Winter 1997

# Ecological database development and analyses of soil variability in northern New England

Michael Anayo Okoye

*University of New Hampshire, Durham*

# ECOLOGICAL DATABASE DEVELOPMENT AND ANALYSES OF SOIL VARIABILITY IN NORTHERN NEW ENGLAND

BY

**MICHAEL A. OKOYE**
B.S., University of Ibadan, Nigeria (1987)
M.S., Oklahoma State University, USA (1993)

**DISSERTATION**

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

**Doctor of Philosophy**

in

**Natural Resources**

**December, 1997**

UMI Number: 9819682



UMI
300 North Zeeb Road
Ann Arbor, MI 48103

This dissertation has been examined and approved.

Dissertation director, Russell G. Congalton
Associate Professor of Remote Sensing and GIS

John D. Aber
Professor of Forest Resources

Christine V. Evans
Associate Professor of Soil Science

Jeffrey H. Gove
Adjunct Assistant Professor of Biometrics

Fay A. Rubin
Coordinator, New Hampshire Geographically Referenced Analysis and Information Transfer Systems

November 15 1997
Date

## DEDICATION

To Kennie (Kenechukwu) and Nneka, my children and best friends, and to the many persons who helped me to make this dream come true.

## ACKNOWLEDGMENT

The stack of privation and poverty that characterized my family would be considered dismal even by African standard. I am still surprised at the unpredictable ways and inexplicable providence that have made it possible for me to climb to this academic pinnacle. I lack words to adequately thank God from Whom all blessings flow, and Who alone can "...raise the poor from dust and ashes and make him (or her) sit in high places with kings" (the Bible: I Samuel 2:8). To God, I give all the glory for the great work He has done.

I would like to thank my major advisor, Russell Congalton, for his advisement and especially for his prompt responses to my numerous requests for support letters in my arduous and prolonged search for means to fund this study. I very much thank my committee members: John Aber, Christine Evans, Jeffrey Gove, and Fay Rubin, for putting up with me, and also for their advice and support throughout this program. In addition, Dr. Aber's personal concern and intervention as the director of our Ph.D. program made it financially possible for me to continue my study beyond the first semester. Dr. Gove was a friend in many ways, and a mentor from whom I learned more than just science. My special thanks to Richard Hallett (now, Dr.) and Marie-Louis Smith who initiated the FIA project, and provided me the opportunity to have a dissertation project. Although these two were also students, they provided me valuable advice and support, and were very generous in sharing useful scholarly information and materials which they had.

One of the bright sides of my financial struggle at this university was the many great persons I had the privilege of meeting. These people provided me energy, encouragement, financial and other support that made it all possible for me to go through this program. Distinguished among these were Professor Tom Fairchild, a great friend in deed; Dr. Karol LaCroix and Dr. Harry Richard, deans of the Graduate School; Dr. Pat Bedkar and Mr. Emery Booska, the associate and assistant deans (respectively) of the College of Life Sciences and Agriculture. I would also want to thank Ms. Diana Wright---the ex-secretary to the NR Ph.D. Program, and my friends and personnel of the USDA Forest Experiment Station, Durham for their immense support. I am thankful to my many friends and relations who had, in one form or the other, stood for and with me during this program and other equally important life struggles. My most hearty thank you goes to my wife, Katie, and more especially to Kennie (Kenechukwu), my son and vice president, and Nneka, my daughter and deputy vice president, for the inspiration they have brought into my life.

Partial funding for developing the ecological database from FIA data was provided by the Northeastern Forest Experiment Station, Radnor, PA, through project fund # RWU 4154.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

**ECOLOGICAL DATABASE DEVELOPMENT AND ANALYSES OF SOIL VARIABILITY IN NORTHERN NEW ENGLAND**

**by**

**Michael A. Okoye**

**University of New Hampshire, December, 1997**

The 1983 Forest Inventory and Analysis (FIA) data of the states of Maine, New Hampshire and Vermont (the study area) contain large amounts of field-measurements of many ecologically important variables. Despite the vast potential usefulness of the FIA data for scientific research, the data were until now, literally unused except for a few administrative purposes, because of problems in the way the data were organized, summarized, and coded for storage. The primary objective of this research was to solve the problems that had thus precluded these FIA data from use in scientific applications, and present the data in a form that is readily accessible and usable for research. This objective was achieved by adapting the un-summarized data in a relational database management system (RDMS) and geographic information systems (GIS). RDMS-GIS technologies would make these data amenable to more types and multiple spatial scales of analyses than previously possible, thus providing the scientific community with an unusually large, high-quality, and spatially referenced data set.

The FIA data also contain field and laboratory measurements of soil properties made at the geo-referenced FIA plot locations. These soil data also provided the basis for other studies in this dissertation. These studies included analyzing the spatial variability of selected soil attributes in the study area; evaluating the nature of the differences in specific soil properties among the ecological land classification map (ECOMAP) section and subsection units; and assessing the variability of specific soil properties in the NRCS-State Soil Geographic Database (STATSGO) of the study area. Both the ECOMAP and the STATSGO studies involved the use of GIS techniques and multivariate statistical methods for map unit analyses.

This dissertation also included more theoretical investigations relating to applied statistics and soil science. One of these addressed the unanswered question of whether or not it is necessary to use non-linear transformations prior to computing variability statistics from non-normally distributed soil data, and explored the use of coefficient of variation as a semi quantitative index of nonnormality in soil variables. Another study looked at why and how error matrices and related statistics can be used as an effective, comprehensive quantitative method of evaluating soil classification and soil map quality.

# CHAPTER 1

# INTRODUCTION

## 1.1 Background and Perspective Of Studies

Many studies in the natural and environmental sciences critically depend on reliably measured spatial data. Such data are often not available especially for large areas because of prohibitively high costs. This dissertation consists of three major parts involving five separate studies. However, each study depends on, and uses the 1983 USDA-Forest Service Forest Inventory and Analysis (FIA) data for the states of Maine, New Hampshire and Vermont (the study area). FIA data consist of field-measured ecological data from over 4,000 geo-referenced locations (Figure 1). These data include about 100 important variables that comprehensively characterize the forest sites, soils, forest composition, land cover, etc. in these states. These data are potentially valuable for a number of scientific applications including ecological modeling and ecologically-based management of natural and environmental resources at regional scales. But despite the vast potential usefulness of these multi-million dollar data, FIA data are literally unused except for a few administrative purposes, because of problems in the way the data were organized, summarized and computer-coded for storage.

This dissertation research started primarily as an effort to solve these problems and make these FIA data available and usable for scientific applications. The other

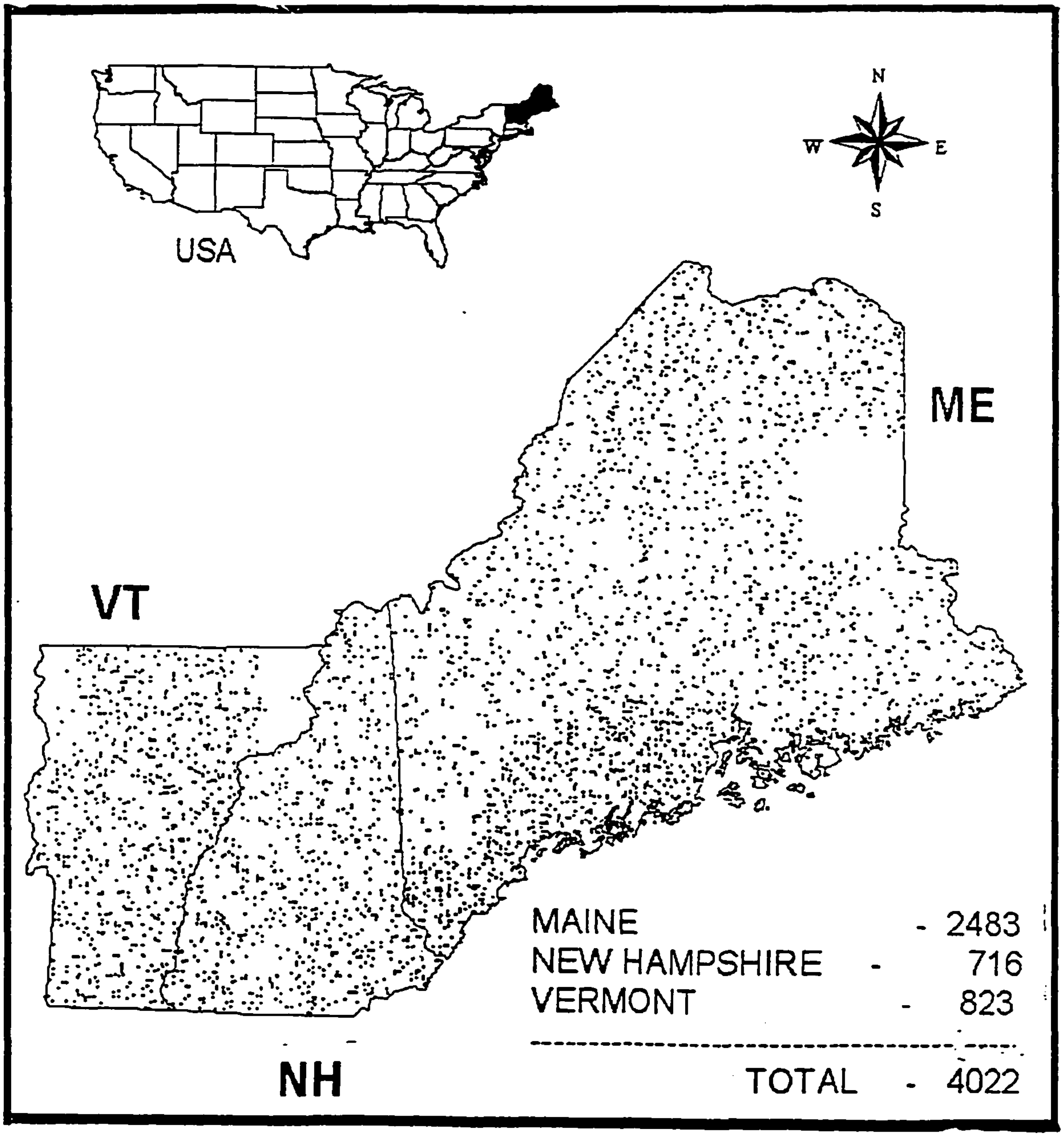



Figure 1-1 The Study Area and FIA Sample Plot Distribution

purposes of the dissertation research were to use the soil components of the FIA data to explore and describe the spatial variability of selected soil properties in the study area; evaluate the nature of the differences in specific soil properties among the ecological land classification map (ECOMAP) sectional and sub-sectional map units; and evaluate the reliability of the NRCS-State Soil Geographic Database (STATSGO) of the study area. The scope and goals of each of the dissertation studies are briefly described below.

## 1.2 Delimitation and Scope

### 1.2.1 Development of Northern New England Ecological Database from 1983 FIA.

Traditionally, FIA results are reported as aggregated summaries on county basis. With such summarization and the use of political boundary as scale, essential details and the spatial variability of the ecological variables were lost. The un-summarized data and especially the "non-forestry" information (e.g., soil chemistry data, geographic coordinates, etc.) were computer-coded in a format that discourages and/or prohibits their further use, and stored away from accessibility. Chapter 3, *Development of Northern New England Ecological Database from 1983 FIA,* describes how the un-summarized form of the ecologically important variables in the 1983 FIA data of Maine, New Hampshire and Vermont were reorganized in a relational database management system (RDMS), and structured in a fashion that allows the databases to be readily interfaced to both raster- and vector-based geographic information systems (GIS). The adaptation of the FIA data to RDMS-GIS technologies make these data amenable to diverse types and multiple spatial scales of analyses. The study also involved the development of comprehensive hardcopy and computer on-line documentation (i.e., user's guide) for the developed ecological databases. It is hoped that the database structure developed in the study will become a prototype and be adopted for subsequent FIA survey data. This will ensure consistency between survey projects and compatibility of inter-survey data, thereby permitting FIA data in the future to serve as a valuable tool for change analyses (change detection studies) of many of the dynamic ecological variables.

