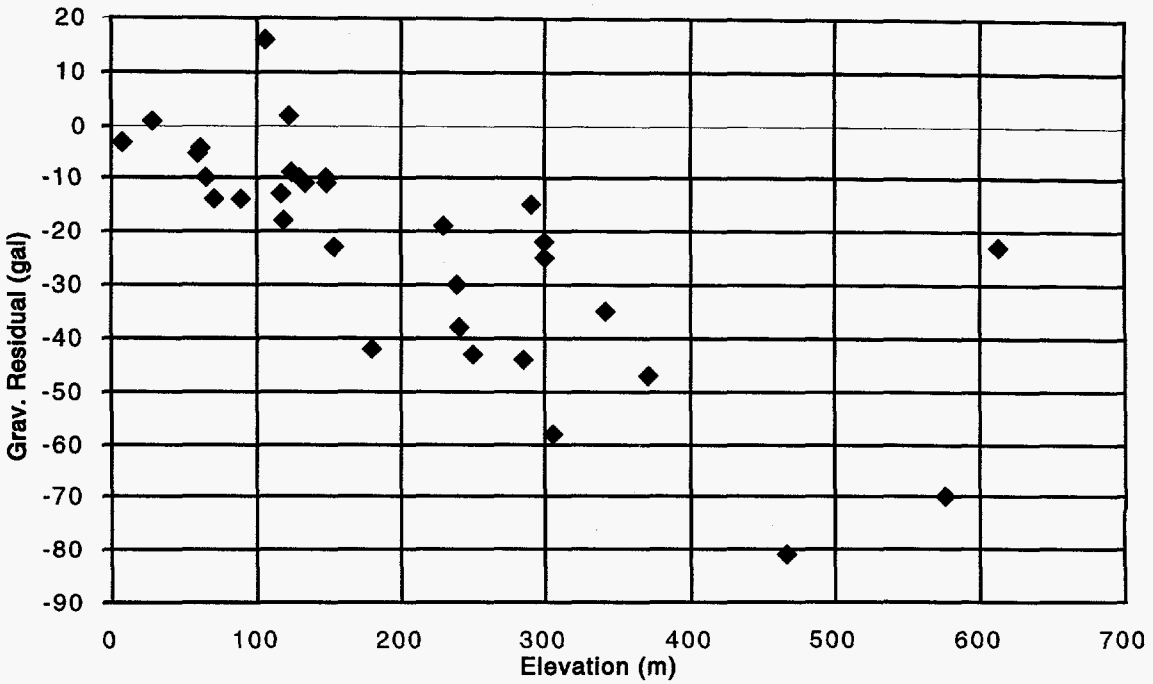

Figure 5.43. Plot of gravity residual versus the topographic elevation for the dataset of M≥4.5 events.

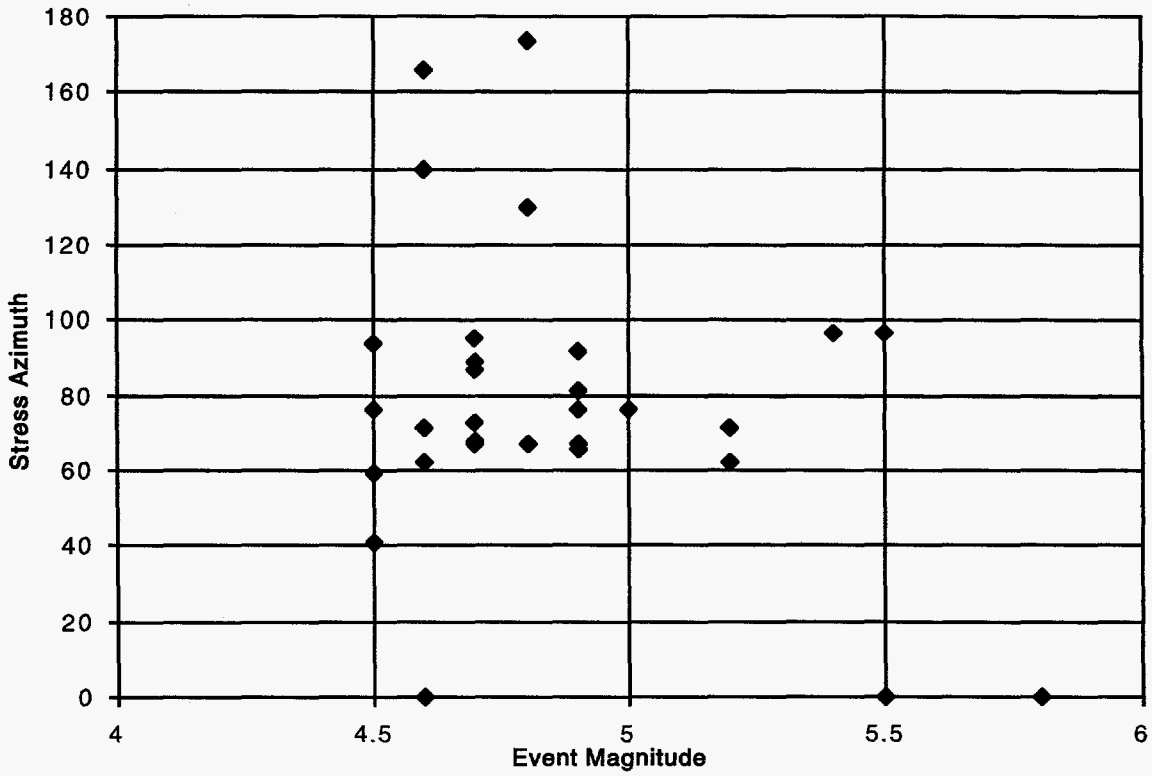

Figure 5.44. Plot of maximum compression azimuth of the nearest stress measurement versus the event magnitude for the dataset of M≥4.5 events.

Table 5.4
Factor Analysis Results for the Dataset of M≥4.5 Events

	Factor 1	Factor 2	Factor 3	Factor 4
Magnitude Excluded				
Unnormalized Dataset				
Eigenvalue	2.545	1.709	0.975	0.819
Stress Az.	-.2036	-.1456	-.0683	.9574
Elev.	.0769	-.9164	.1441	.2087
Grav. Res.	.1244	.9157	.1079	.0068
Magn. Res.	.8653	.0628	-.1346	-.0934
Log 10 (Flt. Dst.)	.6846	.4553	.3195	.0053
Log 10 (Flt. Len.)	-.9094	.1644	-.2246	.2395
Log 10 (Rvr. Dst.)	.0834	-.0083	.9636	-.0651
Normalized Dataset				
Eigenvalue	2.545	1.709	0.975	0.819
Stress Az.	-.2038	-.1459	-.0688	.9572
Elev.	.0774	-.9163	.1433	.2092
Grav. Res.	.1244	.9156	.1085	.0066
Magn. Res.	.8654	.0631	-.1347	-.0947
Log 10 (Flt. Dst.)	.6844	.4539	.3227	.0074
Log 10 (Flt. Len.)	-.8100	.1652	-.2242	.2387
Log 10 (Rvr. Dst.)	.0833	-.0081	.9634	-.0657
Magnitude Included				
Unnormalized Dataset				
Eigenvalue	2.668	1.767	1.046	0.932
Stress Az.	-.2091	-.2720	-.7744	.1051
Elev.	-.9231	.0844	-.1660	.1556
Grav. Res.	.9177	.1215	-.0519	.1234
Magn. Res.	.0652	.8414	.1381	-.1101
Log 10 (Flt. Dst.)	.4455	.6462	.0975	.3577
Log 10 (Flt. Len.)	.1344	-.8474	-.1513	-.1802
Log 10 (Rvr. Dst.)	-.0056	.0897	.1247	.9257
Magnitude	-.0787	.0580	.7901	.3094
Normalized Dataset				
Eigenvalue	2.668	1.767	1.046	0.932
Stress Az.	-.2099	-.2712	-.7742	.1051
Elev.	-.9231	.0847	-.1659	.1548
Grav. Res.	.9176	.1213	-.0515	.1246
Magn. Res.	.0655	.8417	.1388	-.1099
Log 10 (Flt. Dst.)	.4437	.6456	.0964	.3620
Log 10 (Flt. Len.)	.1352	-.8478	-.1509	-.1802
Log 10 (Rvr. Dst.)	-.0053	.0893	.1248	.9250
Magnitude	-.0792	.0579	.7897	.3101

5.6.2 Cluster Analysis

The cluster analysis was run on both the unnormalized and the normalized datasets. However, we report here only results from cluster analyses of the normalized data because those are not biased by the variables with large numerical values. Table 5.5 lists both the strongest clusters as well as several events that did not join clusters until the other events had all clustered. The most obvious observation from Table 5.5 is that most of the clusters are from events in the same region, and often from the same earthquake sequence. For instance, the two 1940 Ossipee, New Hampshire events clustered, as did the two 1944 events at Massena, New York. The 1939 and 1940 events from southern Illinois that clustered took place almost exactly 1 year apart and were of comparable sizes and locations; however, Stover and Coffman (1993) only list the 1939 event and show no M5 event in Illinois in 1940. The 1940 event may be a mistaken reentry of the 1939 southern Illinois shock in the earthquake catalog of Seeber and Armbruster (1991).

The most striking pattern in Table 5.5 is the fact that most of the clusters involve events only from one region (northeastern U.S., southeastern U.S. or central U.S.). Such a pattern was also observed by Barstow et al. (1981) in their cluster analyses. Also, all of the events that did not cluster occurred in relatively isolated pockets of seismic activity, such as western New York and Giles County, Virginia. Thus, this cluster analysis did not find any set of events displaying geological and geophysical characteristics that are common to the stronger earthquakes throughout the study region as a whole.

5.7 Analysis of Full Earthquake Dataset

The earthquake catalog for our study region (Figure 5.1) contained a total of 847 earthquakes. The magnitude range in this catalog is 3.0 to 5.8. We carried out factor analyses, cluster analyses and discriminant function analyses to see how the results for the full dataset compared to the results from the M≥4.5 dataset. Our goal in this part of the study was to see if those localities that experienced M≥4.5 events show any geological and/or geophysical characteristics that are different from other localities where only smaller events have taken place during the period of the catalog. If any significant differences can be found, then perhaps stronger earthquakes may be confined only to these localities. On the other hand, if no significant differences are found, then the possibility of a M≥4.5 event anywhere a smaller event has taken place cannot be ruled out.

5.7.1 Factor Analysis

Factor analysis was applied to both unnormalized and normalized data for the full dataset to see if the relationships among the variables found for the M≥4.5 events also exists for the full dataset. In these runs only eigenvalues of 1.0 or greater were retained in the analysis, accounting for more than 70% of the variance of the data in all cases tested. The eigenvalues and varimax-

rotated eigenvectors (or factors) are listed in Table 5.6 for runs that included the same variables (including event magnitude) as were used with the M≥4.5 data. Once again, the differences between the normalized and unnormalized data are very small.

Table 5.5
Cluster Analysis Results for M≥4.5 Dataset

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
				Cluster 1			
387	-71.37	43.87	12/20/1940	10	5.5	VII	NEW HAMPSHIRE_U.S.A.
388	-71.28	43.91	12/24/1940	8	5.5	VII	NEW HAMPSHIRE_U.S.A.
				Cluster 2			
461	-74.72	44.96	9/5/1944	12	5.8	III	NEW YORK_U.S.A.
463	-74.65	45	9/5/1944	1	4.6		NEW YORK_U.S.A.
825	-80.13	33.05	8/3/1959	1	4.6	VI	SOUTH CAROLINA_U.S.A
				Cluster 3			
365	-90.14	38.18	11/23/1939	0	4.9	V	ILLINOIS_U.S.A.
385	-90.10	38.2	11/23/1940	0	5	VI	ILLINOIS_U.S.A.
				Cluster 4			
199	-84.27	40.43	9/20/1931	5	4.7	VII	OHIO_U.S.A.
294	-84.27	40.49	3/2/1937	2	4.9	VII	OHIO_U.S.A.
296	-84.28	40.47	3/9/1937	3	4.8	III	OHIO_U.S.A.
1584	-71.62	43.51	1/19/1982	7	4.7	V	NEW_HAMPSHIRE_U.S.A.
				Cluster 5			
190	-73.78	43.47	4/20/1931	5	4.8	VII	NEW_YORK_U.S.A.
562	-70.58	44.84	10/5/1949	20	4.7	V	MAINE_U.S.A.
1009	-90.94	37.48	10/21/1965	5	4.9	VI	MISSOURI_U.S.A.
				Cluster 6			
90	-90.20	36	5/7/1927	0	4.7	VI	ARKANSAS_U.S.A.
1279	-88.07	38.55	4/3/1974	15	4.7	VI	ILLINOIS_U.S.A.
1354	-90.48	35.58	3/25/1976	17	4.9	VI	ARKANSAS_U.S.A.
				Cluster 7			
873	-88.64	37.9	6/27/1962	0	5.4	VI	ILLINOIS_U.S.A.
1096	-89.85	38.02	3/31/1968	1	4.5	V	ILLINOIS_U.S.A.
1114	-88.37	37.91	11/9/1968	21	5.5	VII	ILLINOIS_U.S.A.
				Events that did not cluster			
127	-82.83	36.11	11/3/1928	0	4.6	VI	TENNESSEE_U.S.A.
146	-78.40	42.91	8/12/1929	9	5.2	III	NEW_YORK_U.S.A.
1146	-80.93	37.45	11/20/1969	5	4.6	VI	VIRGINIA_U.S.A.
1233	-89.08	37	6/19/1972	13	4.5	IV	ILLINOIS_U.S.A.
1268	-83.99	35.89	11/30/1973	12	4.6	VI	TENNESSEE_U.S.A.

Table 5.6

Factor Analysis Results for the Full Event Dataset

Magnitude Included

Unnormalized Dataset

	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalue	1.887	1.704	1.067	1.012
Stress Az.	-.0010	.0038	.9394	-.0263
Elev.	.0987	-.9123	.0565	.0351
Grav. Res.	.0259	.9145	.0436	-.0039
Magn. Res.	.6236	.0087	-.3347	-.1654
Log 10 (Flt. Dst.)	.4840	.2368	-.2259	.2018
Log 10 (Flt. Len.)	-.7682	.1633	-.0362	.0421
Log 10 (Rvr. Dst.)	.7599	-.0376	.2184	.0931
Magnitude	.0112	-.0402	-.0113	.9670
		Normalized Dataset		
Eigenvalue	1.887	1.704	1.067	1.012
Stress Az.	-.0019	.0034	.9397	-.0264
Elev.	.0983	-.9124	.0564	.0351
Grav. Res.	.0256	.9145	.0434	-.0042
Magn. Res.	.6239	.0085	-.3338	-.1650
Log 10 (Flt. Dst.)	.4847	.2371	-.2247	.2034
Log 10 (Flt. Len.)	-.7677	.1635	-.0354	.0426
Log 10 (Rvr. Dst.)	.7597	-.0372	.2191	.0928
Magnitude	.0106	-.0406	-.0111	.9667

Two differences were found between the eigenvectors or factors found from the full dataset and those found from the M≥4.5 dataset. Once again, magnetic residual is positively correlated with log10 (distance to nearest fault) and negatively correlated with log10 (length of nearest fault), but for the full dataset it is also positively correlated with log10 (distance to nearest river). This may be due to the tendency of rivers to follow major fault trends, or it may support the hydroseismicity hypothesis that water percolating into fault zones helps trigger intraplate seismicity in the CEUS. The second difference between the full dataset and the M≥4.5 datasets is that in the full dataset, magnitude is quite independent of the other variables, as is the azimuth of the closest stress measurement. The inverse relationship between these two variables in the M≥4.5 data disappears when the full dataset is analyzed.

Because we were concerned that the factor analysis results of the full dataset might be biased by the inclusion of the events of M≥4.5, we did a second set of factor analyses (normalized and unnormalized data) using a dataset with only events of M<4.5. Those results are listed in Table 5.7. The results are very close to those from the full dataset, a clear indication that the M≥4.5 events are not biasing the factor analysis results from the full dataset.

Table 5.7
Factor Analysis Results for the Event Dataset of M<4.5

	Magnitude Included			
	Unnormalized Dataset			
	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalue	1.885	1.687	1.074	1.013
Stress Az.	.0029	.0107	.9398	-.0114
Elev.	.0957	-.9133	.0529	.0410
Grav. Res.	.0241	.9142	.0457	.0063
Magn. Res.	.6186	.0067	-.3274	-.1257
Log 10 (Flt. Dst.)	.4847	.2256	-.2330	.1959
Log 10 (Flt. Len.)	-.7626	.1651	-.0576	.0946
Log 10 (Rvr. Dst.)	.7662	-.0401	.2135	.0781
Magnitude	-.0187	-.0383	-.0003	.9710
	Normalized Dataset			
Eigenvalue	1.886	1.687	1.079	1.008
Stress Az.	.0014	.0099	.9397	-.0208
Elev.	.0946	-.9132	.0540	.0351
Grav. Res.	.0229	.9140	.0475	-.0006
Magn. Res.	.6194	.0064	-.3316	-.1266
Log 10 (Flt. Dst.)	.4845	.2262	-.2264	.2019
Log 10 (Flt. Len.)	-.7626	.1653	-.0522	.0929
Log 10 (Rvr. Dst.)	.7665	-.0398	.2151	.0772
Magnitude	-.0171	-.0401	-.0068	.9700

5.7.2 Cluster Analysis

A number of cluster analyses were carried out on subsets of the full dataset to see if there are similar geological and geophysical observables between the larger events in one region and any events from the smaller earthquake seismicity in another region. For the purposes of these tests, we divided our study area into four subregions: northeast (north of 40°N and east of 80°W), southeast (south of 40°N and east of 86°W), New Madrid (south of 39°N and west of 86°W), and northcentral (north of 39°N and west of 80°W). We carried out several tests where we clustered the M≥4.5 events in one region with all the seismicity in another region. Cases where large events from one region cluster with events in another region could indicate possible locations of larger earthquakes in the second subregion. Once again, only the normalized data were analyzed in the cluster analyses.

Four cluster analyses were carried out: M≥4.5 events in the northeast with all the the seismicity in the southeast, M≥4.5 events in the southeast with all the the seismicity in the northeast, M≥4.5 events in New Madrid with all the the seismicity in the northcentral subregion, and M≥4.5 events in New Madrid with all the the seismicity in the southeast. The first two tests looked for similarities in the seismicity along the Atlantic coast and Appalachian trends. The third test looked for indications that the New Madrid seismicity may extend northeast into Indiana, Kentucky and Ohio, while the final test looked for clustering in the southeastern part of the continent.

In general, all four cluster analyses showed that the M≥4.5 events within one subregion readily clustered among themselves but only a few weak clusters were found with the events from other subregions. Table 5.8 shows that there was only one case where a large event in the northeastern subregion clustered with events in the southeast subregion. Curiously, that cluster paired the 1944 M5.8 earthquake at Massena, New York with events at the Charleston, South Carolina area. There were two cases where M≥4.5 events in the southeast subregion clustered with events in the northeast (Table 5.9). In one cluster, a larger event at Charleston, South Carolina again paired with events at Massena, New York. In the other cluster an event in southcentral South Carolina, northwest of the Charleston area, paired with an event in southeastern Massachusetts.

One weak cluster was found between M≥4.5 events in the New Madrid subregion and events in the northcentral subregion (Table 5.10). That cluster paired the 1939 southern Illinois earthquake (with its 1940 mistaken reentry, as noted earlier) with two small earthquakes at a locality in central Ohio. Several clusters emerged when M≥4.5 events in the New Madrid subregion were analyzed with the seismicity in the southeastern subregion (Table 5.11). Most of these clusters emerged well into the analysis, so the similarities of the observables for the events in each of these clusters is not necessarily very great.

Table 5.8

Cluster Analysis Results for M≥4.5 Events in the Northeast Subregion with all Events in the Southeast Subregion

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
				Cluster 1			
461	-74.72	44.96	9/5/1944	12	5.8	III	NEW_YORK_U.S.A.
463	-74.65	45	9/5/1944	1	4.6		NEW_YORK_U.S.A.
450	-80.2	33	12/28/1943	0	3.3	IV	SOUTH_CAROLINA_U.S.A
896	-80.19	32.97	5/4/1963	5	3.3	IV	SOUTH_CAROLINA_U.S.A
1269	-80.27	32.97	12/19/1973	6	3	III	SOUTH_CAROLINA_U.S.A
1312	-80.22	33	4/28/1975	12	3	IV	SOUTH_CAROLINA_U.S.A

Table 5.9

Cluster Analysis Results for M≥4.5 Events in the Southeast Subregion with all Events in the Northeast Subregion

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
				Cluster 1			
825	-80.13	33.05	8/3/1959	1	4.6	VI	SOUTH_CAROLINA_U.S.A
462	-74.9	44.98	9/5/1944	0	3.4		NEW_YORK_U.S.A.
463	-74.65	45	9/5/1944	1	4.6		NEW_YORK_U.S.A.
464	-74.9	44.98	9/5/1944	0	3.3		NEW_YORK_U.S.A.
468	-74.9	45	10/31/1944	0	4		NEW_YORK_U.S.A.
499	-74.88	44.9	9/4/1946	0	3		NEW_YORK_U.S.A.
507	-74.9	44.9	12/25/1946	0	3		NEW_YORK_U.S.A.
745	-74.9	44.9	2/20/1957	0	3.3	IV	NEW_YORK_U.S.A.
769	-74.9	44.9	1/11/1958	0	3.3	IV	NEW_YORK_U.S.A.
848	-74.8	45	4/20/1961	0	3.2	V	NEW_YORK_U.S.A.
853	-74.9	44.9	9/29/1961	0	3.1	IV	NEW_YORK_U.S.A.
936	-74.9	44.9	3/29/1964	0	4.3	V	NEW_YORK_U.S.A.
				Cluster 2			
1222	-80.58	33.31	2/3/1972	2	4.5	V	SOUTH_CAROLINA_U.S.A
106	-71.6	41.2	1/13/1928	0	3.3	IV	MASSACHUSETTS_U.S.A.

Table 5.10

Cluster Analysis Results for M≥4.5 Events in the New Madrid Subregion with all Events in the Northcentral Subregion

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
				Cluster 1			
365	-90.14	38.18	11/23/1939	0	4.9	V	ILLINOIS_U.S.A.
385	-90.10	38.2	11/23/1940	0	5	VI	ILLINOIS_U.S.A.
276	-83.2	41.2	1/31/1936	0	3.1	IV	OHIO_U.S.A.
1306	-83.2	41.3	2/3/1975	0	3.3	IV	OHIO_U.S.A.