### 1.2.2 Variability Of Soil Properties In Northern New England Based On FIA Data.

The 1983 FIA data also contain field and laboratory measurements of some physical and chemical soil properties, made from the B-horizons of soil profiles at the geo-referenced plot locations. These soil data include variables that are known to affect land use and are important for environmental studies and resource management. In Chapter 4, *Variability Of Soil Properties In Northern New England Based On FIA Data*, these FIA data were used to explore and describe the spatial variability of some of these soil properties in the study region. The spatial scale chosen for this study was the section ecological map units of the National Hierarchical Framework of Ecological Units (ECOMAP, 1993), recently adopted by the USDA-Forest Service (Avers et al., 1994). ECOMAP is a geographically-based ecological regionalization, classification and mapping system for stratifying the earth into progressively smaller areas of increasingly uniform ecological potential. ECOMAP is based on multiple biotic and environmental factors which include climate, physiography, geology, soils, water, and potential natural communities. Based on available national, regional and state resource maps and information, and through the participation of numerous individuals from federal and state agencies and non-governmental organizations, an ecological map and characterization data of the eastern United States have been produced (Keys et al., 1995). These maps present 10 section and 30 subsection ecological units for the Northern New England or study region (see Figure 4-1).

The primary goal in this study was to provide summary statistics of important soil attributes within each ECOMAP Section in the study area. These statistics include central tendency and variance statistics (including coefficients of variation), estimate of confidence intervals of means, and the number of observation (sample sizes) required to estimate the population means of soil properties at different levels of precision. This type of information is important for soil-based resource and land use management, and is needed in much pedological and environmental research requiring field sampling on regional scales. Also, although ECOMAP subsections are expected to reflect differences in "soil types", no studies have been done to empirically assess the nature of these differences and/or to express how these map units differ in terms of specific soil properties. A secondary objective in this study was to evaluate the differences among subsection ecological map units with respect to specific soil attributes. This objective is analogous to evaluating the potential suitability of ECOMAP subsections as a basis for partitioning and describing field variation of soil properties and on a regional scale, and for extrapolating soil attribute information from place to place in the study area.

### 1.2.3 The Legitimacy Of Variability Statistics Computed From Non-normal Soil Data

Statistical analyses and exploration of the large data and several soil variables carried out in Chapter 4 provided a unique opportunity to re-evaluate some the conclusions frequently made about the distributions of soil variables. More importantly, the course of the study brought my attention to a major unanswered question and controversy in soil variability studies. One of the consistent conclusions is that natural soil populations are rarely normal or symmetrically bell-shaped about the mean, but are mostly positively skewed, often in a lognormal fashion. However, much confusion still exists about whether or not to use transformations prior to the computation of traditional variability statistics, namely, the mean, standard deviation, coefficient of variation, confidence intervals of the mean, and optimum sample sizes for the estimation of the mean. The soil science literature reveals conflicting recommendations, and the publication of studies advocating contradicting approaches to this problem. This chapter of the dissertation was developed posteriorly to the study in Chapter 4, to address this apparent contradiction in the soil science literature, among other objectives.

Through the review of pertinent statistical and soil science literature, and extensive statistical analyses of real soil data sets, this study showed that it is desirable but not necessary to achieve normality in soil data to validly compute traditional variability statistics. It discusses the limitations of the use of non-linear transformations and why it is not advisable to employ them prior to the computation of variability statistics on soil data. The study also showed that valid and more appropriate interpretation of variability statistics would require certain types of information about the nature and degree of non-normality in the untransformed soil data set, but that the coefficient of variation (CV) and other distribution characteristics can be used to provide such information. It shows why the CV is a better index of non-normality in soils than both the qualitative and commonly used quantitative tests of normality. This study provides practical guide on how to use the CV and other information to more validly interpret variability statistics computed from non-normally distributed and untransformed soil data.

### 1.2.4 Multivariate Analysis of Map Unit Variability In NRCS-STATSGO: A Case Study in Northern New England.

Soil survey has traditionally been the most practical method for partitioning field variation or grouping similar and separating different soils on a regional scale (Trangmar et al., 1985). However, within the last two decades, concerns about the reliability of soil survey or accuracy of soil map information have gained increased importance among scientists and users of soil surveys and land evaluation data. The literature is replete with documentation of the causes of these concerns. For instance, Moore et al. (1993) stated that conventional soil maps neither delineate all of a field's inherent variability nor represent specific soil attributes; and the inferred homogeneities do not exist for many physical and chemical attributes that affect environmental modeling and soil-specific management. These and other problems of soil survey have created the need to quantitatively evaluate soil map quality, and further characterize the variability within soil map units.

The USDA-NRCS soil survey data are presently being automated or computerized nation-wide, and made available as one of three types of digital geographic databases, reflecting different levels of details (SCS, 1991). From the most to the least detailed, the digital soil databases are the Soil Survey Geographic Database (SSURGO), the State Soil Geographic Database (STATSGO), and the National Soil Geographic Database (NATSGO). STATSGO is compiled at a scale of about 1:250, 000 and is designed to be used "primarily for regional, multi-state, river basin, state, and multi-county resources planning, management and monitoring" (SCS, 1991, p. 2). Data for STATSGO are distributed as complete coverage for a state, and are available for most states of the US. Digital soil databases like STATSGO, and GIS technology have introduced new users to, and expanded the functions of soil survey information, and they greatly facilitate operations of familiar soil-based analyses especially on a regional scale. However, these digital soil databases are subject to all the potential errors of soil survey as well as other errors which are introduced in the further process of digitization or automation.

The goal of this study was to assess the "reliability" of NRCS-STATSGO data and elucidate the nature of the variability of specific soil properties in STATSGO map units in the study region. Among other analyses, multivariate statistical methods (i.e., multivariate analysis of variance and discriminant function analysis) were used to ascertain how STATSGO map units differ on the basis of specific soil attributes, and to assess the relative efficiency with which specific soil properties were mapped in STATSGO. Results of this study provide scientists and others who must use the readily available NRCS STATSGO data some ideas of when and for what soil properties the data are adequate, and the degree of variation in soil properties to expect within a given map unit and between related map units. Also, map unit variability studies like this are important in order for us to better understand soil genesis and improve soil survey methodology. Assessing map unit variability in STATSGO provides some evaluation of the efficiency of the traditional *compilation* methodology for making small-scale, large-area soil surveys in similar sites.

### 1.2.5 The Use Of Error Matrix In Evaluating Classification Accuracy And Soil Map Quality

The reliability of soil survey or accuracy of soil maps has become a critical issue to many users of soil survey information. Published research in soil survey and land evaluation has continued to reveal the need to find an effective quantitative method of evaluating and expressing the reliability of soil classification and soil map quality. At the same time, the remote sensing community has made significant advancement within the last two decades in the area of accuracy assessment of classification through the use of error matrix and discrete multivariate statistical analyses. The error matrix and related statistics are state-of-art, quantitative techniques that provide comprehensive information about the accuracy of maps or classifications from remotely sensed data. The thrusts of this part of my study are that 1) the art and science of classification in remote sensing are markedly similar to those of soil classification and mapping, and 2) therefore, the use of the error matrix techniques could also be adapted in soils to significantly improve the present methods of assessing soil map quality.

The objective of this study was to introduce the use of error matrix and related statistics in evaluating soil classification and soil map quality. The study discusses pertinent error matrix concepts, and demonstrate their applications in soils through the analysis of real data from STATSGO classification. It reviews the present methods of evaluating soil map quality, and shows why the use of the error matrix techniques could be a solution to the age-long search for an effective, comprehensive and quantitative method for evaluating and communicating soil map quality.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 USDA-Forest Inventory and Analysis (FIA) Data

### 2.1.1 Background History.

Forest Inventory and Analysis (FIA) is a continuing endeavor mandated by Congress in the Forest and Range Renewable Resources Planning Act of 1974 and the McSweeney-McNary Forest Research Act of 1928. Its objective is to periodically determine the extent, conditions, and volume of timber, growth, and depletion of the Nation's forest land (Hansen et al., 1992). Initial inventory efforts began in the West in 1930, and by the 1960s, inventories were completed for all of the 48 conterminous states, and more than once for many of the more heavily forested states (Birdsey & Schreuder, 1992). These initial inventories were conducted on state-by-state basis, and were concentrated essentially on providing volume data on the timber resources of most states and regions.

Between the 1960s and 1970s, significant changes and rapid expansion in natural resource inventory were introduced. The 1974 Resources Planning Act emphasized the need for FIA to provide information about the various resources occurring on forest and range lands, i.e., forage, timber, water, wildlife habitat, recreation (Birdsey & Schreuder, 1992). Today, FIA procedures are standardized, and data are collected and published by each of the USDA Forest Service regional experiment stations for a number of specific states. Statistics from each experiment station are presented in manner that permits aggregation with those from the other stations in order that uniform regional and national statistics may be produced.

The Northeastern Experiment Station at Radnor, Pennsylvania is responsible for the FIA of 14 northeastern states including Maine, New Hampshire and Vermont. In this region, inventories are usually conducted every 5 to 15 years (Hansen et al., 1992). For the states of Maine, New Hampshire and Vermont (the study area), the last survey was completed in 1983 and the next survey which started in 1995 is being completed. The 1983 survey was the fourth inventory conducted by the Northeastern Forest Experiment Station for New Hampshire and Vermont (USDA-FS, 1982), and the third for Maine (USDA-FS, 1981). The inventory data were collected at over 4,000 geo-referenced plot locations (Figure 1), and include about 100 measured variables (see Tables 1(a-e), Chapter 3) on soil, geology, land-use, forestry and related resources, in addition to tree-level forest composition data.

### 2.1.2 Present Problems.

The purpose of FIA surveys is to gather data for use in management planning and policy making, and to provide expert advice and assistance in solving resource questions (Hansen et. al, 1992; USDA-FS, 1992). FIA data have been used primarily for the evaluation of forestry resources and tracking of merchantable timber volumes by county. To serve these administrative purposes, FIA survey findings are traditionally summarized and reported on a county basis. With such summarization and the use of political boundary scale, essential details and the spatial variability of the forestry variables are lost. The un-summarized data, most "non-forestry" data and all locational information were computer-coded in a format that discourages and/or prohibits their further use, and then stored away from accessibility. The limited access to the un-summarized data is partly to ensure that the public is kept from visiting the plot locations. There is an agreement between the Forest Service and the land owners that specific information on plot locations will be kept confidential. There is on-going discussion about this agreement and how it affects the attempts to use FIA for scientific research. For now, potential investigators that wish to use the FIA data for research have to be considered on a case by case basis. Access and permission to use the data are granted but with conditions and restrictions about visiting the plot locations

Although FIA has been going on for 60 years and some areas of the USA have been surveyed six times, many significant changes have occurred from one survey to another (Birdsey & Schreuder, 1992). Hansen et al. (1992) affirmed that inconsistency in data collection and processing methods creates data incompatibility among FIA projects and precludes analysis of data from more than one FIA project. There is, therefore, the need to make efforts towards the development of a uniform data collection method and sampling plan between FIA surveys. Consistency between survey projects would permit FIA data in the future to serve as a valuable tool for change analysis (change detection studies) of many of its dynamic ecological variables.