Table 5.11

Cluster Analysis Results for M≥4.5 Events in the New Madrid Subregion with all Events in the Southeast Subregion

Event #	Long. (E)	Lat. (N)	Date	Depth (km)	Mag.	Int.	Location
				Cluster 1			
1009	-90.94	37.48	10/21/1965	5	4.9	VI	MISSOURI_U.S.A.
928	-85.39	34.67	2/18/1964	1	3.3	V	TENNESSEE_U.S.A.
1745	-85.2	34.75	10/9/1984	12	4.2	VI	TENNESSEE_U.S.A.
				Cluster 2			
204	-89.80	34.1	12/17/1931	0	4.7	VII	MISSISSIPPI_U.S.A.
1184	-80.66	33.36	5/19/1971	1	3.7	IV	S.CAROLINA_U.S.A.
1196	-80.63	33.34	7/31/1971	4	3.8	III	S.CAROLINA_U.S.A.
1198	-80.7	33.4	8/11/1971	0	3.5		S.CAROLINA_U.S.A.
1222	-80.58	33.31	2/3/1972	2	4.5	V	S.CAROLINA_U.S.A.
1223	-80.58	33.46	2/7/1972	0	3.2		S.CAROLINA_U.S.A.
1224	-80.58	33.46	2/7/1972	0	3.2		S.CAROLINA_U.S.A.
1396	-80.7	33.37	8/4/1977	9	3.1		S.CAROLINA_U.S.A.
1399	-80.69	33.39	8/25/1977	0	3.1	V	S.CAROLINA_U.S.A.
				Cluster 3			
873	-88.64	37.9	6/27/1962	0	5.4	VI	ILLINOIS_U.S.A.
485	-81.38	33.75	7/26/1945	5	4	V	S.CAROLINA_U.S.A.
				Cluster 4			
1114	-88.37	37.91	11/9/1968	21	5.5	VII	ILLINOIS_U.S.A.
1354	-90.48	35.58	3/25/1976	17	4.9	VI	ARKANSAS_U.S.A.
1387	-78.63	37.9	2/27/1977	0	3.4	V	VIRGINIA_U.S.A.
				Cluster 5			
32	-88.20	38	4/27/1925	0	4.9	VII	ILLINOIS_U.S.A.
1655	-84.89	32.64	10/31/1982	0	3.1		GEORGIA_U.S.A.

Over all, the larger earthquakes in one subregion tend not to cluster with the seismicity in other regions. A somewhat similar conclusion was found by Barstow et al. (1981). In particular, the cluster analyses do not support the notion that the New Madrid seismicity can be linked in a continuous band through Indiana and Ohio to the seismicity along the St. Lawrence River, as suggested by Woolard (1969). On the other hand, the clustering of events in the Charleston, South Carolina area with events at Massena, New York may indicate that an event like the 1886 M7 Charleston earthquake could be possible in northern New York state.

5.7.3 Discriminant Function Analysis

The strategy followed in the discriminant function analysis of the full dataset was to divide the events in the catalog into two groups, one group with M≥4.5 (the 33 events of Table 4.3) and the second group with M<4.5 (814 events). Our goal was to see if there was some discriminant function that could be found that would separate the locations of the stronger events from those only smaller events. Such a discriminant would be useful in identifying which locations may have an enhanced probability of a large earthquake.

While an almost identical discriminant function was found from the normalized and the unnormalized data (Table 5.12), it badly failed the chi-squared test, with almost a 60% probability that the two groups cannot be discriminated by the function. A plot of the discriminant function values for the events in the analysis shows complete overlap between the groups of events with virtually identical means (Figure 5.45). Thus, this analysis indicates that there are not differences, at least among the variables tested here, between localities that have experienced larger events and those localities that have only had smaller events during the period of the catalog we used.

Table 5.12

Results of the Discriminant Function Analysis on the Full Event Dataset

	Unnormalized Data Discriminant Function Weights	Normalized Data Discriminant Function Weights
Stress Az.	.1121	.1132
Elev.	.1351	.1380
Grav. Res.	.3100	.3139
Magn. Res.	.4835	.4831
Log 10 (Flt. Dst.)	-.1864	-.1896
Log 10 (Flt. Len.)	.6685	.6704
Log 10 (Rvr. Dst.)	-.5162	-.5121
Chi-square value	5.728	5.695
Significance	.5718	.5758

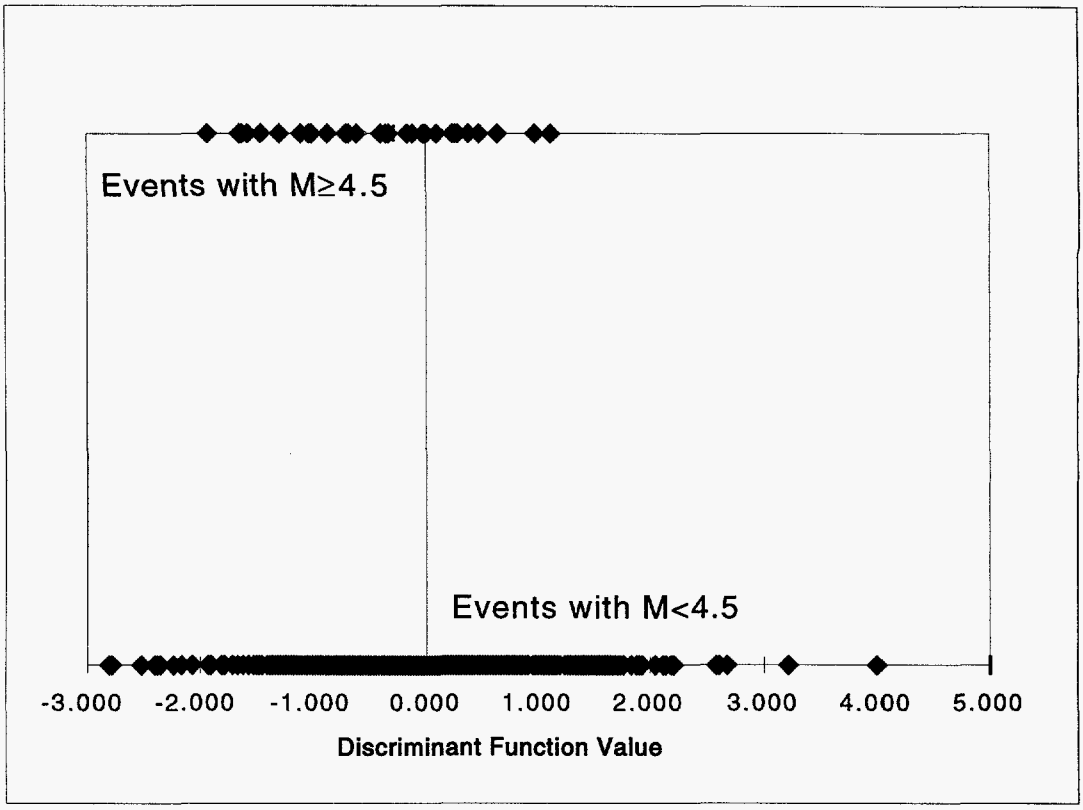

Figure 5.45. Plot of the discriminant function scores for the events with M≥4.5 and M<4.5 from the full dataset. The normalized data were used in the plot, along with the discriminant function from Table 5.12.

5.8 Analysis of Spatial Cells Dataset

In order to compare the geological and geophysical characteristics of seismically active and inactive parts of our study area, we followed the lead of Barstow et al. (1981) by dividing the region into cells and conducting multivariate statistical analyses on the observables within those cells. While Barstow et al. (1981) used circular cells, we found it easier with the GIS system to divide the area into roughly rectangular cells (.5° x .5°) following the geographic coordinate system (Figure 5.46). We then found the center of each cell and measured our geological and geophysical observables from the cell centers. For each cell we accumulated the values of the observables as listed in Table 5.2. We then classified each cell as seismic if it contained one or more epicenter from our catalog, while it was classed as non-seismic if it did not contain any such events. Because our catalog only contained events of M3.0 or greater, any cell that has experienced earthquakes where the largest event was less than magnitude 3.0 in the NCEER catalog is classified as non-seismic by the definition we applied here. A total of 864 cells were included, 103 of which were classified as active in our analysis. We conducted factor analyses and discriminant function analyses on both unnormalized and normalized cell datasets.

5.8.1 Factor Analysis

As as was done in the earlier factor analyses, we analyzed both normalized and unnormalized datasets. Seven variables were included in these computations: maximum compressive stress direction, topographic elevation, gravity residual, magnetic residual, log10 (distance to the nearest fault), log10 (length of the nearest fault), and log10 (distance to the nearest river). We found five eigenvectors with eigenvalues greater than 0.7 accounting for more than 80% of the variance of the data. These along with their corresponding eigenvectors are listed in Table 5.13.

Table 5.13
Factor Analysis Results for the Grid Cell Dataset

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
			Unnormalized Dataset		
Eigenvalue	1.836	1.317	1.037	0.897	0.749
Stress Az.	.0315	-.0011	.0231	.9832	.0102
Elev.	-.7670	-.2696	.0844	.1026	.0468
Grav. Res.	.8577	.0391	.1239	.1274	.0472
Magn. Res.	.0036	.0380	-.0415	.0101	.9963
Log 10 (Flt. Dst.)	.1646	.6271	-.5047	-.1953	.0014
Log 10 (Flt. Len.)	.0784	.0653	.9251	-.0095	-.0430
Log 10 (Rvr. Dst.)	.1804	.8731	.1637	.0839	.0467
			Normalized Dataset		
Eigenvalue	1.661	1.574	1.247	1.020	0.780
Stress Az.	.0839	.0057	-.0039	.9721	-.0351
Elev.	-.8185	-.0014	-.0019	-.4339	-.1780
Grav. Res.	.8236	-.0019	-.0041	-.1761	-.4256
Magn. Res.	-.0884	-.0058	.0053	-.1342	.9724
Log 10 (Flt. Dst.)	-.0008	-.0581	.9461	-.0022	.0055
Log 10 (Flt. Len.)	.0021	-.6420	-.6228	.0047	-.0016
Log 10 (Rvr. Dst.)	.0004	.9434	-.0716	.0080	-.0064

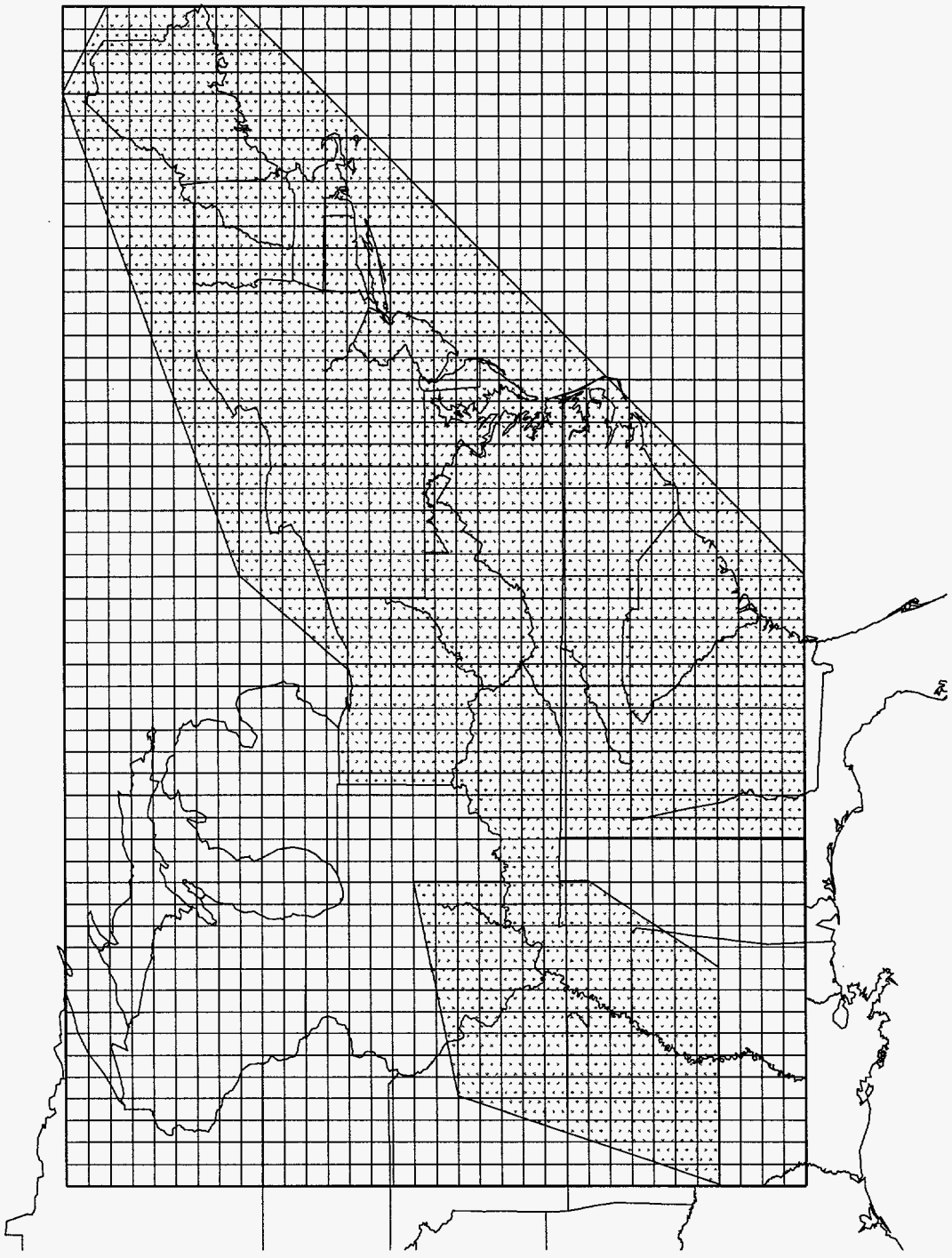

Figure 5.46. GIS representation of the 0.5° x 0.5° grid cell pattern for the statistical analyses. Dotted area indicates the region where the statistical analysis was performed.

In contrast to the factor analysis results reported in the previous sections, there were some differences in the results between the normalized and unnormalized cell datasets. For the unnormalized data, azimuth of the nearest stress measurement, magnetic field residual, and log10 (length of the nearest fault) are all quite independent of the other observables. The negative correlation of topographic elevation and gravity residual is again seen, while the fifth varimax-rotated eigenvector indicates that log10 (distance to nearest fault) is directly correlated with log10 (distance to the nearest river). In the case of the normalized dataset, azimuth of the nearest stress azimuth shows a weak negative correlation with topographic elevation, while magnetic field residual shows a weak negative correlation with gravity residual. Gravity residual and topographic elevation are again inversely related, as are log10 (distance to the nearest river) with log10 (length nearest fault) and log10 (distance to nearest fault) with log10 (length nearest fault).

These results confirm some of the results seen earlier, namely the independence of the stress direction from the geology and geophysics and the inverse relationship between gravity residual and elevation. They also provide statistical evidence that rivers and large faults are related. On the other hand, they do not show that either gravity or magnetic residuals are related to nearby major faults. The statistical independence of gravity residuals and faults was also found in the factor analyses reported above, but in those same analyses some relationship between magnetic residuals and faults was observed. This difference between the factor analysis results for the event and the grid cell datasets hints that CEUS earthquakes tend to associate with faults that show strong magnetic residuals.

5.8.2 Discriminant Function Analysis

The objective of the discriminant function analysis was to find a linear function of the observables that most separated seismic from non-seismic cells. Except for insignificant differences, the same discriminant function was found from both the unnormalized and the normalized data (Table 5.14). The chi-squared fit of the data showed that the null hypothesis (the two different cells types cannot be discriminated) is rejected with better than 99% confidence. The weights of most of the variables in the discriminant function are of nearly the same size. The only exceptions are the gravity residual, which has a higher weight, and the magnetic residual, which is given very little weight by the analysis. Figure 5.47 shows a plot of the discriminant function scores of each of the cells analyzed. The mean value for the seismic cells is about 0.6 while that for the non-seismic cells is very close to 0. Unfortunately, there is a great deal of overlap in the discriminant function scores between the two groups. Even though the chi-squared value suggests that the function is able to discriminate between the two populations, it is not easy to characterize a grid cell as either seismic or non-seismic based on its discriminant function score. A grid cell with a discriminant function score below -2 is most likely non-seismic, while one with a disrciminant function score above +3 is more likely than not to be seismic. However, a score of 0.5 could be either seismic or non-seismic.

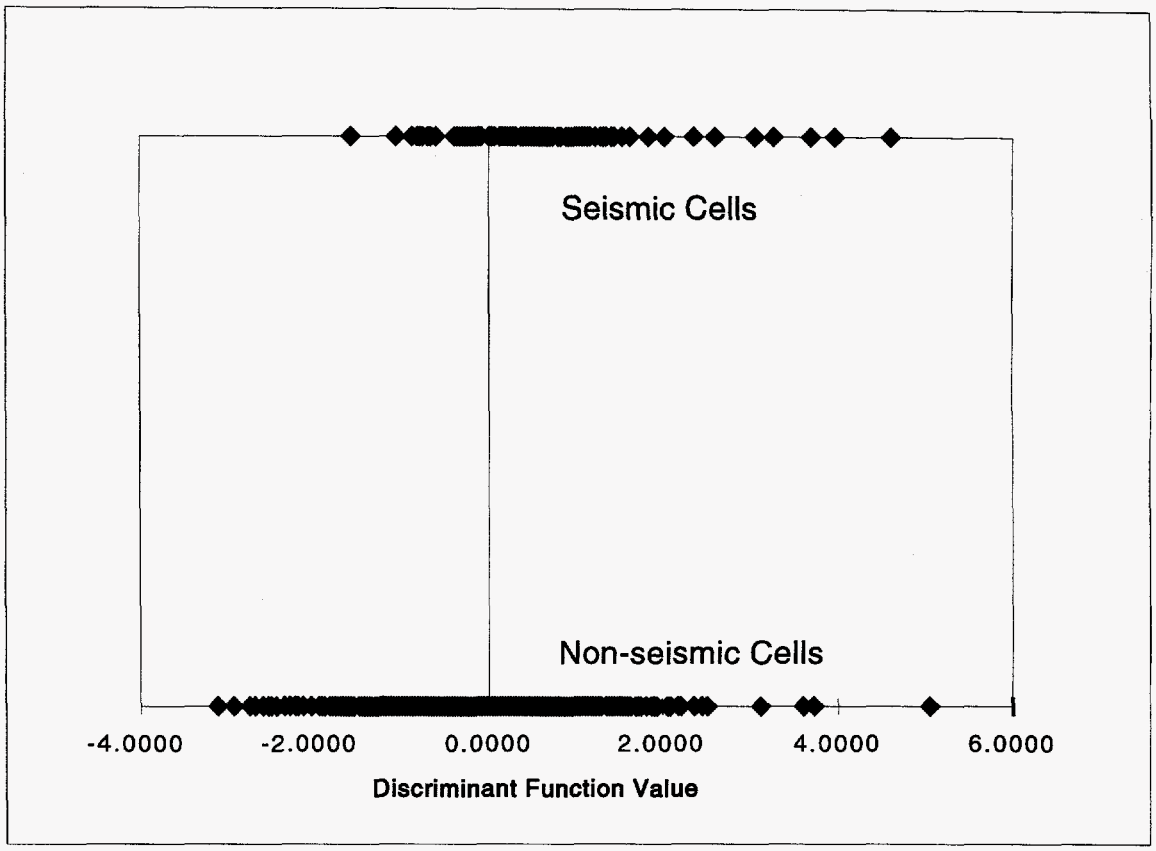

Figure 5.47. Plot of the discriminant function scores for the seismically active grid cells and the inactive grid cells (as defined in this study). The normalized data were used in the plot, along with the discriminant function from Table 5.14.

While the chi-squared result indicates a robust discrimination among the two types of cells, this can be tested further by randomly reassigning the cell types and redoing the analysis. This allows us to test whether or not the discriminant function found some spurious pattern in the data. We generated three additional cell datasets, where we randomly reassigned the activity parameter (1=active, 0=inactive) to the cells. In this process we kept the same ratio of active to inactive cells (103 active to 761 inactive). We then ran the discriminant function analysis on each of these randomized datasets.

The discriminant functions and chi-squared significance values for the three randomized cell datasets are given in Table 5.15. Again, there is virtually no difference between the results of the normalized and of the unnormalized data. The discriminant functions are different from that of the real data, and the chi-squared fits are not as good (about 50% chance of nondiscrinimation in one case and about 25% chance of nondisrcrimination in the other two cases). We interpret this test as indicating that the discriminant function found from the real dataset is not a chance result but rather does show some tendencies toward real differences between seismic and non-seismic grid cells.

Table 5.14
Results of the Discriminant Function Analysis on the Grid Cell Dataset

	Unnormalized Data Discriminant Function Weights	Normalized Data Discriminant Function Weights
Stress Az.	.4610	.4609
Elev.	-.3177	-.3186
Grav. Res.	-.7207	-.7209
Magn. Res.	.0936	.0944
Log 10 (Flt. Dst.)	.3703	.3709
Log 10 (Flt. Len.)	-.2378	-.2372
Log 10 (Rvr. Dst.)	.4641	.4631
Chi-square value	55.578	45.496
Significance	.0000	.0000

Table 5.15
Results of the Discriminant Function Analysis on the Randomized Grid Cell Dataset

Randomization 1

	Unnormalized Data Discriminant Function Weights	Normalized Data Discriminant Function Weights
Stress Az.	.1536	.1562
Elev.	.2112	.2120
Grav. Res.	.8692	.8690
Magn. Res.	-.3159	-.3148
Log 10 (Flt. Dst.)	-.4029	-.4010
Log 10 (Flt. Len.)	.0049	.0067
Log 10 (Rvr. Dst.)	.1591	.1595
Chi-square value	8.671	8.642
Significance	.2772	.2794

Randomization 2

	Unnormalized Data Discriminant Function Weights	Normalized Data Discriminant Function Weights
Stress Az.	.3450	.3447
Elev.	-.3113	-.3411
Grav. Res.	.2698	.2695
Magn. Res.	-.2486	-.2473
Log 10 (Flt. Dst.)	.4875	.4888
Log 10 (Flt. Len.)	.6489	.6489
Log 10 (Rvr. Dst.)	-.3322	-.3319
Chi-square value	7.076	7.044
Significance	.4210	.4243

Randomization 3

	Unnormalized Data Discriminant Function Weights	Normalized Data Discriminant Function Weights
Stress Az.	.2968	.2954
Elev.	-.1431	-.1461
Grav. Res.	-.6022	-.6011
Magn. Res.	-.0402	-.0407
Log 10 (Flt. Dst.)	-.0070	-.0101
Log 10 (Flt. Len.)	-.2142	-.2152
Log 10 (Rvr. Dst.)	.8974	.8980
Chi-square value	9.098	9.067
Significance	.2457	.2479

5.9 Results of Statistical Analyses

Several major conclusions can be drawn from the results of the statistical analyses. First, a linear combination of the geological and geophysical variables analyzed here has a slight ability to discriminate between seismic and non-seismic cells in the CEUS. If the seismic cells (as defined here) are those where larger earthquakes are most likely to take place, then the discriminant function may be helpful in seismic zonations of the CEUS for seismic hazard analyses. A second conclusion is that the statistical analyses failed to find any geological or geophysical characteristics differentiating those localities that experienced larger ($M \geq 4.5$) earthquakes (between 1924 and 1984) from other localities that only had smaller earthquakes. This result is consistent with the notion that a larger earthquake can occur in any part of the study area that has experienced minor earthquake activity.