### 2.1.3 Quality and Potential Scientific Uses

Being field-measured is one the unique qualities of FIA data. Hansen et al. (1992, p.3) describes the high accuracy standards with which the USDA Forest Service carries out forest inventory plans. FIA inventories are said to be designed to meet the specified sampling errors of 67-percent confidence limit (one standard error) at the state level. A 3-percent error per 1 million acres of timberland is the maximum allowable sampling error for an area. A 5-percent error per 1 billion cubic feet of growing stock on timberland is the sampling error goal for volume, removal, and annual growth (Hansen et al., 1992). There are strong reasons to think that FIA tree-level (species types, volume and conditions), vegetative and other landcover and landuse records are reasonably accurate. FIA surveys were conducted by Forest Service personnel to whom such inventory must have be familiar routine. Pertinent Forest Service publications (e.g., USDA-FS 1980 & 1982) show high level preparation and training of the Forest Service personnel prior to FIA surveys.

Field measured or field-verified regional data are scarce for most natural and environmental resources. The paucity of, and critical need for reliable and extensive data sets for validating ecological models at the regional scales are documented by Aber et al. (1993). Remote sensing and field extrapolation techniques have been the traditional means of obtaining regional-scaled data. One of the advantages of remote sensing is that it gives a complete census of the object of interest. However, research (e.g., in soils and forestry) has shown that remotely sensed data often do not have the same level of reliability as field measurement. In fact, field measured data are usually required to assess the accuracy or reliability of data from remote sensing and extrapolative survey procedures. Congalton (1988; 1991) showed that an integral cost in most remote sensing studies is for the acquisition of field data as "training sites" and for accuracy assessment of classification results. Being actual field measurements, FIA data present a rare source of important data for many scientific studies at the regional scale.

The regional coverage, intensive field sampling and comprehensive nature of FIA data make them usable for regional ecologically-based resource management and analysis. They are also potentially valuable for ecological model validation, and for cross-validation of traditional sources of environmental and natural resource data. As a reliable source of actual field measurements, FIA may have the potential of being used as "ground truth" or as reference data for calibration of airborne remote sensors and for classification and accuracy assessment of remotely sensed data at a regional scale. If used in these ways, FIA data will reduce the cost and effort required to acquire data for remote sensing studies. The idea of making the un-summarized FIA data available and useable for scientific applications was presented at the Second International Conference/ Workshop on Integrating Geographic Information Systems and Environmental Modeling (Smith & Hallett, 1993). The response and interests generated at this conference convinced us that there will be many new users of these data if the un-summarized data are presented at the plot level with corresponding geo-referenced coordinates for each plot.

## 2.2 Relational Database Management and Geographic Information Systems

Adaptation of FIA un-summarized data to a relational database management system (RDMS) and geographic information systems (GIS) technologies will allow diverse types of data analyses and multiple spatial scales to be applied to these data. RDMS and GIS are the most up-to-date and efficient computer-based systems for organizing, storing, managing, manipulating, analyzing and presenting geographic information.

Geographic information is data about objects and phenomena where space or locational position is an important characteristic or is critical to the analysis (Aronoff, 1993). Geographic information typically has two components: the spatial features which show the dimension (e.g., area), shape and location in space, and descriptive or attribute data associated with the spatial features. Attribute data are usually organized in a tabular form as independent tables of related information. RDMS is the data model "most widely accepted" for handling non-spatial attribute data in GIS applications (Aronoff, 1993), and most GIS are built to readily accept geographic data from standard RDMS. RDMS allows one to define relationships between different tables, extract or combine data from these tables, and to use Boolean logic and mathematical operations to formulate queries in an unlimited ways (Aronoff, 1993; Burrough, 1987). RDMS is also used to display and present query results in a variety of ways (Borland, 1994).

One of the most important benefits of a GIS is its spatial analysis function; the ability to organize and integrate large volumes and multiple types of spatial information from a range of sources (Lillesand & Kiefer, 1994), and analyze these to show expected or previously unidentified relationships within and among these data sets. GIS-RDMS interfaces are used when there is need to manipulate and analyze data in both a spatial and a tabular sense thereby providing the scientist with a richer data model than the traditional tabular data structures alone (Lanfear, 1989), and also allowing query results to be cartographically displayed and spatially visualized.

## 2.3 Soil Spatial Variability

### 2.3.1 Importance of Soil Variability Information.

Humankind depends on soils for a multitude of agricultural and non-agricultural uses. Many of these land uses are known to be discriminatory on properties of soils. Weismiller et al. (1977) stated that soil information is the bedrock of any sound decision on land use planning, optimization of agricultural production, and conservation and management of many natural/environmental resources. According to Lillesand & Kiefer (1993), soil information forms a primary source of resource data about an area. Way (1985) states that when development activities are undertaken, land planners must be concerned with and understand the properties of soils if planned land use are to be in harmony with the environment.

Spatial variability is change in a given variable over distance. Soil scientists have recognized variation in soil from place to place for many years (Webster, 1985; Arnold & Wilding, 1991), and much effort has been devoted to understanding and describing this phenomenon. The need to understand, describe, document or report soil spatial variability is well documented (Wilding & Drees, 1983; Wilding, 1984; Arnold & Wilding, 1991). Knowledge of soil variability is essential to properly monitor and understand much of long-term ecological research data (Nash & Daugherty, 1990), and soil maps have become valuable tools for natural resource management (Moore et al., 1993). Values associated with soils and their combination in space are vital for tax assessment, land values, route locations, preservation of areas deemed important for society such as fragile land, wilderness, prime farm land, and wetlands, and the identification, inventory, and evaluation processes underlying policy decisions concerning land uses (Arnold & Wilding, 1991). Soils are also sinks, sources, and filtering membranes, as well as blocks of memory (Arnold & Wilding, 1991), and play vital roles in mitigating the effects of natural and anthropogenic perturbations of ecosystems (Lammers & Johnson, 1991). Understanding of soil variability is very important in studies to predict tree growth and timber production from forest site attributes (Blyth & Macleod, 1978). Grigal et al. (1991) stated that variability in soil properties is a particularly vexing problem for those attempting to assess either the present status or changes in ecosystems. They added that this variability can affect both precision of estimates and the ability to detect true underlying relationships. In a study he entitled: Soil Variability---A Serious Problem in Soil-Site Studies in the Northeast, Mader (1963) stated that "the degree of variability in forest soil and limits of accuracy of mean plot values for soil variables is an important problem needing evaluation for soil-site studies in the Northeast". Wilding and Drees (1983, p. 84) and Boul et al. (1989, p. 358) list major reasons why pedologists continue to pursue soil spatial variability. These reasons include the following:

(1) To estimate central tendency and variance statistics for specific soil classes and class differentiae

(2) To quantify soil genesis studies, including both the effects of pedogenic process and of external soil-forming factors

(3) To more quantitatively determine the composition of soil mapping units

(4) To develop better sampling designs and statistical models for soil survey and pedogenic applications

(5) To determine optimum allocation of sampling units for the most efficient statistical design

(6) To differentiate between systematic variations (such as change in one or more soil-forming factors) and random variation (associated with sample selection, collection and laboratory analyses)

(7) To improve efficiency and quality of soil surveys

(8) To determine spatial variability in three dimensions so that soil formation and soil behavior can be easily visualized

(9) To determine more precise and quantitative information about land tracts that can be applied by land users (both agricultural and nonfarm) to improve decision making.

### 2.3.2 Random versus Systematic Soil Variation.

Spatial variations of soil properties are categorized into two components: systematic versus random variations. Random variations are observed differences in soil properties which cannot be readily attributed to a known cause, and thus cannot be explained. Systematic variability is a gradual or marked change (or sign of trend effects) in soil properties as a function of landforms (e.g. mountains, basins, plains, terraces, valleys moraines, etc.) and Jenny's (1941) soil-forming factors. Soil forming factors are climate, geologic parent material, topography, biota (especially vegetation and soil management by man) and time or age of soil in the landscape (van Wambeke & Dudal, 1978; Wilding & Drees, 1983). Where these factors are similar, similar soils are formed, and the cumulative and differing effects of these factors on soil formation are expressed as observable properties (Hartung et al, 1991).

Systematic variability implies that soils with discrete sets of properties have a degree of predictability on the landscape (Miller et al., 1979; Soil Survey Staff, 1980a, b; Witty & Arnold, 1987); it is the scientific basis of soil survey and mapping. Hudson (1980) stated that "Soil mapping is possible because of observable discontinuities between landscape units, and the strong covariance between landscape units and soils. These relationships make it possible to accurately delineate bodies of soil with limited observations". Soil properties have observable relationships with soil forming factors (Jenny, 1941), and a degree of predictability on the landscape (Soil Survey Staff, 1980a, b). Visible changes in slope, vegetation, surface color, and drainage pattern enable a soil scientist to locally extrapolate soil/landscape relationships previously established (Wilding, 1984).

To summarize, random variability occurs simultaneously and concurrently with systematic variability. Systematic variability is explained heterogeneity while random variability includes what is left, and the relative proportions of the systematic and random components of variation will be inversely related and dependent on our present level of knowledge and the nature and scale of investigation. Wilding & Drees (1983) state that when the soil system is investigated in greater detail, a part of the variation originally considered random may be recognized as systematic, and if our state of knowledge were perfect, perhaps all variation in soil properties would be recognized as systematic.

### 2.3.3 Soil Classification, Soil Survey and Soil Maps.

Soil classification and soil survey have been the most practical methods for investigating field systematic variation or grouping similar and separating different soils on a regional scale (see Trangmar et al., 1985). Soil survey is a technique for determining soil resources and describing their spatial distribution on the landscape. During soil survey, the land surface is divided into parcels. Within each parcel, the land is considered to be of the same kind or of a few kinds of soils that can be listed and described. Usually, soil surveys are made using a particular soil classification system which guides the naming of delineated areas and the placement or location of boundaries that are not readily visible by external features (Arnold, 1983). Soil classification is the systematic arrangement of soils into groups or categories on the basis of their characteristics (SSSA, 1987). Soil classification is used to help soil scientists predict the behavior of one kind of soil for which experimental data are lacking, by its relationship to the other kinds of soils for which knowledge and experience exist (Van Wambeke, 1982). A conceptual group of soils having defined or specific ranges in particular soil properties constitutes a soil class or taxonomic unit (TU). A soil taxonomic unit could be described as a well-defined, highly structured sets of taxonomic criteria (Markewich & Cooper, 1991), or a defined portion of a multi-dimensional array of sets of soil properties that are known from studying pedons or other sampling units of the landscape (Arnold, 1983).

The results of soil survey are usually portrayed as a soil map. Choroplethic maps are the most common kind. Parcels or geographic delineations similar in nature are grouped into classes called map units (MU). Names for the map units which also constitute the legend for the map are chosen from the TU that best describes the typical or modal soil profile apparently found in the map unit (Webster, 1979). In summary, soil survey identifies bodies of soils that can be recognized as natural units, predicts and delineates their areas on maps, and identifies the delineated areas in terms of defined kinds of soils or conceptual soil classes called taxonomic units. Hence the ranges in soil properties assigned to MUs are based on typical values of the TUs that are supposedly predominant within the landscape delineations. Map units of the USDA-Natural

Resources Conservation Service (NRCS) soil surveys are described with Soil Taxonomy (Soil Survey Staff, 1975), the US system of soil classification.

Soil spatial variability is anisotropic (multi-directional) in nature and occurs in a continuum that ranges from sub-microscopic to megascopic in scale (Wilding, 1984; Upchurch et al., 1988). It is impossible to observe or sample the soil at every point on the landscape. Therefore, the soil scientists are restricted by resource and other pragmatic constraints to actually observe or sample a limited number of spots during soil survey of an area. From the knowledge of soils in these places, they interpolate or predict the properties of soils in other unsampled locations. Wilding (1984) argued that the soil scientist needs only enough observations to determine soil/landscape relationships and to confirm predictions of soil models established from these relations. The predictive approach of soil survey has been praised for substantially reducing the amount of money, time and effort required for physically visiting and sampling many spots (see Hartung et al., 1991; Bie & Beckett, 1971) while still producing information reliable enough for many uses (Webster, 1985; Hudson, 1980 and 1990; Hartung et al., 1991). Soil survey technology has thrived because the "procedure has undoubtedly been successful" (Webster, 1985), and "practical experience has convinced us that soil maps are reliable and provide valid interpretations" (Hudson, 1990).