A third conclusion from the factor analyses is that no combination of the geological and geophysical observables at a seismically active locality gives any clue about how large an event could occur at that locality. The possibility of larger magnitude events at a locality cannot be assessed with the variables we examined. Finally, the geological and geophysical characteristics of the earthquakes in each subregion of the study area show little resemblance to those characteristics of the earthquake locations in any other of the subregions. Thus, the relationship of the seismicity with the local geology in one subregion probably contains little information that can be used to identify seismically active geologic features in another subregion.

Some of these conclusions support the results reported by Barstow et al. (1981). They also found discriminant functions that separated the seismic and non-seismic cells in their analysis. While their set of observables was quite different from ours, the two studies taken together show that the regional geologic structures, and probably the past geologic history, do play a controlling role in determining which areas in the CEUS are seismically active. In particular, those variables in the two studies that are most important for discriminating seismic versus non-seismic cells appear to include: gravity residual; magnetic residual; distance to nearest river, fold, fault, and arch; length of nearest fault; and basement elevation. All of these observables can depend directly or indirectly on the existence of major basement faults, supporting the general idea that preexisting faults or other zones of basement weakness are rupturing due to the modern tectonic stress field. The lack of dependence of stress direction on location of the earthquake activity is evidence that the stress driving the earthquakes is the regional plate tectonic stress field. Finally, both Barstow et al. (1981) and this study found that the observables for the earthquakes in one subregion do not cluster well with those from another region. Our one exception to this general conclusion is that the Massena, New York area did cluster in two separate tests with the Charleston, South Carolina area. Both of these localities experienced damaging earthquakes within the last 120 years.

6. Discussion and Conclusions

The manipulation and analysis of the many data types necessary to address the problem of investigating what tectonic features control where large earthquakes will occur in the CEUS is a significant problem in database management. Furthermore, all of the information is map-based, so they require a database system capable of handling geographic data. In this study, we demonstrated how a Geographic Information System (GIS) can be used to aid in investigating the correlation of seismicity with tectonic features in the CEUS. The major advantage of a GIS is that it allows the machine manipulation of large geographic datasets, such as those of geological and geophysical observations. Subsets or combinations of data can be extracted and analyzed to look for hidden patterns and relationships within the data. Many different combinations of subsets of data from the database can be analyzed relatively easily. Analyses of the data can be redone if new or additional data are obtained, and analyses can be easily reproduced to verify results.

GIS provides a means to easily and quickly manipulate spatial geoscience data. Conversions between vector and raster data formats, distances between geographic features and coordinate system transformations have proven very useful for this project. Consider, for example, the problem of locating the nearest fault to a an earthquake epicenter. Traditionally, complex search algorithms would have to be constructed. With GIS, these parameters are automatically computed when the the fault database tables are "joined" to the earthquake database tables. So, using GIS becomes a matter of data manipulation by function, not by direct computation. Furthermore, raw data can be easily converted to a more compact geographical representation. For example, topographic point data consists of 650,000 points in our study area. However, we can quickly represent these data as contours, surface maps or shaded-relief maps and edit them spatially (e.g., include all events in area A and B but not C). These spatial representations and manipulations are a powerful analysis tool that is available through GIS.

Although GIS systems can be very helpful in solving the types of problems adressed in this study, GIS systems are not simply turnkey operations, but they demand highly trained and knowledgeable experts to operate them properly. Critical to the successful application of a GIS is the proper planning and creation of the database. Since a strength of a GIS system is the ability to query the database to extract desired subsets of the data, one must plan the database so that the proper data subsets can be extracted in the query process. The spatial resolution of the datasets is also an important consideration; it must be detailed enough to show the desired information but low enough resolution to be efficient for GIS program execution and disk storage. The uniformity of the coverage of the data across the area of interest is also of concern, especially when mathematical analyses are to be performed on the data.

The objective of this study was to use the GIS technology to provide the infrastrucuture for a quantitative assessment of the potential for a given geological or geophysical feature (or combination of features) in the CEUS to generate large earthquakes. Our approach was to use

multivariate statistical analyses of the datasets we have accumulated in our GIS database to look for evidence of active structures. As part of this work, we also are learned about the practical and inherent advantages and limitations of using GIS for multivariate statistical analysis. Thus, this research not only generates results in its own right, but also provides guidance on how to better handle and analyze such extensive datasets in future studies.

Several major conclusions can be drawn from the results of the statistical analyses. Perhaps the most important is that multivariate statistical analysis techniques can yield information about seismotectonically active structures in the CEUS. Both Barstow et al. (1981) and this study found discriminant functions that separate, albeit with significant overlap, seismic cells from non-seismic cells. Since each study used different sets of variables and different cells, it is not clear which variables contribute the most discriminant information. This is a technique that should be explored further with new and better datasets.

Another important conclusion is that regional geologic structures, and probably the past geologic history, do play a role in determining which areas in the CEUS are seismically active. The statistical analyses in this study and that of Barstow et al. (1981) represent tools that begin to define some of the characteristics of which geologic structures are capable of generating earthquakes. Those characteristics include such observables as distance to the nearest river, magnetic field residual, and proximity to folds, faults and plutons. Many of the observables used in this study or by Barstow et al. (1981) indicate the existence of major basement faults, supporting the general idea that preexisting faults or other zones of basement weakness are rupturing due to the modern tectonic stress field.

Any hypothesis that claims to explain the potential for large earthquakes must ultimately be demonstrated to explain the occurrence of known earthquakes (including paleoseismic events). In this study, we set up several ways to test whether or not a given hypothesis can be shown to be associated with the actual occurrences of earthquakes. The hypothesis tests that we have assembled and described in this report represent only one set of possible examples of how our GIS system can be used for such purposes. Datasets with more complete spatial coverage, finer resolution, newer data, and other geological and geophysical parameters than those used in this study could well reveal more differences between seismic and non-seismic areas. They may even be capable of discriminating between areas with the potential of generating strong earthquakes and those without such a potential. With the assembly of our current GIS system, future studies can extend the work done here.

7. References

Barstow, N.L., K.G. Brill, Jr., O. W. Nuttli, and P.W. Pomeroy. 1981. An Approach to Seismic Zonation for Siting Nuclear Electric Power Generating Facilities in the Eastern United States. U.S. Nuclear Regulatory Commission publication NUREG/CR-1577, 143 pp.

Bernreuter, D.L., J.B. Savy, R.W. Mensing, and J.C. Chen. 1989. Seismic Hazard Characterization of 69 Nuclear Power Plant Sites East of the Rocky Mountains, U.S. Nuclear Regulatory Commission, Technical Report NUREG/CR-5250, Prepared by Lawrence Livermore National Laboratory.

Budnitz, R.J., G. Apostolakis, D.M. Boore, L.S. Cluff, K.J. Coppersmith, C.A. Cornell and P.A. Morris. 1997. *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts,* U.S. Nuclear Regulatory Commission, Technical Report NUREG/CR-6372, Vol. 1, Prepared by Lawrence Livermore National Laboratory.

Cooley, W.W. and P.R. Lohnes. 1971. *Multivariate Data Analysis,* John Wiley & Sons, New York, 364 pp.

Costain, J.K., G.A. Bollinger, and J.A. Speer. 1987. "Hydroseismicity: A Hypothesis for the Role of Water in the Generation of Intraplate Seismicity." *Seismological Research Letters,* 58, No. 3, 41-64.

Davis, J. C. 1986. *Statistics and Data Analysis in Geology* (2nd Edition), John Wiley and Sons, New York, 646 pp.

Ebel, J.E. and J.A. Spotila. 1992. The Relationship Between Earthquakes and Regional Geology in New England, EOS, *Trans. Am. Geophys. Un.*, 74, 288.

Electric Power Reserarch Institute. 1986. *Seismic Hazard Methodology for the Central and Eastern United States*, 10 volumes, EPRI Report NP-4726, Electric Power Reserarch Institute, Palo Alto, CA.

Frankel, A. 1995. Mapping Seismic Hazard in the Central and Eastern United States, *Seismological Research Letters,* 66(4), 8-21.

Johnston, A.C. 1989. The Seismicity of 'Stable Continental Interiors', in *Earthquakes at North-Atlantic Passive Margins: Neotectonics and Postglacial Rebound*, S. Gregersen and P.W. Basham, (eds.), 299-327.

7. References

Johnston, A.C., K. J. Coppersmith, L. R. Kanter, and C.A. Cornell. 1994. The Earthquakes of Stable Continental Regions, *Electric Power Research Institute Report*, TR-102261-VI, Palo Alto, California.

Kafka, A.L. and P.E. Miller. 1996. Seismicity in the Area Surrounding Two Mesozoic Rift Basins in the Northeastern United States, *Seismological Research Letters*, 67(3), 69-86.

Kane, M.F. 1977. Correlations of Major Eastern Earthquake Centers with Mafic/Ultramafic Basement Masses, *U.S. Geological Survey Professional Paper 1028.*, pp. 199-204.

King, P.B. and H. M. Beikman. 1974. Geologic Map of the United States, *U.S. Geological Survey Professional Paper 901*, 40 pp.

Leet, L.D. 1942. Mechanics of Earthquakes Where There is No Surface Faulting, *Bull. Seism. Soc. Am.*, 32, 93-96.

Press, F., and R. Siever. 1978. *Earth* (2nd Edition), W. H. Freeman and Co., San Francisco, 649 pp.

Schruben, P.G., R.E. Arndt, and W.J. Bawiec. 1994. Geology of the Conterminous United States at 1:2,500,000 Scale - A Digital Representation of the 1974 P.B. King and H.M. Beikman Map, *U.S. Geological Survey Digital Data Series DDS-11* (CD disk).

Seeber, L. and J.G. Armbruster. 1991. *The NCEER-91 Earthquake Catalogue: Improved Intensity-based Magnitudes and Recurrence Relations for U.S. Earthquakes East of New Madrid,* National Center for Earthquake Engineering Research, NCEER-91-0021.

Stover, C.W. and J.L. Coffman. 1993. Seismicity of the United States 1568-1989, *U.S. Geological Survey Profressional Paper 1527,* 418 pp.

Sykes, L.R. 1978. Intraplate Seismicity, Reactivation of Preexisting Zones of Weakness, Alkaline Magmatism, and Other Tectonism Postdating Continental Fragmentation, *Rev. Geophys. Space Phys.,* 16, 621-688.

Talwani, P., 1988. The Intersection Model for Intraplate Earthquakes, *Seism. Res. Lett.,* 59, 305-310.

Telford, W.M., L.P. Geldart, R.E. Sheriff, and D.A. Days. 1976. *Applied Geophysics.* Cambridge U. Press, New York, 860 pp.

Wheeler, R.L. 1985. Evaluating Point Concentrations on a Map: Earthquakes in the Colorado Lineament, *Geology*, 31, 701-704.

Wheeler, R.L., 1996. Earthquakes and the Southeastern Boundary of the Intact Iapetan Margin in Eastern North America, *Seism. Res. Lett.,* 67, No. 5, 77-83.

Woolard, G.P., 1969. Tectonic Activity in North America as Indicated by Earthquakes, in *The Earth's Crust and Upper Mantle,* P.J. Hart, ed., American Geophys. Union, Monograph 13, 125-133.

Zoback, M.L. 1992, First- and Second-Order Patterns of Stress in the Lithosphere: The World Stress Map project: *Journal of Geophysical Research*, 97, pl 11, 703-11, 728.

Zoback, M.R. and M. Zoback. 1980. State of Stress in the Conterminous United States, *J. Geophys. Res.,* 85, 6113-6156.

Appendix A

A.1 Description of Datasets Collected and Archived for This Study

The following data sets were collected and archived for this study and are available at Weston Observatory:

Seismicity

USGS/NEIC Global Hypocenter Database, USGS CD-ROM, Version 3.0, 2100 B.C. - 1992. This database provides information on over 900,000 natural and artificial earth tremors from 2100 BC to 1995. Each event is listed by source, date, time, latitude, longitude, magnitude, depth, intensity and related phenomena. These data are compiled from 46 catalogs from worldwide sources. Consequently, this is one of the most, if not the most, comprehensive global seismic datasets available today. Earthquakes in the CEUS have been extracted from these files and converted to GIS point coverages.

USGS Extended PDE Data Catalog, USGS CD-ROM, January 1993 - June 1995. These data extend the USGS/NEIC Global Hypocenter Database, Version 3.0, to June 1995.

National Center for Earthquake Engineering Research (NCEER) seismicity data. This is an earthquake catalog for the central and eastern United States for the years 1627 through early 1985. We downloaded the following NCEER data from the Internet. Years: 1924-1984; Magnitudes: m≥3; 25°N≤lat.≤50°N; 110°W≤lon.≤65°W.

New England Seismicity. Seismicity data for the New England states are available as a GIS data set. The spatial accuracy of the data is different for different periods of time. The earthquake epicenters are known to an accuracy of 3-5 km for events since about 1975. The epicentral accuracy is about 10 km between about 1933 and 1975, and before 1933 it is 20 km or worse.

Regional Magnetic Field

National Geophysical Data Grids: Gamma-Ray, Gravity, Magnetic and Topographic Data for the Conterminous United States (USGS DDS-9, 1993). This dataset is extracted from the DNAG gridded magnetic data. These data are gridded at 2 km intervals with a resolution of 100 nanoteslas.

Decade of North American Geology (DNAG), Geophysics of North America CD-ROM, Magnetic data set. Units: Gammas; Grid Spacing (lat.): 2.5 min; Grid Spacing (lon.): 2.5 min; 25°N≤lat≤50°N; 110°W≤lon≤65°W.

Regional Gravity Field

Geophysical Data Grids: Gamma-Ray, Gravity, Magnetic and Topographic Data for the Conterminous United States (USGS DDS-9, 1993). These data are in ASCII format as non-gridded and gridded data designed for use on a PC. The included softwarehas a helpful DOS interface for extracting the desired data, but this interface fails to work properly on newer PCs.

Decade of North American Geology (DNAG), Geophysics of North America CD-ROM, Gravity data set. Units: mgals; Grid Spacing (lat.): 2.5 min; Grid Spacing (lon.): 2.5 min; 25°N≤lat≤50°N; 110°W≤lon≤65°W.

Regional Topography

National Geophysical Data Grids: Gamma-Ray, Gravity, Magnetic and Topographic Data for the Conterminous United States (USGS DDS-9, 1993). These data are in ASCII format as non-gridded and gridded data designed for use on a PC. The included software has a helpful DOS interface for extracting the desired data, but this interface fails to work properly on newer PCs.

Decade of North American Geology (DNAG), Geophysics of North America CD-ROM, Topography data set. Units: Meters; Grid Spacing (lat.): 5 min; Grid Spacing (lon.): 5 min; 25°≤lat≤50°; 110°≤lon≤65°W.

Regional Stress

Decade of North American Geology (DNAG), Geophysics of North America CD-ROM, Stress data set. This dataset is avaialable on the DNAG CD-ROM.

World Stress Map Data. This global dataset is described in Zoback (1992). The entire global stress database, as published in the *Journal of Geophysical Research* special issue on Stress in the Lithosphere (July, 1992), is available via anonymous FTP at andreas.wr.usgs.gov.

Regional Geology

Geology of the Conterminous United States at 1:2,500,000 Scale—A digital representation of the 1974 P.B. King and H.M. Beikman Map (USGS DDS-11, 1994). The regional geology is from "Geology of the Conterminous United States at 1:2,500,000 Scale"—a digital representation of the 1974 P.B. King and H.M. Beikman Map (USGS). While this is a good geology summary for the region as a whole, its resolution is insufficient for us to examine the

relationship between the geology and individual epicenters on a local scale. We explored getting bedrock geology from the individual states and met with very mixed success. Standard GIS, as incorporated in ARC/Info, is not well suited for handling geologic maps. It is not easy to incorporate formations, faults, folds, discontinuities, strikes, dips, metamorphic grade, etc., into the types of features (i.e., points, lines, or polygons) that ARC/Info supports. Furthermore, creating the links from the mapped features to the database is difficult. Thus, many states do not yet have their bedrock maps in a useful GIS form.

Regional Faults

Additional data on regional faults are available in the Weston Observatory archives.

Hydrography

1:100,000 Scale Digital Line Graph (DLG) Data Hydrography and Transportation (USGS Geo Data, 1993). These data are stored on a CD-ROM and include states in the Mississippi Valley, Northern Great Lakes and East Coast. The hydrography data are at a scale that is usable at state and regional levels. The transportation data contain roads, railroads and other transportation lines. Most states store all of their water-related features in a single GIS database without differentiating between major and minor water bodies. Defining the difference between a major and minor water body must be done with care.

Hydrography Data Set Provided by ESRI. These additional hydrography data were obtained from ESRI.

Political Boundaries

USGS data set of political boundaries, roads and hydrography for the conterminous US (stored on a CD-ROM). These data, accessible on the Internet at in http://edcftp.cr.usgs.gov/pub/data/DLG, provide 1:2,000,000 scale political boundaries, administrative boundaries, streams, water boundaries, hypsography, roads and trails, railroads, and other cultural features. These data were provided in the Digital Line Graph (DLG) standard, which we were not able to read immediately by either the Alpha-CARIS system nor by the PC-based ARCAD GIS. In addition, all GIS systems vary in their method of importing "standard" data, and several "standards" are available. This dataset, not only was a crucial supporting dataset for this project, but also provided a means of understanding and developing an efficient method of importing any future DLG data into Weston Observatory's GIS.

Political Boundary Data Set Provided by ESRI. These additional political boundary data were obtained from ESRI.

Gamma-Ray Data

The gamma-ray data acquired for this project are from the same data sets as the gravity, magnetic, and topography data (USGS DDS-6, 1996 and DNAG).

Stratigraphic Nomenclature

Stratigraphic Nomenclature Databases for the United States, its possessions and territories (USGS DDS-6, 1996). These data are stored on a CD-ROM.

A.2 Coordinate Systems

Storing GIS coverages in longitude and latitude coordinates is a logical standard in GIS. Thus, all data were to be first converted to longitude and latitude, and then any other conversions and data preparation would be done in that coordinate system. However, much of the data (i.e., gravity, magnetics, topography and seismicity) needs to be converted into GIS point coverages, then contoured. Accurate contouring must be done with data at regularly spaced intervals. Longitude and latitude, however, are not regularly spaced coordinate points. Longitude and latitude, however, are not regularly spaced coordinate points. These data must therefore be converted to a coordinate system that is as regularly spaced as possible prior to contouring.

The USGS data are provided in the Albers Equal Area (AEA) projection. In general, map projections are true to distance and shape near the center of the map, with increasing distortion toward the map edges. Selecting a map projection depends on how one wants to mange the distortions. AEA is a projection that attempts to represent map data in equal area sections that minimizes distortions of distance at the expense of shape distortions. As the USGS considers the contiguous United States as the area of interest for its data set, their projection is centered near the middle of the contiguous United States while our study area is centered in the middle of the eastern half of the contiguous United States. Nevertheless, we judged the distortions of the USGS AEA projection to be small enough that we could use it in our study.

Appendix B

The Role of GIS Technology in This Study

B.1 Introduction

The operational objective of this project was to develop a method of identifying seismically active features based on a broad range of geo-scientific data. These data are voluminous, often consist of disparate data types, and are usually geographically oriented (i.e., they can be located, drawn or identified on a map). GIS, a map-oriented database analysis tool, is a logical choice for this kind of project. In effect, GIS is simply a two-dimensional graphical front end of a database reporting system. GIS is designed to organize and report data using geographic maps supported by tables and charts. The strength of GIS is in its ability to spatially manipulate data in order to visualize and analyze spatial relationships among disparate data sets. Adding GIS to the seismic hazard analyst's toolbox provides a powerful complement to more "traditional" types of analysis. This project could be done entirely without the use of GIS, or entirely in the GIS environment, or in a combination of both GIS and non-GIS systems. We decided to use both GIS and non-GIS approaches for this project. Each approach has its advantages and disadvantages; the choice of which approach to use depends critically on the application in hand and on the expertise of the user.

Because GIS requires a well-defined set of queries, it is crucial to plan how GIS will be used prior to building the GIS project. The GIS press often quotes that about 60% of all GIS projects fail due to poor design. A key to the success of a GIS project is, therefore, to identify poor design early. Also, GIS capabilities are evolving at a fantastic rate—what is impossible or difficult today may be routine tomorrow. In our work, we frequently needed to change course as we learned more about the nature of the data we were using, as new technologies became available, and as our own experience grew. Continual change is a hallmark of GIS work—radical midcourse corrections are common and necessary.

B.2 Project Methodology

The various (and often disparate) geo-scientific data sets used for this study are fused using GIS technology. GIS provides a means to organize data that would otherwise be difficult to manage using standard analysis techniques (e.g., compute the distance from a point to the nearest fault). To accomplish this task, a system of hardware, software, and data management tools was developed for this project. Finally, these geographically organized datasets were exported in tabular format for statistical analysis.

Below we describe the methodology we used for organizing and analyzing the data for this project, as well as the history of this project. Our description begins with the issues involved in creating a GIS environment and then moves on to describing our analysis methods. This description is more or less chronological. One of the primary purposes of this appendix is to illustrate what worked, what proved difficult, and why we ended up with our final GIS datasets. Thus, the reader can take advantage of our successes and avoid the pitfalls we encountered.

Originally, our plan was to collect regional data for the CEUS and to supplement that with detailed data from the individual states. The intention was to use a two pronged approach: a coarse data analysis over the entire area of interest, and a fine data analysis over the area that we were most familiar with and where we had the most detailed data. As the project evolved, however, we discovered that we could manage the regional aspects of our original plan but including the detailed local analysis exceeded our resources.

In developing a computing intensive environment, which is required for GIS, it is crucial to carefully plan and balance the design and use of computing hardware, operating system, GIS software, training and data management. During the three project years, there were significant changes in the applicable hardware and software that was available on the market. GIS is rapidly evolving technology, requiring significant computing resources and training. The principal commercial suppliers have major financial investments in their complex software and have developed involved (and expensive) training seminars. These training seminars are a crucial part of successful GIS practice, but they also tie the user to specific GIS software packages. Also, most datasets are in unique data formats, which may or may not be easily readable by specific GIS software. The practical implications were that we had to alter course several times in our choice of hardware platform and software packages.

We found that using the highest levels of GIS software (e.g., Arc/Info) was inefficient for managing this project's requirements. This is because the high-level GIS programs are rather complex—thus the overhead of using this software is high. For this reason, we moved to a mid-level of GIS that was feasible for doing the work at hand (i.e., ArcView). Although there remain significant incompatibilities between mid-level and high-end GIS software, the mid-level software packages proved the most cost-effective for this project.