## 2.4 Limitations of Soil Survey and The Need To Further Characterize Map Unit Variability

The possibility of reliable soil resource information at a reasonable cost is certainly the greatest merit of soil survey methodology, and the prime reason for the continued and predominant use of this approach especially for large area studies. However, this approach has limitations which are of significant concerns to many users of soil survey results. The literature is replete with documentation of the shortcomings of traditional soil surveys (Butler, 1980; Holmgren, 1988; Nash & Daugherty, 1990; Netttleton et al., 1991; Moore et al, 1993). First, the reliability of the predictions obtained from soil survey varies widely as a function of the soil scientists' experience, knowledge and abilities. It also depends on the complexity of or abruptness of change in the mapping area (Hartung et al., 1991), and the degree of correlation among different soil properties and their relations in the landscape (Webster, 1985). Rogowski & Wolf (1994) considered, the assignment of properties derived from "typical" or modal soil profiles to the entire map unit without regard for the inherent spatial and temporal variability of field soils, as the most serious limitation of the current survey process. Moore et al. (1993) reported that the inferred homogeneity of soil maps does not exist for many soil physical and chemical attributes, and ranges given for some attributes often vary by an order of magnitude (see also Wilding, 1984). They attributed this problem to the fact that the nearest sampled pedon or soil used to derive mapping unit attributes could be kilometers from point of interest.

Other concerns of surveys include uncertainty regarding the placement of soil boundaries, presence of inclusions [mixing of soils that are taxonomically and interpretively dissimilar within the MUs] and lack of a mechanism to quantify spatial variability within map units (Rogowski & Wolf, 1994). Webster (1979) stated that soil maps do not show soil data but merely serve as indices to data; they show the limits, as soil boundaries, within which data can be safely used for prediction. As a result of such variation within sampling units, soil survey cannot be expected to reliably predict variation of all properties, particularly those that are easily influenced by soil management (Arnold, 1983; Trangmar et al., 1985). Moore et al. (1993) added that the approach lacks quantitative framework and does not delineate all of a field's inherent variability nor represent specific soil attribute variability. Webster (1985) explained that although the soil survey procedure has undoubtedly been successful, nowadays scientists increasingly require quantitative estimates of soil properties for regions. They need confidence limits, probabilities, and frequency analyses on the composition of map units and information on how the inclusions within a given map unit influence the interpretation (Miller, 1978; Brubaker & Hallmark, 1991). They want to know the probability that knowledge about variability can be extrapolated from one mapping unit to the next (Wilding, 1988). These types of information are rarely included in traditional soil surveys, but they can be determined from more intensive sampling, and field and laboratory measurements of specific soil attribute, even after the survey had been completed.

For the purposes of this dissertation, the limitations of soil survey (discussed above) can be categorized rather arbitrarily into two groups. First are problems that are naturally concomitant of the predictive nature of soil survey methodology, and are direct tradeoffs of the advantage of economy in field sampling and laboratory data analysis.

These "problems" include functions that the conventional soil survey methodology is simply not capable or appropriate to perform. Most soil classes are polythetic, depending on values of multiple soil attributes. It is practically impossible to efficiently separate the variability of all these soil attributes in a field of any reasonable size by soil survey method. Arnold (1983) noted that the art and scale used in map making, and the recognition of intermingled soil bodies having contrasting qualities preclude delineating areas containing the same limits of variability as taxonomic classes. Soil survey cannot be expected to reliably predict variation of all properties, particularly those that are easily influenced by soil management (Arnold, 1983; Trangmar et al., 1985). The main thrust from these problems is the obvious need to augment soil survey (i.e., interpolated) data with more empirical or observed and quantitative estimates of specific soil attributes. Lammers & Johnson (1991) explained the need for an alternate strategy [to soil survey] that captures local-scale soil variability and provides a mechanism for maintaining integrity across scales of extrapolation. Quantitative, precise and multi-scale analysis of spatial variability of individual soil properties requires actual measurements of the soil properties of interest at reasonably intensive scale. Such measured data are sparse (Burgess & Webster, 1980) and practically non-existent for large areas or regions.

The second type of problems indicate the necessity to improve the traditional soil survey information to accommodate new and more sophisticated pedocentric needs. Concerns about the reliability of soil survey or accuracy of soil maps became much more important within the past two decades. Since then there has been a steady proliferation of research studies and published literature on soil spatial variability. This trend appears to parallel our increasing concerns about the environment, and the increasing number of soil-based studies of global change, environmental quality and ecological management. Lammers & Johnson (1991) observed that scientists from many disciplines are recognizing the vital role that soils play in mitigating the effects of natural and anthropogenic perturbations of ecosystems. It could be argued therefore that many of the inadequacies of soil survey (mentioned above) have "evolved" essentially from recent changes in land use, and from the "paradigm shift" in soil geographic research from the traditional predominant focus on agricultural production. As Nordt et al. (1991) put it "land use today is frequently more intensive and, as a result there is greater demand for more precise statements... so that management decisions can be made with a higher degree of confidence" (see also Brubaker & Hallmark, 1991). Many of these new and more sophisticated land uses require quantitative expression of spatial variability. The descriptive and qualitative measures of variability which soil maps carry, though adequate for agricultural soil management, are often not so adequate anymore. And as Bouma (1988) put it, these don't stand up in court. This is the reason for the observed growing pressure by modern users of soil surveys for quantification of spatial variability and assignment of confidence limits for soil composition, specific soil properties, and soil performance within mapping units (see Miller, 1978; Wilding & Drees, 1983). Again, collection of statistical data (actual field observations and laboratory analysis of specific soil attributes) by transect or other types of sampling (see Brown & Huddleston, 1991) is required to provide such quantitative characterization of map units and their variability.

## 2.5 NRCS State Soil Geographic Database (STATSGO)

The USDA-Natural Resources Conservation Service (NRCS) formerly known as Soil Conservation Service (SCS) has the Federal leadership in a national effort to provide digital soil data for use in geographic information systems (GIS). NRCS has established three soil geographic databases representing kinds of soil maps at differing levels of detail. Soil Survey Geographic Data Base ( SSURGO) is the most detailed of these digital soil databases, and is made from NRCS standard county soil surveys at scales typically between 1:15,000 to 1:24,000. Soil maps for STATSGO are compiled by generalizing the more detailed SSURGO maps. Where SSURGO maps are not available, data on geology, topography, vegetation, and climate are assembled and used, together with remotely sensed satellite images. Soils of like areas are studied, and the probable classification and extent of the soils is determined. Map unit composition for STATSGO is determined by sampling areas on the more detailed maps and expanding the data statistically to characterize the whole map unit. Then, using the US Geological Surveys 1:250,000 quadrangle series as a map base, the soil data are digitized to comply with national guidelines and standards (see SCS 1991, p. 2; 1994). STATSGO, therefore, is not only subject to all the potential errors of soil survey discussed earlier, but more errors are introduced in the further process of automation. Jordon et. al., (1986) as cited in Day et al. (1988) stated that in the US, approximately 80% of published soil surveys and 50% of soil surveys in progress are on spatially distorted base maps that do not meet National Map Accuracy Standards. There are other potential sources of serious errors (Lunetta et al., 1991; Heuvelink et al., 1989; Burrough, 1987) in a geographic information system.

As Aronoff (1993) notes, error is introduced and propagated at every step in the process of generating and using geographic information. Yet as Hammer et al. (1991) noted, soil-based applications of GIS technology introduce new demands upon soil surveys and produce new users of soil survey information, many of whom may be unaware of either the potentials or limitations of soil survey information. To serve these numerous and often crucial demands well, it is important that the reliability of soil geographic databases be assessed, and the variability of specific soil attributes within their map units be further characterized.

## 2.6 Methods Of Assessing Map Unit Variability

Brubaker & Hallmark (1991) contains an excellent treatise on the methods for evaluating map unit composition. These methods have been used by Protz et al., 1968; Amos & Whiteside, 1975; Bascomb & Jarvis, 1976; Campbell, 1978; Steers & Hajek, 1979; Bigler & Liudahl, 1984; Edmonds & Lentner, 1986; Hopkins et al., 1987; Nordt et al., 1991). Quantification of map unit reliability involves selecting unbiased samples (usually by transecting but also by stratified random sampling) from delineations of map units to be studied. These samples are then used to estimate either (1) the compositional purity of the map unit in terms of TU content or (2) to evaluate the variability of individual soil properties. In the former, the objective is to determine proportion of soils within the MU that are in the same taxonomic class as the named soil or TU. Confidence intervals are then calculated using either the Student's t-distribution or a binomial method (see Wilding & Drees, 1983; Upchurch et al., 1988; and Burrough, 1991). A good soil survey was required to have a mapping unit purity of 85% or better (Soil Survey staff, 1951), but many studies such as McCormack & Wilding, 1969; Amos & Whiteside; Edmonds & Lentner, 1986; and Hopkins et al., 1987 reported taxonomic purity of 50% and less. This number increases significantly when the taxonomic purity is examined at higher levels of soil taxa, or when interpretive (instead taxonomic) purity is examined (West et al., 1981; Nordt et al., 1991). In interpretive purity, soils that were taxonomically dissimilar but had similar interpretations are allowed to be included in the map unit. The problem here is that the definitions of similar and dissimilar soils (Soil Survey Staff, 1983) used in taxonomic purity are subjective, user-biased and dependent on intended land use (Nordt et al., 1991). According to Miller et al. (1979), and Wilding & Drees (1983), taxonomic purity of map units is not a proper measure of quality or precision of soil survey.

A better approach (and also the method employed in this study) is to assess the variability of specific soil properties. This method uses parametric or nonparametric statistics to analyze the between and within map units variances for selected soil properties, and to compute summary statistics including coefficients of variation (CV) for these soil properties within map units. The results indicate the "quality" of soil map units, revealing if values of soil attributes are within the limits expected for the reference taxa they represent, as well as showing the relative efficiency with which the spatial variability of the selected soil properties is mapped. An assessment based on individual soil properties is more useful to many users (Trangmar et al., 1985), and specialists and map interpreters (Ragg & Henderson, 1980). The probability estimates of soil variability and individual soil properties provided by this approach are needed if we are to extrapolate properties from one delineation to the next (Wilding, 1988; Nordt et al., 1991).

## 2.7 Statistical Analyses Used in Soil Variability Studies

### 2.7.1 Central Tendency and Variance Statistics.

The statistical procedures for expressing the variability of a specific soil variable within an area of land (e.g., a map unit or study area) are discussed in Webster (1977), Warrick & Nielsen (1980), Wilding & Drees (1983), and overviewed more recently by Webster & Oliver (1990), and Upchurch & Edmonds (1991). These include the estimation of the mean, variance, coefficients of variation and frequency distributions for the soil population represented by the sample data set. Warrick & Nielsen (1980) stated that "···a population is more completely defined by its frequency distribution. Given the frequency distribution, we can determine all sorts of things---including averages, dispersions, and even the probability that a randomly drawn value will be within specified limits". The CV is a useful and meaningful index to compare variability among different soil properties (Wilding & Drees, 1983), while standard error of the mean and confidence limits of sampled data are used to make probability statements concerning the expected variance or limit of accuracy for randomly drawing a given size sample, and to determine the number of samples or observation necessary to estimate the mean within specified limits at desired confidence levels.