B.3 Project Description and Chronology

The three year period of this project breaks down into four phases. The following chronological description of this project illustrates how GIS technology can contribute to the task of seismic hazard analysis for the CEUS:

Phase 1. Work began on the project by defining the following three tasks for the initial phase of the project:

1. *Contacting the GIS centers of each of the New England states, the USGS, and private vendors to locate and acquire data.* We expected that the most time-consuming part of the project would be the assembly of the GIS database, and so began this activity immediately upon commencement of the project. For the regional data, we sought data from the USGS, who had much of the data that we needed, although much of it was not in GIS format. For the New England data, we began by acquiring data from the GIS centers of the six New England states. We found that each vendor had a different procedure for ordering data: some by phone, some by written request, and some only after a negotiation with the administrator of the GIS database. We also found that cost, availability, format and delivery time of the data sets varied widely among vendors. We began placing orders for data during this time period and made plans for ordering additional data. Data management was the key issue at this stage of the project.

2. *Specifying and purchasing the computer for the project.* Previous experience in dealing with GIS datasets drove home the need to specify and purchase an appropriate computer for the project as soon as possible. We needed a computer with large disk space and fast disk access and transfer rate, a fast CPU and large amounts of memory. Our choice was constrained by compatibility with existing GIS software. Hardware and software management became the key issues at this stage of the project.

3. *Begin converting data to GIS formats for the analysis.* We began the initial steps of constructing the GIS database with the first data sets that we received from the New England states. This work involved transferring the data into our initial GIS computer (a 486-based PC), decompressing it as necessary, and then attempting to read the data into a GIS program. The project began by using ARCCAD, a GIS add-on to AutoCad for data manipulation. ArcView (Version 1) was used for the initial viewing of the data. The most difficult and most important aspect of this phase was to assemble these datasets into a GIS database. For each dataset, several steps had to be taken to incorporate the data fully into the GIS. First, the data had to be transformed into a set of maps, all with a common map projection, registration and scale. Then the database information for each map had to be assembled. Finally, the GIS links from the map features to the database had to be constructed. Our goals were to display as much data as possible on a single map but at the same time to minimize the distortion of that map. The GIS software available at that time required selecting a baseline coordinate system. While a single map for the CEUS based on UTM coordinates seemed reasonable, we considered dividing up the study area into three or four separate regions and constructing GIS maps for each region separately. That would have allowed us to use different UTM projections for each area, minimizing the distortion for that area. Working with smaller GIS datasets dramatically improves system performance. However, multiple databases increases the complexity of data management. As desirable as advanced, goal-oriented planning is to GIS, practical considerations dictate that GIS planning necessarily is an iterative process between meeting scientific goals and optimizing technical resources. Again, data management played the key role at this stage of the project.

Phase 2. Initially, we selected Computer Aided Resource Information System (CARIS) as our GIS platform of choice. CARIS is the GIS designed by Universal Systems, Ltd. (USL) in Fredericton, New Brunswick, Canada. The selection was logical as CARIS is an excellent product, and it was explicitly developed for use in geo-scientific and geotechnical applications. At that time, USL's best developed version of CARIS was designed for a UNIX environment. CARIS for Windows NT was in the development stage. Although we preferred the simpler system management for Windows NT, the UNIX operating system was far more developed and stable. Consequently, we advanced to a UNIX-based workstation designed to run CARIS. During this phase of the project, we continued to import, convert and manage the required databases. Data and information management continued as the central issue in GIS design and execution. Importing the scientific databases, contouring them as required, and converting to GIS line coverages was the principal focus point.

Two problems surfaced that forced us to reconsider adopting Arc/Info to carry out this project. First, due to a competitive battle of proprietary data formats, CARIS did not acquire the capability to import Arc/Info formatted data. Second, CARIS requires an external database—it does not have even a rudimentary database that can keep track of more than one parameter. For a PC-based system, this is not a big problem, as PC-based databases are inexpensive and straightforward to use. However, databases for UNIX systems are expensive and complex. This became a difficult problem as the cost and effort of installing a UNIX-based database system exceeded our project resources. Arc/Info, which was not as well developed for geo-scientific use at the time, did have "INFO"—a built-in basic database. Consequently, we transferred our project entirely to Arc/Info. By this time, Arc/Info was available in Version 7, and the associated ArcView was available in Version 2. Both were significant improvements over the previous versions.

Arc/Info had all the tools we needed for this project, but its methodologies and command-line structure is complicated and non-intuitive. ArcView was an intuitive, point-and-click GIS viewer but did not have the data analysis and graphical tools that we needed. As we better understood the data that were available and how to use them in GIS environments, we began to divide the data management into three components: 1. Convert ASCII point data to GIS point data. 2. Convert GIS point data to GIS line data via contouring. 3. Convert GIS point data to GIS raster data via Arc gridding methods. When converting point data into two-dimensional representations, it was not clear to us whether contoured maps or raster "images" would prove more useful. As we could find only minimal information in the literature about other research in these areas, we needed to develop this area ourselves. At this stage of the project, we were trying to find the best techniques, constrained by the software capabilities. Since each approach generated its own large data sets, data management once again became a key issue.

Phase 3. With the release of ArcView 3, we found our capability significantly advanced. By this time it had become very obvious that there are two options in developing serious GIS projects: One option is to have an experienced GIS technician who is very experienced with the

line-oriented command approach used in high-end GIS software. The other option is to find a menu-driven GIS software that assumes a knowledgeable but less experienced GIS user. The former is really a full-time, highly trained GIS expert who may or may not also have experience in geophysics. However, the large number of commands, their switches and difficult to interpret error messages are a barrier to the researcher who needs the functionality of GIS, and needs to find how to *perform a specific function, not how to use a command*. Since, our project needed the latter, one of the crucial aspects of our successful use of GIS was in the availability of menu-driven systems. ArcView 3, with the addition of "Spatial Analyst", appeared to include *all* of the functionality that we would need for this project. Discussions with ESRI confirmed this, so we took a deliberate step back from our Arc/Info efforts with the expectation that ArcView 3 would solve our immediate data management problems. When ArcView 3 arrived and was finally functional, the work for this project progressed at a much faster rate than was previously possible. In short order, we read in ASCII data in almost any format and quickly converted all our data sets into GIS formats. With "Spatial Analyst", we easily gridded data, created contours that were accessible as GIS line coverages, and had easy access to various important GIS functions (e.g., calculating distance to nearest fault). Soon, however, we hit the limits of ArcView's menus. ArcView is built entirely on the programming language called "Avenue." Avenue is a "C-like", structured programming language that allows access to far more GIS functionality than ArcView's menus that were provided by the developer. All of the menu items included in ArcView are written in Avenue. ESRI (ArcView's developer) included what they believed would be the most popular functions in the default set of menus. For this project, we found reasonable workarounds for the necessary but missing functions.

Phase 4. At this stage of the project, we arrived at a point where we had the datasets in formats and with content that we could use for scientific analysis. The datasets produced were a combination of point, line, polygon, gridded raster, contoured gridded raster, and various combinations of all of these. The results of the analyses are reported in other sections of this report. For the GIS development portion of this project, this was the final stage of effort. To complete this phase, we set out to produce a comprehensive database that included all the major components of our project's efforts. The resulting database was exported as a tabular set of data. We developed two tabular data bases: (1) a set of geophysical parameters associated with the earthquake epicenters, and (2) a set of geophysical parameters associated with the centers of a set of 0.5° by 0.5° degree grid squares in our study area. These databases were then exported for use in the statistical analyses that are described in the body of this report.

B.4 Computer Hardware

Originally, we ordered an IBM PC, based on the Pentium 90 processor, and including 2 GB of Fast-Wide SCSI-2 disks, 500 MB of Enhanced IDE disk, 32 MB memory and 64-bit accelerated graphics. Unfortunately, the problems with the Pentium processor chips that surfaced at the beginning of 1995, plus major delivery problems by the vendor (IBM), delayed

our order indefinitely. Eventually, IBM informed us that the PC that we ordered was taken off the production line and would not be delivered. We then began to explore the possibility of purchasing another machine to handle the data from this project. We decided that a Digital Equipment Corp. (DEC) Alpha 400/233 computer with 64 MB memory and 3 GB disk space would be able to handle our needs. This decision was based on the ability of this particular hardware configuration to run either OSF/1 (Open Software Foundation UNIX, the then current version of DIGITAL UNIX), or Windows NT. The GIS software (CARIS and Arc/Info) were only available in UNIX versions at the time, but both vendors were developing Windows NT versions. The Digital Alpha workstation purchased for this project and located at Weston Observatory now shares its files with a Digital Alpha server located in the Department of Geology and Geophysics on the Boston College campus as well as with the Pentium-based PC and Macintoshes widely in use in Boston College. The work for this project involved all of those resources.

B.5 Preparing Data for Analysis

The volume of data that needed to be considered and managed in this project grew at a phenomenal rate. As we developed this project, and applied GIS technology to a difficult geophysics problem, we found that new data management techniques were required. Throughout the project, we noted that GIS datasets cannot be easily provided in a "generic" form that can be immediately applied to arbitrary set of scientific problems. In more traditional research, converting data formats for specific applications is a relatively minor component in the analysis process. Analysis using GIS requires intensive efforts in data formatting and management—the data *must* be specifically configured for the analysis efforts at hand. Furthermore, GIS and related datasets are usually available as "here are all the data available." Finding, or acquiring, limited datasets is difficult and often impossible. For example, we tried to build a regional hydrography GIS coverage from USGS hydrography data at a 1:100,000 scale. In order to do so, we would have to construct several thousand smaller coverages and join all of them into a single set. In this process, we would have to extract just the data we needed. Consequently, project-specific datasets must be culled from huge data sets. We deemed this an impossible task without virtually unlimited resources. Even when data are obtained from a single source, (e.g., USGS), they are often in various formats, designed for different computing systems and sometimes not in easily downloadable ASCII formats. These data then need to be converted into a form that can be imported into a GIS system. Frequently, software is available on the data CDs to decode and map the data. However, many of these CD-based programs do not work properly on new computers. For example, to speed up processing such programs may write directly to the video hardware—whereas "Windows-based" operating systems require that graphics information be written to the operating system, not to the hardware.

Throughout the project we focused on importing, converting and managing the databases required for this project. Some datasets converted easily, some stubbornly refuse to be reduced

to a useful size and content. A main issue for this project was figuring out how to convert the scientific point data (e.g., seismicity, gravity, magnetics and topography) into useful line or polygon GIS coverages. Fortunately, current versions of GIS software have significantly simplified this problem. Another significant issue involves the simplification of the hydrology, roads and political boundary datasets. These datasets are enormous in size and in number, and they proved to be very difficult to simplify. The roads and political boundaries are useful primarily for orienting the analyst. One solution to reduce the size of these datasets is to locate and convert other, non-GIS datasets for this purpose. With Arc/Info Version 7, ESRI provided several datasets that proved useful: major roads, rivers, lakes and drainage systems. As an alternative to the massive job of reducing the USGS hydrography data, we used the abbreviated ESRI dataset.

B.6 Conclusion

Primarily data management, and secondarily software and hardware management, are crucial to successful GIS projects. We cannot overemphasize the need to plan and execute such management carefully and to be able to adopt it to changing or unforseen circumstances during the course of a study. We strongly encourage any GIS project planner to allocate sufficient time and resources for this effort so that the project can progress smoothly and quickly—and produce visible and demonstrable results.

Appendix C

C.1 Variables Extracted for Data Analysis

From the datasets listed in Appendix A, we constructed ASCII tables of values to be used in the statistical analysis. The overall methodology was to create two datasets — one based on earthquake epicenters, and one based on regularly spaced (0.5° by 0.5°) grid points. In each case we built an initial dataset of values identifying and describing the events or grid cells. All related data were "joined" to these initial identidication values—the computed or linked parameters were generally from the nearest data point, line or area to the event or grid cell location. For example, the gravity value was taken from the nearest gravity measurement to the event or grid cell location. A river, fault or other linear feature was taken from the nearest one to the initial event or grid cell location value. For areal data, such as geology, we identified the geology of the mapped area that covered the base point.

Computed parameters, specifically gravity and magnetic gradients, needed special processing. First the point coverage was converted to a raster image. The raster image was then searched for equal intensity points, from which we produced contour lines. These contour lines were converted into GIS line coverages. Finally, the value of the nearest contour line was adopted as the required value associated with the event or grid cell point. In all cases, the name and distance to the nearest entity was maintained in the database.

C.2 List of first fields for the database based on seismic epicenters:

Recno - Record number
Lon - Longitude of seismic event
Lat - Latitude of seismic event
Year - Year of seismic event
Month - Month of seismic event
Day - Day of seismic event
Time - Time of seismic event
Depth - Depth of seismic event
Magnitude - Magnitude of seismic event
Type - Type of seismic event
Intensity - Mercalli intensity of seismic event
Felt_area - Felt area of seismic event
Distance - Distance to Fault
Fnode - GIS node attribute
Tnode - GIS node attribute
Lpoly - GIS node attribute
Length - Length of fault

Kbf_II_ - USGS fault ID number
Kbf_II_id - USGS fault ID number
Desc - Description of Fault
Ltype - USGS plotting line type for fault
Area - USGS geology area
Perimeter - USGS geology perimeter of area
Kbge_II_ - USGS geology ID number
Kbfe_II_id - USGS geology ID number
U - USGS geology geological unit number
Code - USGS geology code
Unit_lc - USGS geology rock type
Color - USGS geology mapping color
Dxf_text - USGS geology rock type (same as Unit_lc but upper case)
Distance2 - Distance to nearest stress measurement
Lon2 - Lon of nearest stress measurement
Lat2 - Lat of nearest stress measurement
Depth2 - Depth of nearest stress measurement
Type - Type of stress measurement
Azimuth - Stress azimuth
Quality - Quality of stress measurement
Regime - Stress regime
Location - Location of stress measurement
Distance - Distance to nearest topographic data point
Recno2 - Record number in GIS topography file
Lon2 - Longitude of nearest topographic data point
Lat2 - Latitude of nearest topographic data point
Topo - Topographic value
Distance2 - Distance to nearest topographic slope contour
Id - ID of topographic slope contour line
Contour - Topographic value of nearest topographic slope contour line
Distance3 - Distance to nearest gravity data point
Recno3 - Record number in GIS gravity file
Lon3 - Longitude of nearest gravity data point
Lat3 - Latitude of nearest gravity data point
Grav - Gravity value
Distance4 - Distance to nearest gravity slope contour
Id2 - ID of gravity slope contour line
Contour2 - Gravity value of nearest gravity slope contour line
Distance5 - Distance to nearest magnetic data point
Recno4 - Record number in GIS magnetic file
Lon4 - Longitude of nearest magnetic data point
Lat4 - Latitude of nearest magnetic data point

Mag - Magnetic value
Distance6 - Distance to nearest magnetic slope contour
Id3 - ID of magnetic slope contour line
Contour3 - Magnetic value of nearest magnetic slope contour line
Area - Area of Lake where seismic event occurred
Name - Name of Lake where seismic event occurred
Distance - Distance to nearest river
Name2 - Name of nearest river
System - Name of river system
Distance2 - Distance to nearest drainage system
System2 - Name of nearest drainage system
Distance3 - Distance to nearest major highway
Length - Length of highway
Type2 - Type of highway
Admin_class - Type of highway system
Toll_rd - Toll road?
Rte_num1 - Route number
Rte_num2 - Secondary route number
Route - Name of route
Distance4 - Distance to nearest city
City_fips - ID code of city
City_name - Name of city
State_fips - ID code of state
State_name - Name of state
State_city - ?
Type3 - Type of city
Capital - Is the city a state capital?
Elevation - Elevation of city
Distance5 - Distance to nearest nuclear plant
Plant - Name of nearest nuclear plant
Lon_d - Longitude in integer degree of nearest nuclear plant
Lon_m - Longitude in integer minutes of nearest nuclear plant
Lon_s - Longitude in integer seconds of nearest nuclear plant
Lon2 - Longitude in decimal degrees of nearest nuclear plant
Lat_d - Latitude in integer degree of nearest nuclear plant
Lat_m - Latitude in integer minutes of nearest nuclear plant
Lat_s - Latitude in integer seconds of nearest nuclear plant
Lat2 - Latitude in decimal degrees of nearest nuclear plant
Geology - Local geology at nearest nuclear plant
Parm1 - Parameter 1 of nearest nuclear plant
Parm2 - Parameter 2 of nearest nuclear plant
Parm3 - Parameter 3 of nearest nuclear plant

Parm4 - Parameter 4 of nearest nuclear plant
State - State where nearest nuclear plant is located

C.3 List of first fields for the database based on the 0.5°latitude by 0.5°longitude degree grid:

(Note: All distances are from center of grid)

Label - Indexing label A(-94°W) -> BB(-67°W); 1(48°N) -> 34(31°N)
X - Converted longitude (Label A -> 1) through (BB -> 54)
Y - Converted latitude (Label 1 -> 1) through (34 -> 34)
Lon - Converted longitude to center of grid "pixel"
Lat - Converted latitude to center of grid "pixel"
Count - Number of seismic events in grid "pixel"
Max_mag - Maximum magnitude of seismic events in grid "pixel"
Distance - Distance to Fault
Fnode - GIS node attribute
Tnode - GIS node attribute
Lpoly - GIS node attribute
Length - Lenght of fault
Kbf_II_ - USGS fault ID number
Kbf_II_id - USGS fault ID number
Desc - Description of Fault
Ltype - USGS plotting line type for fault
Area - USGS geology area
Perimeter - USGS geology perimeter of area
Kbge_II_ - USGS geology ID number
Kbfe_II_id - USGS geology ID number
U - USGS geology geological unit number
Code - USGS geology code
Unit_lc - USGS geology rock type
Color - USGS geology mapping color
Dxf_text - USGS geology rock type (same as Unit_lc but upper case)
Distance2 - Distance to nearest stress measurement
Lon2 - Lon of nearest stress measurement
Lat2 - Lat of nearest stress measurement
Depth2 - Depth of nearest stress measurement
Type - Type of stress measurement
Azimuth - Stress azimuth
Quality - Quality of stress measurement
Regime - Stress regime

Location - Location of stress measurement
Distance - Distance to nearest topographic data point
Recno2 - Record number in GIS topography file
Lon2 - Longitude of nearest topographic data point
Lat2 - Latitude of nearest topographic data point
Topo - Topographic value
Distance2 - Distance to nearest topographic slope contour
Id - ID of topographic slope contour line
Contour - Topographic value of nearest topographic slope contour line
Distance3 - Distance to nearest gravity data point
Recno3 - Record number in GIS gravity file
Lon3 - Longitude of nearest gravity data point
Lat3 - Latitude of nearest gravity data point
Grav - Gravity value
Distance4 - Distance to nearest gravity slope contour
Id2 - ID of gravity slope contour line
Contour2 - Gravity value of nearest gravity slope contour line
Distance5 - Distance to nearest magnetic data point
Recno4 - Record number in GIS magnetic file
Lon4 - Longitude of nearest magnetic data point
Lat4 - Latitude of nearest magnetic data point
Mag - Magnetic value
Distance6 - Distance to nearest magnetic slope contour
Id3 - ID of magnetic slope contour line
Contour3 - Magnetic value of nearest magnetic slope contour line
Area - Area of Lake where seismic event occurred
Name - Name of Lake where seismic event occurred
Distance - Distance to nearest river
Name2 - Name of nearest river
System - Name of river system
Distance2 - Distance to nearest drainage system
System2 - Name of nearest drainage system
Distance3 - Distance to nearest major highway
Length - Length of highway
Type2 - Type of highway
Admin_class - Type of highway system
Toll_rd - Toll road?
Rte_num1 - Route number
Rte_num2 - Secondary route number
Route - Name of route
Distance4 - Distance to nearest city
City_fips - ID code of city

Appendix C

City_name - Name of city
State_fips - ID code of state
State_name - Name of state
State_city - ?
Type3 - Type of city
Capital - Is the city a state capital?
Elevation - Elevation of city
Distance5 - Distance to nearest nuclear plant
Plant - Name of nearest nuclear plant
Lon_d - Longitude in integer degree of nearest nuclear plant
Lon_m - Longitude in integer minutes of nearest nuclear plant
Lon_s - Longitude in integer seconds of nearest nuclear plant
Lon2 - Longitude in decimal degrees of nearest nuclear plant
Lat_d - Latitude in integer degree of nearest nuclear plant
Lat_m - Latitude in integer minutes of nearest nuclear plant
Lat_s - Latitude in integer seconds of nearest nuclear plant
Lat2 - Latitude in decimal degrees of nearest nuclear plant
Geology - Local geology at nearest nuclear plant
Parm1 - Parameter 1 of nearest nuclear plant
Parm2 - Parameter 2 of nearest nuclear plant
Parm3 - Parameter 3 of nearest nuclear plant
Parm4 - Parameter 4 of nearest nuclear plant
State - State where nearest nuclear plant is located

NRC FORM 335 (2-89) NRCM 1102, 3201, 3202

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

(See instructions on the reverse)

1. REPORT NUMBER (Assigned by NRC, Add Vol., Supp., Rev., and Addendum Numbers, if any.)

NUREG/CR-6573

2. TITLE AND SUBTITLE

"Investigating Seismotectonics in the Eastern United States Using a Geographic Information System"

3. DATE REPORT PUBLISHED

MONTH	YEAR
February	1998

4. FIN OR GRANT NUMBER

W6370

5. AUTHOR(S)

J.E. Ebel, A.R. Lazarewicz, A. L. Kafka

6. TYPE OF REPORT

Technical

7. PERIOD COVERED *(Inclusive Dates)*

August 1994 - September 1997

8. PERFORMING ORGANIZATION - NAME AND ADDRESS *(If NRC, provide Division, Office or Region, U.S. Nuclear Regulatory Commission, and mailing address; if contractor, provide name and mailing address.)*

Boston College
Weston Observatory
381 Concord Rd.
Weston, MA 02193 -1340

9. SPONSORING ORGANIZATION - NAME AND ADDRESS *(If NRC, type "Same as above"; if contractor, provide NRC Division, Office or Region, U.S. Nuclear Regulatory Commission, and mailing address.)*

Division of Engineering Technology
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555-0001

10. SUPPLEMENTARY NOTES

E. Zurflueh, NRC Project Manager

11. ABSTRACT *(200 words or less)*

A Geographic Information System (GIS) database has been assembled to use in regional analyses looking for seismotectonically active features in the central and eastern U.S. (CEUS). Included in the database for the region are topography, earthquakes, stress measurements, gravity residual field, magnetic residual field, major rivers and regional geology, especially faults. Observables from this database were extracted for the seismically active areas of the northeastern, southeastern and central U.S. for use in multivariate statistical analyses. These analyses indicate that the earthquakes of the CEUS do tend to associate with faults and other deformation structures, but that the geologic characteristics are not very similar between earthquakes in different regions. The discriminant function analysis shows some ability to differentiate between seismic and nonseismic areas.

12. KEY WORDS/DESCRIPTORS *(List words or phrases that will assist researchers in locating the report.)*

GIS, Statistical Analyses, Seismicity, Geology, Database

Central and Eastern U.S., Earthquakes

13. AVAILABILITY STATEMENT

unlimited

14. SECURITY CLASSIFICATION

(This Page) unclassified

(This Report) unclassified

15. NUMBER OF PAGES

16. PRICE