### 2.7.2 Geostatistical Techniques

Although, classical statistical procedures give valuable information about the soil population for a soil property, they say nothing about where the samples are located, and they assume independence among sample points. Soil properties are distributed in space and their values are related to their spatial location. Variables are not independent if values at points close together approximate one another and increasingly differ as the distance separating them increases. Such variables are described as having spatial structure or showing spatial dependence or spatial continuity. Geostatistics is a set of statistical tools which are extensions of classical statistics with the assumption of sample independence removed (Upchurch & Edmonds, 1991). The relationship or spatial structure among values at different location in the study area is mathematically described by the variogram. Based on the variogram, the statistical interpolation procedure of kriging is used to estimate values at any unsampled location within the study area (see Isaaks & Srivastava, 1989). Because kriging takes into account the spatial dependence in the data, its estimations have minimum variance or error. The variance or error of estimation by kriging depends only on the degree of spatial dependence and the configuration of the observation points in relation to the point or area (block) to be estimated. This error is itself estimated during kriging, and therefore can be known. Kriging is described as a "best linear unbiased estimation (b.l.u.e)" method. It is "unbiased" since it tries to have the mean residual or error of estimation that is equal to zero; and it is "best" because it aims at minimizing the variance of errors (see Isaaks & Srivastava, 1989). Kriging is also termed an optimal interpolation procedure because "the sparsest sampling intensity that can achieve a desired precision could be derived for a given soil attribute" (Odeh et al., 1990). McBratney and Webster (1983) reported 3.5 to 9 fold gain in efficiency of sampling effort required by geostatistical method (for a given estimation variance and standard error) over that estimated by classical method. Also, with kriging, very precise contour maps can be drawn for space-distributed variables, reducing sampling and analysis costs (Vieira et al, 1983).

The use of geostatistical methods has gained much support among soil scientists for examining the spatial variation of soil properties (e.g. Burgess & Webster, 1980a, b; Yost et al., 1982a, b; McBratney et al., 1982; McBratney & Webster, 1983; Uehara et al., 1984; Yates & Warrick, 1987). However, most of these studies have been for small areas where the luxury of intensive grid sampling could be afforded. The commonly reported correlation distances for soil properties are under a few hundred meters (Wierenga, 1984), although Cipra et al. (1972) reported some correlation between soil chemical properties sampled 45 km apart for a loess-derived soil, and Yost et al. (1982a) reported ranges of 32 km for some cations and pH, and 42 km for phosphorus from samples taken at 45-cm depth. Yost et al (1980a) remarked that there have been few application of these methods over distances of several kilometers as might be useful in mapping of soils and soil properties over areas which might be independently managed. This is most probably due to the large number of samples required for adequately computing variograms for such large areas. Recently, Webster & Oliver (1992) observed that many of these geostatistical studies had the variograms computed from insufficient sample sizes. Such inadequate sample sizes result in erratic variograms and large estimation variances.

Although it could not be conveniently accommodated in this dissertation, an interesting geostatistical study would be to determine optimal sampling efforts (number of samples required to adequately sample an area) for various soil attributes, and see how these compare to those determined by classical statistical methods. It was apparent that to carry out such a study, some of the selected soil variables would require additional samples or increased sampling density in order to yield stable variograms. SSURGO data seem an appropriate source from which to create supplementary data or more closely spaced samples of soil attribute values, but SSURGO data are not available for much of the study area. Consistent and well-defined study area is important in geostatistical studies because degree of soil spatial dependence or correlation length not only depends on the soil property but also on area of study, and may be a function of time (see Wierenga, 1984). The proposed geostatistical study would entail many trials to fit variograms for many of the soil variables and each of the ECOMAP Sections or Subsections. It therefore seemed appropriate to defer these geostatistics studies to the immediate future following this dissertation when such studies will be more feasible and appropriately done.

### 2.7.3 Multivariate Statistics

Soil classes are usually polythetic---class membership is based on observations of several variables, no one of which is either [absolutely] necessary or sufficient to define the class (Webster & Burrough, 1974). Soil variables are usually intricately interrelated, and it is difficult for soil map units to efficiently reflect spatial variation in all soil variables simultaneously. To assess the effectiveness of soil classification and/or evaluate the reliability of a soil survey usually involves comparing two or more map units for differences on a set of soil attributes. Classical univariate analysis of variance (ANOVA) used for assessing group differences on a single dependent variable is inappropriate when more than two variables must be considered simultaneously. Multivariate analysis of variance (MANOVA) and discriminant function analysis (DFA) are commonly used multivariate statistical analyses to answer questions about how two or more groups or classes differ from one another on the basis of multiple criteria (considered simultaneously). They also provide means for assessing the contribution or relative effectiveness of each of the variables in distinguishing the groups or predicting group membership. Except for some nuances, MANOVA is practically the same as DFA; they are used to answer the same types of research questions but stated differently (see Tabachnick & Fidell, 1996).

MANOVA evaluates the differences among centroids for a set of dependent variables when there are two or more groups or levels of an independent variable (Tabachnick & Fidell, 1996). A centroid is the multivariate equivalent of the mean of a variable. MANOVA is used to test the hypothesis that groups are significantly different; to tests if it is worth treating the map units as different from one another (Norris, 1970). Like MANOVA, discriminant function analysis (also called Multivariate Discriminant Analysis) is also a technique for analyzing the differences between groups or interpreting ways in which groups differ. With DFA, "one is able to discriminate between groups on the basis of some set of characteristics, assess how well the properties discriminate, and which characteristics are the most powerful discriminators (Klecka, 1980; see also Norris, 1970). Horton et al. (1968) used a computer program much like MANOVA (or multivariate analysis of covariance) to show that the top, slope, and depression areas of a gilgaied landscape in Queensland are significantly different taken over all properties, though not over most properties taken individually. In a similar research, Little et al. (1968) showed that soil materials on a certain valley fill were significantly different from each other taken over their contents of trace-elements and other soil properties.

Unlike MANOVA, DFA has been used very frequently in soil classification research, to both measure and test differences between soil groups (see Horton et al, 1968; Webster & Burrough, 1974; Pavlick & Hole, 1977; Webster, 1977; Duning et al., 1986). Recently, Bell et al. (1992) gave a detailed review of the application of this method in pedology. A major advantage of DFA over MANOVA is that the former also offers classification procedures to evaluate how well individual subjects (i.e., soil units or profiles) are classified into their appropriate groups (i.e., soil map units), on the basis of their scores on the independent variables (i.e., soil properties) (see Tabachnick & Fidell, 1996). Webster & Burrough (1974) and Webster (1978) discussed the advantages of using DFA as an allocation tool for soil classification. Norris and Loveday (1971) found that soil profiles classified using multivariate discriminant techniques were more consistent than classification by surveyors using their mental concepts of the modal soil for each group.

## CHAPTER 3

## DEVELOPMENT OF NORTHERN NEW ENGLAND ECOLOGICAL DATABASE FROM 1983 FIA

### 3.1 Introduction

Regional coverage, inclusion of locational attributes (in latitude and longitude), intensive field sampling and comprehensive nature of the 1983 FIA data of Maine, New Hampshire and Vermont are qualities that make these data potentially valuable for a number of scientific applications. These extensive environmental and ecological data are certainly useable for ecological modeling (e.g., ecological model validation), and ecologically-based management of natural and environmental resources at the regional scales. Being field measured and reliable data, they may have the potential to be useful as reference data for the calibration of airborne remote sensors and for classification and accuracy assessment of remotely sensed data, and cross-validation of other traditional but interpolated sources of environmental and natural resource data. However, despite these potential scientific uses, and the paucity of like regional ecological data, FIA data have merely sat in archives; being unused for research because of poor data summarization and presentation, and trouble-some computer storage format of the un-summarized data.

To be accessible and readily usable for the identified and other potential scientific research, un-summarized FIA data needed further processing and major reorganization, and adaptation to relational database management and geographic information systems technologies. This chapter discusses the research effort undertaken to achieve this goal. The specific objectives of the research were to:

(1) Develop 1983 FIA data of Maine, New Hampshire and Vermont into an ecological relational database (using *Paradox 5.x for Windows*)

(2) Create interfaces for raster- and vector-based geographic information systems (GIS) for the developed data tables

(3) Develop comprehensive hardcopy and computer on-line documentation on the contents, source, limitations, uses, etc. of the ecological databases.

During each stage in the development of the databases, necessary steps were taken to check for obvious errors such as anomalous or suspicious data, and also to correct errors that might have been introduced in the process of the database development.

## 3.2 Methods

The development of this ecological database involved the following steps: First, computer programs were written in FORTRAN to read the data tapes obtained from the Northeastern Regional Forest Service office (Radnor, Pennsylvania) into comma-separated ASCII format. The maze of data (close to 100 data files in all) were then imported into *Paradox* (relational database management system) *5.x for Windows*, and evaluated for any inconsistency or obvious errors (e.g., misplaced decimals, mis-coding, - etc.) resulting from data reading and retrieval processes. Then all related data files for each data type were added together for each state. Again, effort was exercised to ensure that the number of records summed as expected after the addition procedure. Data fields were then named appropriately following related document and FIA survey manuals. Some components of the FIA data were received from a different source and at a different time. For instance, the soil chemistry data and geographic coordinates (along with other data) were received about two years earlier, and not directly from Northeastern Forest Experiment Station. With later support and collaboration from the Forest Service, we sought to get original and perhaps more complete and valid copies of the FIA database directly from the Forest Service. These later data from Forest Service included about 70 data files many of which were not among the initial data set we had received. However, there were no data for the state of Vermont, and no soil chemistry or geographic coordinate data at all in this later batch of data. Therefore, it was necessary to combine data from both sources during the database development; we had to resort to the old data set for the "missing" components of the data received from the Forest Service.

The next step was to link or associate the data in each data file in each state with appropriate geographic coordinates. Each FIA data file (including the locational data) had three variables or keys (i.e., Unit, County, and Plot #) which when combined or *indexed,* uniquely identified each FIA plot location. However, we had two initial concerns. First we were not sure if the locational data (from old data-source) would match the later incomplete data when indexed by Unit/County/Plot#. Fortunately, they matched perfectly, thereby assuring that the potentially developed database would have the necessary spatial component. Also, since we had to supplement the data received from Forest Service with data (e.g., latitude, longitude and soil chemical properties) from the other source, we needed to verify that the latter were valid data. Fortunately again, both sources of data contained many variables (e.g., soil moisture class, soil series, rooting depth, organic depth, parent material, soil geology, soil drainage, B-texture etc.) in common. When these common data fields were compared (using Paradox query procedures) they were exactly same, plot for plot in the two data sources. This gave us confidence that the old data set was also valid and could be used. In all, only plot locations that were unique when indexed by State, County and Plot #s, and could also be associated with unique latitude and longitude values were used in the database development. These include about 2270 "New Ground" 1/5th acre plots (FIA Sample Kind = 3, see US-FS, 1982, p. 15) for ME, and about 700 for NH and 800 for VT. FIA "Remeasured" or permanent plot data were not used as these did not have associated latitude and longitude data.

To keep the plot locations uniquely and permanently identified without the use of compound keys (i.e., Unit/County/Plot#), an *Auto-increment* field was added to the indexed and *sorted* data files. The auto-increment values (originally named Serial#) were then concatenated with the state name and code (e.g., NH and 23 respectively for New Hampshire) to form *alphanumeric* and *numeric* data fields. With these fields each of the about 4000 FIA plots in the study area was uniquely identified. For instance, the FIA plot in New Hampshire which was #100 when indexed and sorted as described above, became permanently and uniquely identified in the study area as NH100 and 23100 in these fields which were named UNIQID-A and UNIQID-N respectively. These fields would later also serve as the interfaces for linking the developed data tables to *vector*- and *raster*-based GIS respectively.

Next, the ecologically relevant variables were selected from the clutter of data within each state, and re-organized into tables of related items such as soils, forest composition, site and locational data, tree-level data table, etc. In all, each state has five (5) data tables containing the following information:

1) Geographic coordinates and other locational information about each plot site
2) Soil and soil-related variables
3) General site characteristics
4) Forest composition data, and
5) Tree-level measurements.

Contents of each of the data tables are provided in Tables 3-1 to 3-5. Miscellaneous suggestions (see Bowers, 1988) were followed to develop the data into efficient, intelligible and easy-to-use databases. Finally, a hardcopy meta-data or user's guide (compiled from existing FIA documents) about the contents, definitions of variables, methods of survey, limitations of data and other relevant information about the FIA data was developed for the ecological database. A computer on-line, abridged version of the user's guide was also designed and built into the database. The on-line meta-data is in a relational database *Memo* format. It was intended to accompany all developed data table(s); to provide quick information and reference when the unabridged hardcopy user's guide is not readily within reach.

## 3.3 Results and Discussion

The names of the data tables and the variables they contain are given below. Often, a variable appears in more than one table if and when such duplication makes the affected tables more complete and sufficiently independent. Also, every table has two additional fields which contain alpha-numeric and numeric unique identifications for each plot. One of these two fields is required as the *relate item* or common field for interfacing the tables to either a raster or vector-based GIS. Item#s (in the tables) are used to easily locate the variable in the User's Guide. The full or more descriptive names and the abbreviated field names for the data variables appear in the second and third columns (respectively) of the tables.

It may be important to note that many of the FIA plots (shown in Figure 1.1) were not sampled for each and every of the variables shown in the tables, and some of the tables or groups of variables had many more samples than others. For instance, only Maine had soil series data, and very limited number of plots in the study area had soil chemical data especially CEC, OM and TN. Also, as at the time of this report, much of the data for the state of Vermont were yet to be received from the Forest Service.

| ITEM# | DESCRIPTION | FIELD NAME |
|---|---|---|
| ItemA 00 | Unique Identity | UniqidA |
| ItemN 00 | Unique Identity (numeric) | UniqidN |
| Item 990 | Latitude | Latitude |
| Item 999 | Longitude | Lngtude |
| Item 1 | State | State |
| Item 2 | Unit | Unit |
| Item 3 | County | County |
| Item 4 | Town or Sub-county | Town |

Table 3-1: Geographic coordinates and other locational information about plot sites

| ITEM# | DESCRIPTION | FIELD NAME |
|---|---|---|
| ItemA 00 | Unique Identity | UniqidA |
| ItemN 00 | Unique Identity (numeric) | UniqidN |
| Item 14 | Land use Class | Landuse |
| Item 15 | Disturbance since Photo | Disturb |
| Item 17 | Previous Land use | Prev-lnduse |
| Item 18 | Previous Date | Prev-Date |
| Item 19 | Month of Current Tally | Crrnt-Date |
| Item 25 | Aspect | Aspect |
| Item 26 | Terrain position | Trn-positn |
| Item 27 | Percent Slope | %Slope |
| Item 28 | Percent exposed Soil | %expo-soil |
| Item 65 | Distance to Nearest Road | Dist-to-Road |
| Item 66 | Recreation Opportunity | Recreatn-oppty |
| Item 75 | Water on Plot | Water-on-plt |
| Item 73 | Equipment Limitation | Eqmt-Lmtn |
| Item 74 | Surface Boulder class | Surf-Bouldr |
| Item 79 | Elevation | Elevation |

Table 3-2: General site characteristics or description of plot locations

| ITEM# | DESCRIPTION | FIELD NAME |
|---|---|---|
| ItemA 00 | Unique Identity | UniqidA |
| ItemN 00 | Unique Identity (numeric) | UniqidN |
| Item 06 | Sample Kind | Smpknd |
| Item 14 | Land use Class | Landuse |
| Item 16 | Owner Class | Owner |
| Item 17 | Previous Land use | Prev-landuse |
| Item 18 | Previous Date | Prev-Date |
| Item 19 | Month of Current Tally | rmt-Date |
| Item 24 | Point History | Pnt-hist |
| Item 29 | Less than 1ft seedling | 1ft-Seedling |
| Item 30 | Cover class | Cover |
| Item 31 | Crown Closure | Crown-closr |
| Item 32 | Foliage Condition | Foliag-Condt |
| Item 33 | Canopy height | Canopy-hgt |
| Item 34 | Stratum Volume | Stratum-vol |
| Item 35 | Life form volume | Life-form-vol |
| Item 41 | Stem Count | Stemcount |
| Item 64 | Stand Area class | Stand-area |
| Item 67 | Forest Type | Forest-type |
| Item 68 | Plot Origin | Plot-Origin |
| Item 69 | Plot age | Plot-Age |
| Item 7 | Photo Interp Class | PI Class |
| Item 70 | Timber Management | Timber-Mgt |
| Item 71 | Harvest History | Harv-History |
| Item 72 | Time Since Harvest | Since-Harvst |
| Item 76 | Browse Line | Browse-lne |
| Item 77 | Forest Openings | Forst-Opns |
| Item 78 | Edge on Plot | Edge-on-Plt |
| Item 104-107 | Site Index Trees | Site-Index |
| Item 104 | Site Index Tree Species | S-I-Tree-Spp |
| Item 105 | Site Index Tree D.B.H | S-I-Tree-D.B.H |
| Item 106 | Site Index Tree Total Height | S-I-Tree-Height |
| Item 107 | Site Index Tree Age | S-I-Tree-Age |
| Item 108 | Gross Cubic-foot Volume | G.C.F.Vol |
| Item 109 | Photo Edge Information | Phot-Edg-Info |

Table 3-3: FIA forest composition data

| ITEM# | DESCRIPTION | FIELD NAME |
|---|---|---|
| ItemA 00 | Unique Identity | UniqidA |
| ItemN 00 | Unique Identity (numeric) | UniqidN |
| Item 06 | Sample Kind | Smpknd |
| Item 14 | Land use Class | Landuse |
| Item 16 | Owner Class | Owner |
| Item 39 | Species | Species |
| Item 40 | Diameter at Breast Height | D.B.H. |
| Item 41 | Stem Count | Stemcount |
| Item 42 | Cavities | Cavities |
| Item 45 | Sawlog Length | Sawlog-lgth |
| Item 46 | Bole Length | Bole-length |
| Item 47 | Board-foot cull | Board-ft-cull |
| Item 48 | %Soundness(board-ft-cull) | %Sndness-brd |
| Item 49 | Cubic Foot cull | Cubic-ft-cull |
| Item 50 | %Soundness(cubic-ft-cull) | %Sndness-cub |
| Item 51 | Crown Ratio | Crown Ratio |
| Item 52 | Crown Class | Crown Class |
| Item 53 | Crown Availability | Crown-Avail |
| Item 54 | Primary Damage/Agent | Prmry-Damage |
| Item 57 | Tree Class | Tree-Class |
| Item 58 | Merchantability Class | Merchbty-class |
| Item 59 | Tree History | Tree-History |
| Item 60 | Previous Tree Number | Prev-Tree-# |
| Item 61 | Previous D.B.H | Prev-D.B.H. |
| Item 62 | Previous Merchantability | Prev-Merchbty |

Table 3-4: Tree-level data in FIA

| ITEM# | DESCRIPTION | FIELD NAME |
|---|---|---|
| ItemA 00 | Unique Identity | UniqidA |
| ItemN 00 | Unique Identity (numeric) | UniqidN |
| Item 79 | Elevation | Elevation |
| Item 80 | Depth of Organic Layer | Organic-Dept |
| Item 81 | Rooting Depth | Rootg-Depth |
| Item 82 | Depth to Mottling | Motlng-Dept |
| Item 83 | Subsurface Soil Texture | Subsoil-Text |
| Item 84 | Depth of Bedrock | Bedrk-Depth |
| Item 85 | Parent Material | Parent-Matl |
| Item 86 | Soil Geology | Soil-Geolgy |
| Item 87 | Soil Moisture or Drainage | Soil-Drnge |
| Item 88 | Lab Soil pH | Lab-Soil-pH |
| Item 89 | Soil Series | Soil-Series |
| Item 90 | Field Soil pH | Fld-Soil-pH |
| Item 91 | Soil extractable sodium (Na) | Extble-Na. |
| Item 92 | Soil extractable calcium (Ca) | Extble-Ca |
| Item 93 | Soil extractable Magnesium (Mg) | Extble-Mg |
| Item 94 | Soil extractable Potassium (K) | Extble-K |
| Item 95 | Soil extractable Phosphorus (P) | Extble-P |
| Item 96 | Soil extractable Aluminum (Al) | Extble-Al |
| Item 97 | Soil extractable Iron (Fe) | Extble-Fe |
| Item 98 | Soil extractable Manganese | Extble-Mn |
| Item 99 | Soil extractable Zinc (Zn) | Extble-Zn |
| Item 100 | Soil extractable Copper (Cu) | Extble-Cu |
| Item 101 | Soil Cation Exchange Capacity | Soil-CEC |
| Item 102 | Total extractable Nitrogen | Tkn- %N |
| Item 103 | extractable H+ | Extble-acid (meg/100g) |

Table 3-5: Soil and soil-related data

The ideas of this study were presented at the Second International Conference/ Workshop on Integrating Geographic Information Systems and Environmental Modeling (Smith & Hallett, 1993). The response and interests generated at this conference clearly showed that there will be many new users of the data if the un-summarized data are presented as shown in this study. There are already research proposals and Ph.D.

dissertations (mainly from University of New Hampshire) in which the results from this study would be a primary data source. The adaptation of FIA un-summarized data to RDMS and GIS will allow these data to be more efficiently analyzed. The kinds and scales of data analysis will increase, making FIA data suitable for many more uses (as mentioned above).

Although FIA has been going on for 60 years and some areas of the USA have been surveyed six times, many significant changes have occurred from one survey to another (Birdsey & Schreuder, 1992). Hansen et al. (1992) affirmed that inconsistency in data collection and processing methods creates data incompatibility among FIA projects and precludes analysis of data from more than one FIA project. This study has built a prototype database structure which if adopted for subsequent FIA surveys, will ensure consistency of data between survey projects. Such consistency among survey projects would permit FIA data in the future, to serve as a valuable tool for spatial and temporal change analyses (change detection studies) of many of the dynamic ecological variables in the data sets.

## 3.4 Summary

Hansen et al. (1992, p.3) describes the high accuracy standards with which the USDA Forest Service carries out forest inventory plans. FIA inventories are said to be designed to meet the specified sampling errors of 67-percent confidence limit (one standard error) at the state level. A 3-percent error per 1 million acres of timberland is the maximum allowable sampling error for an area. A 5-percent error per 1 billion cubic feet of growing stock on timberland is the sampling error goal for volume, removal, and annual growth (Hansen et al., 1992). FIA tree-level (species types, volume and conditions), vegetative and other landcover and landuse records were collected by Forest Service personnel to whom such activities must be familiar routine. Pertinent Forest Service publications (e.g., USDA-FS 1980 & 1982) also show high level preparation and training of the Forest Service personnel prior to FIA surveys.

The importance of, and the critical need for reliable, regional-scale data for ecological modeling and other studies is well established (e.g., Aber et al., 1993). Diekkruger et al., (1995) wrote that "considering the amount of published models it seems that it is much easier to develop a new model than verifying or validating existing computer codes. This is mainly due to the fact that laboratory and field measurements necessary for model verification are expensive...." They added that "testing a model on an independent data set is often not possible because usually those data are not available, unpublished, or not documented." However, the 1983 FIA data of the states of Maine, New Hampshire and Vermont which are a large amount of quality, field measured regional data have not been used for scientific applications due to inappropriate data organization and presentation, and problematic data storage format. This study involved the adaptation of these multi-million dollar data set to relational database management and geographic information systems technologies which are the most up-to-date and efficient computer-base systems for organizing, storing, managing, manipulating, analyzing and presenting data. By so doing, this study has provided the scientific community with regional data of rare quality and proportion, and in forms that are readily amenable to diverse types and scales of data analyses.

FIA is a national endeavor, and it is very probable that similar large and expensive data sets that are potentially valuable for research, are just sitting on the shelves and accumulating dust in other Forest Service regions in the country. It is our hope that the ideas of this study will be quickly adopted in other parts of the country where similar situation already exists. In addition, it is hoped that the US Forest Service will seriously consider the database management structure developed in this study as a model for organizing data from subsequent FIA surveys. The prospect of using FIA data for scientific research (brought about by this study) has already caused the Forest Service to start reconsidering some of its policies that may be adverse to this idea. One of such policies prohibits the Forest Service from giving out the geographic coordinates of FIA plots to scientist and/or disallows people using FIA data from visiting the plot locations. We hope that this discussion and policy re-evaluation remains in the forefront until it is resolved, hopefully in favor of more accessibility and wider use of FIA data than is presently the case.

# CHAPTER 4

# VARIABILITY OF SOIL PROPERTIES IN NORTHERN NEW ENGLAND BASED ON FIA DATA

## 4.1 Introduction

### 4.1.1 Overview

The importance of soil variability studies is well documented (e.g., Wilding & Drees, 1983; Wilding, 1984; Arnold & Wilding, 1991). The nature of soil substrata is a major abiotic factor in ecological land classification systems (e.g., ECOMAP, 1993; Avers et al., 1994), and the importance of soil-site relationship has been recognized in forest site classification systems (e.g., Pregitzer & Barnes, 1984; Corns & Annas, 1986; Zelazny et al., 1989). Knowledge of soil variability is essential in ecological research (Nash & Daugherty, 1990); to predict tree growth and timber production from forest site attributes (Mader, 1963; Blyth & Macleod, 1978); and to assess either the present status or changes in ecosystems (Grigal et al., 1991). Aber & Mellillo (1991, p.139) identified soil chemistry as a major factor determining the availability of nutrients in ecosystems. Taylor (1987) thought that tree types found on New England may be related to the chemical status of the B-horizons of on these soils, but remarked that the chemical characteristics New England forest soils have scarcely been studied.

Quantitative expressions of soil variability require intensive sampling and actual measurements of specific soil attributes. Soil data obtained in this fashion are scarce, and almost non-existent on a regional coverage, due to the prohibitive cost of field sampling and laboratory analyses. However, in the states of Maine, New Hampshire and Vermont, the 1983 USDA Forest Inventory and Analysis (FIA) survey included field and laboratory measurements of many physical and chemical soil attributes. The chemical properties were determined from samples taken from the B-horizons of about 2400 soil profiles dug at sites that were representative of FIA plot locations. Table 2 shows the soil and soil-related variables that were measured in the said FIA surveys, and their units of measurement.

| | |
|---|---|
| Depth of Organic Layer (inches) | Rooting Depth (inches) |
| Depth to Mottling (inches) | Subsurface Soil Texture |
| Depth to Bedrock (inches) | Parent Material |
| Soil Geology | Soil Moisture or Drainage |
| Soil pH | Soil Organic Matter (%) |
| Elevation (feet) | Slope (%) |
| Exchangeable calcium (mg $kg^{-1}$) | Exchangeable Magnesium (mg $kg^{-1}$) |
| Exchangeable Potassium (mg $kg^{-1}$) | Exchangeable Phosphorus (mg $kg^{-1}$) |
| Exchangeable Aluminum (mg $kg^{-1}$) | Exchangeable Iron (mg $kg^{-1}$) |
| Exchangeable Manganese (mg $kg^{-1}$) | Exchangeable Zinc (mg $kg^{-1}$) |
| Exchangeable Copper (mg $kg^{-1}$) | Exchangeable sodium (mg $kg^{-1}$) |
| total Nitrogen (%) | Exchangeable acid (mg $kg^{-1}$) |
| Cation Exchange Capacity (meq/100g) | |

Table 4-1: Soil and soil-related variables in the FIA data

### 4.1.2 Purpose of Study.

The primary purpose of this study was to explore and quantitatively describe the variability of selected soil properties in the study area based on these FIA data. The soil properties selected for this study were those on ratio-interval scales, and include exchangeable basic cations; calcium (Ca), potassium (K), sodium (Na) and magnesium (Mg), and cation exchange capacity (CEC), organic matter (SOM), depth to organic matter (OM_depth), phosphorus (P), total nitrogen (total N), the micronutrient cations; aluminum (Al), iron (Fe), manganese (Mn), zinc (Zn) and copper (Cu). Soil variability is virtually continuous, and generally increases with size of area (Beckett & Webster, 1971; Webster & Oliver, 1990; Grigal et al., 1991). Spatial variability studies of soils on landscape and regional scales are usually more meaningful when such large areas could be partitioned into smaller, more homogeneous sub-areas. Stratification of a large study area minimizes within-unit variances and maximize between-unit variances of soil properties. This increases precision (i.e., relative lack of change in repeated values) of estimates of variation for local areas (see also McBratney et al, 1981 and Stein et. al, 1988). As Chen et al. (1995) stated, spatial sampling efficiency depends on soil variability, and increases as variability decreases. Hence, minimizing within-unit variation also makes sampling designs for soil properties more efficient (see also Campbell 1978; McBratney et al., 1991).

Crepin and Johnson (1993) suggested using topography, underlying geology, and dominant vegetation type for horizontal subdivisions of the landscape for soil sampling. The recently produced ecological map of the eastern United States (Keys et al., 1995) which is based on the National Hierarchical Framework of Ecological Units (ECOMAP, 1993) has partitioned the study area into 10 section and 30 subsection map units (Figure 4-1). ECOMAP sections and subsections are based on biotic and environmental factors many of which are well known factors of soil formation (Jenny, 1941), i.e., they cause or affect soil spatial variation. In fact, ECOMAP subsections are described as smaller areas of sections, with similar surficial geology, lithology, geomorphic processes, subregional

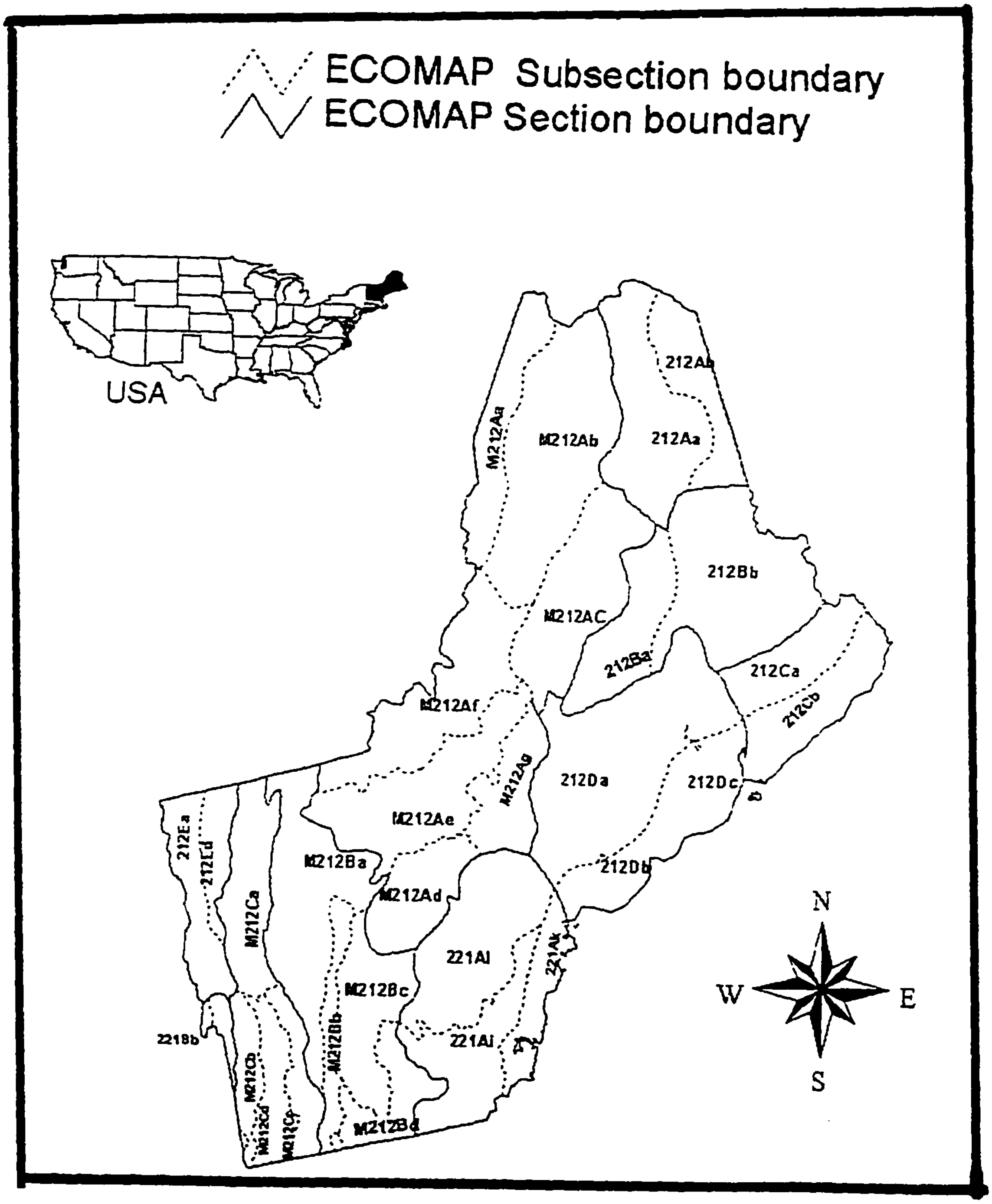



Figure 4-1: ECOMAP section and subsection boundaries in the study area

(Map units' descriptive names are given in Table 4-2 below).

| Section ID | Section Names | Subsect ID | Subsection Names |
|---|---|---|---|
| 212A | Aroostook Hills and Lowlands | | |
| | | *212Aa* | *Aroostook Lowlands* |
| | | *212Ab* | *Aroostook Hills Subsection* |
| 212B | Maine and New Brunswick Foothills and Eastern Lowlands | *212Ba* | *Central Maine Foothills* |
| | | *212Bb* | *Maine-New Brunswick Lowlands* |
| 212C | Fundy Coastal and Interior | | |
| | | *212Ca* | *Maine Eastern Interior* |
| | | *212Cb* | *Maine Eastern Coastal* |
| 212D | Central Maine and Coastal interior | | |
| | | *212Da* | *Central Maine Embayment* |
| | | *212Db* | *Penobscott Bay Coast* |
| | | *212Dc* | *Casco Bay Coast* |
| 212E | St. Lawrence and Champlain Valley | | |
| | | *212Ec* | *Chaplain Glacial Lake & Marine Plains* |
| | | *212Ed* | *Chaplain Hills Subsctions* |
| M212A | White Mountains | | |
| | | *M212Aa* | *International Boundary Plateau* |
| | | *M212Ab* | *St. John Upland Subsection* |
| | | *M212Ac* | *Maine Central Mountains* |
| | | *M212Ad* | *White Mountains Subsection* |
| | | *M212Ae* | *Mahoosic-Rangely Lakes Subsection* |
| | | *M212Af* | *Connecticut Lakes Subsection* |
| | | *M212Ag* | *Western Maine Foothills Subsection* |
| M212B | New England Piedmont | | |
| | | *M212Ba* | *Vermont Piedmont Subsection* |
| | | *M212Bb* | *Northern Connecticut River Valley* |
| | | *M212Bc* | *Sunapee Uplands Subsection* |
| | | *M212Bd* | *Hillsboro Inland Hills and Plains* |
| M212C | Green, Taconic, Berkshire Mountain | | |
| | | *M212Ca* | *Northern Green Mountain Subsection* |
| | | *M212Cb* | *Taconic Mountains Subsection* |
| | | *M212Cc* | *Berkshire-Vermont Upland Subsection* |
| | | *M212Cd* | *Sourthern Green Mountain Subsection* |
| 221A | Low New England | | |
| | | *221Ai* | *Gulf of Maine Coastal Lowlands* |
| | | *221Ak* | *Gulf of Maine Coastal Plain Subsection* |
| | | *221Al* | *Sebago-Ossipee Hills and Plains* |
| 221B | Hudson Valley | | |
| | | *221Bb* | *Taconic Foothills Subsection* |

Table 4-2: Names of ECOMAP sections and subsections map units in the study area. Source: Keys et al., 1995, Map.

climate, potential natural communities and soil groups (ECOMAP, 1993, p.4). ECOMAP sections and subsections are recommended for use in multi-forest, statewide and multi-agency analysis and assessment (Avers et al., 1994), and for data aggregation, generating and testing research hypotheses, and technology transfer and data extrapolation (Smith and Carpenter, 1996).

### 4.1.3 Study Objectives

The objectives of this study were to provide central tendency and variance statistics of the selected soil properties within each ECOMAP section in the study area; estimate confidence intervals of means, and the sample size required to estimate the population means of soil properties within ECOMAP sections; and evaluate the nature of the differences among ECOMAP ecological map units in terms of specific soil properties. Although ECOMAP subsections are expected to reflect differences in "soil types", no studies have been done to empirically assess the nature of these differences and/or to express how these map units differ in terms of specific soil properties. In this study, multivariate analysis of variance (MANOVA) was used to test if subsections within an ECOMAP section were statistically different from one another; determine the soil variables on which they differ; and assess the general effectiveness of ECOMAP ecological map units in partitioning the geographic variation in forest-soils of the study region.

Results from this study are expected to contribute to the development of ECOMAP which is still an on-going and iterative process. Smith & Carpenter (1996) expressed the need for collaborative research efforts to evaluate the validity and utility of the present ecological units, and to better understand and interpret ECOMAP map units. The regional analyses of geographic variation in soil properties will be useful for resource management and environmental research, and will provide important basis for other studies requiring field sampling and/or involving analysis of spatial patterns of soil variation in the study area.

## 4.2 Materials and Methods

### 4.2.1 Routine Soil Sampling and Analysis

Selection of soil profiles, and field sampling procedures used to collect FIA soil data are described in USFS (1982, p. 84). The soil profiles were located on sites that were representative of the overall soil conditions on FIA geo-referenced plots. Following elaborate field guide provided for them, Forest Service personnel made measurements of the field data, and also sampled the B-horizons of the soil profiles, for laboratory determination of chemical properties by soil scientists. Sample preparation, and laboratory analyses were performed by Taylor (1987), following standard procedures, and the results were written to a data-tape, in *FORTRAN*. This study started with the receipt of the data-tape, and the reading of it to retrieve the soil data in a comma-separated ASCII format. Next, the soil chemistry data were imported in a relational database management system, and were relationally joined to other soil variables and the larger FIA data as described in Chapter 3. In all, about 2400 soil units were sampled and analyzed in the study area (Taylor, 1987), but only about 1800 units were properly geo-referenced (had locational attributes) and could be used in this study.

### 4.2.2 Geographic Information Systems Procedures.

The next step in the study involved the use of geographic information systems (GIS) procedures to "combine" both FIA data and ECOMAP data, so as to identify the ECOMAP section and subsection to which each FIA plot falls. The GIS procedures were accomplished using ArcInfo---a GIS software by Environmental Systems Research Institute (ESRI), Inc. First, the ArcInfo GENERATE procedure was used to create a *point coverage* from the latitude and longitude values (in decimal degrees) of FIA plots. There were about 3800 labels in the resulting point coverage. The ECOMAP map was in *Albers Equal Area* projection and covered much (approximately 40 eastern states) of the US (see Keys et al., 1995). Using the ArcInfo CLIP command, and a *clip coverage* with the right map extent and appropriate projection, the study area (ME, NH and VT) was clipped or cut out from the ECOMAP map. Concurrent analysis of two or more maps (of the same area) in a GIS requires that the maps be *registered*. Registration is the process of ensuring that a location in one map (e.g., the point coverage showing FIA sample locations) is perfectly aligned, or corresponds to the same location on another map (e.g., the ECOMAP map of the study area). To accomplish this, the point coverage was projected from Geographic Reference grid or latitude and longitude to Albers Equal Area, using ArcInfo PROJECT procedure. Finally, IDENTITY— a point-in-polygon *overlay* procedure was used to "combine" the *point* features in the FIA coverage with the *polygon* features of ECOMAP map of the study area. With this overlay analysis, every one of the FIA plots in the point coverage also took on additional attributes from the ECOMAP map units in which they were contained. Miscellaneous GIS operations preceding this overlay procedure made it possible to identify both the ECOMAP section and subsection within which each FIA sample plot falls. The data resulting from these spatial analyses were then brought up in *Paradox for Windows* (Borland, 1995), edited to remove unnecessary variables (created by the GIS operations), and then *relationally* joined to the table of soil variables. The resulting composite data were then *exported* to *SPSS for Windows* (by SPSS, Inc.) for statistical analyses.

### 4.2.3 Data Screening and Estimation of Variability Statistics.

As an initial data screening procedure, frequency distribution and summary statistics of each soil attribute were evaluated for the study area using SPSS HISTOGRAM procedure, and the Kolmogorov-Smirnov (KS) procedure was used to test the hypothesis that the distribution of the dataset is not significantly different from normal distribution. None of the soil variables passed the Kolmogorov-Smirnov test of normality, $p < .01$, and most soil variables showed distributions that departed grossly from normality. In some instances it was possible to bring the distribution closer to normal by excluding outliers. Outliers were data values greater than the variable mean plus four standard deviations. Most of the time, however, non-linear transformation was needed to correct for departure from normal distribution. Where transformation was needed, two or more types of plausible transformation were tried, and the one which produced the most normally distributed outcome, indicated by least KS Z value and/or largest p value, was finally used. KS factor is based on the largest absolute difference between the distributions being compared, in this case, the soil variable and a hypothetical normal distribution. Distribution characteristics of soil variables were determined for the entire study area, before, and after the use of non-linear transformation. Summary statistics of soil properties were also computed, for the entire study area, and within each of the three states. The statistics were determined for the untransformed soil data before, and after the exclusion of outliers.

Next, the soil data were aggregated by ECOMAP sections, and SPSS EXAMINE procedure was used to provide central tendency and variance statistics of soil properties within each ECOMAP section. These statistics include the number of observations used in the analysis (n), the mean, minimum and maximum values; standard deviation (SD)---the square root of variance; coefficient of variation (CV) which is SD/mean, expressed as a percentage; and confidence intervals (CI) for the mean. CI is a range of values within which one can have a certain degree of confidence that the true mean lies (see Young et al., 1991), and is calculated as:

$$\mathrm{CI} = \text{mean} \pm t_{\alpha/2} \cdot \mathrm{SE}$$

where t = tabulated Student's t-value, determined by the desired alpha ($\alpha$) level or 100(1-$\alpha$/2)% confidence and appropriate degrees of freedom (V). SE is standard error (also called the standard deviation of the mean, and = SD / n. For this study, CI was calculated for 90% confidence, but any desired confidence can be readily computed for a soil variable of interest from the values of SD and n shown in Table 4-5 (see Results and Discussion).

Next, the optimum number of samples (Ɲ') required to estimate the population mean of each soil variable within specified error margins (E) or deviations from the mean, and a given level of confidence were also estimated. Ɲ' is computed as $(t^2_{\alpha/2} \cdot \mathrm{SD}^2)/E^2$, where SD and $t_{\alpha/2}$ are as defined earlier. The desired margin of error, E, is determined as a percentage of the mean of the soil property. Calculating Ɲ' with the above equation is an iterative process: first the value of t would be based on V or (n-1)... an unknown but "guesstimated" sample size. Knowing t allows $N'$ to be calculated, and this would most probably differ from the initially chosen n. Substituting v or n-1 with $N'$-1, the sample size is recalculated till v and $N'$-1 are the same. Zar (1996, p.107) discusses how to arrive at the final estimate of sample size faster, and notes that the procedure works well even if the initial guess is far from the final estimate. But Barrett & Nutt (1979, p. 65) recommend that for moderately large to large sample sizes (when n > 30), the table of confidence levels and t-values below could be used.

| Confidence Levels | t-values |
|---|---|
| .80 | 1.3 |
| .90 | 1.7 |
| .95 | 2.0 |
| .99 | 2.7 |

This is because t-values are more or less constant (for a given a level) when sample size is large (see also Webster & Oliver, 1990, p. 37). This study followed Barrett & Nutt (1979) recommended procedure, and sample sizes were determined at the 90% confidence level, given 10%, 20% and 30% margins of error.

### 4.2.4 Evaluating ECOMAP Map Units Differences.

Finally, multivariate analysis of variance (MANOVA) was used to test hypotheses about the nature of differences, if any, among ECOMAP map units. First, subsections within a given section were analyzed to determine 1) if the subsections were significantly different from one another, 2) on which soil attributes they differ, and 3) the relative proportion of variance in each soil attribute that was captured or explained by subsectional delineations. These analyses were performed in *SPSS for Windows* using the MANOVA procedure, entering all the variables on one step. Preliminary data screening and pre-analysis tests for violation of the assumptions of MANOVA were done. Often, it was necessary to apply non-linear transformations to the soil data so as to achieve normality and also stabilize the variance. Again, multiple transformations were tried within each ECOMAP section, and the one that gave best result was chosen. Soil data were also tested for multicollinearity (too high inter-correlation among discriminating variables), and for heterogeneity of within group variance-covariance matrices (through Boxes' M test). Similar but less scrupulous analysis was also performed to evaluate the variation in specific soil properties among ECOMAP sections.

## 4.3 Results and Discussion

### 4.3.1 Frequency Distributions and Normality Tests of Soil Variables

Frequency histograms, and the effects of non-linear transformation on the distribution parameters of soil variables were evaluated. Table 4-3 shows the distribution characteristics, that is, KS factor, skewness, kurtosis, and CV, before and after the use of non-linear transformations. The frequency histograms of soil variables were examined for the entire study area, and also within ECOMAP sections . The results are shown on Figure 4-2 and Figure 4-3, respectively. As seen on Figure 4-2, most soil properties were positively skewed, and none passed the Kolmogorov-Smirnov test of normality ($p < .001$). Of all the soil variables, the distributions of CEC and soil pH were the closest to normal, but their KS tests of normality were still highly significant. Most times non-linear transformations, namely, logarithmic and square root, were needed to make the soil data more normal. The logarithmic and square root transformations are only two of the members of a family of power transformations in which the observed values, $\chi_i$, are indexed by a parameter, $\lambda$, such that $\chi^\lambda$ (see Johnson & Wichern, 1982). $\lambda$ is continuous and ranges from negative to positive for $\chi > 0$. Although the appropriate $\lambda$ for any dataset can be objectively obtained by maximizing a normal likelihood function as described in Johnson & Wichern (1982, p. 162), the reciprocal transformation ($\lambda$ = -1), logarithmic transformation ($\lambda$ = 0), and the square root transformation ($\lambda$ = ½) are most commonly used because these are readily available in most statistical and data analysis software. The use of these non-linear

| Soil Variables | n | BEFORE TRANSFORMATION | | | | Transform. | AFTER TRANSFORMATION | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CV | KS | Kurt | Skew | Type | KS * | Kurt | Skew | CV |
| Exch_Acid | 1674 | 75.61 | 4.57 | 6.67 | 2.02 | sqrt | 2.35 | .98 | .38 | 38.77 |
| Exch_Al | 1679 | 90.48 | 5.13 | 4.68 | 1.78 | sqrt | 1.161 ns | .11 | .31 | 49.84 |
| Exch_Ca | 1679 | 173.03 | 11.54 | 34.71 | 4.44 | log Ca+1 | 2.02 | .36 | -.55 | 37.27 |
| CEC | 735 | 43.92 | 2.02 | 4.09 | 1.13 | log | 1.04 ns | 0.38 | 0.01 | 21.29 |
| Exch_Cu | 1679 | 174.14 | 12.58 | 95.77 | 8.34 | sqrt | 6.57 | 8.81 | 1.58 | 70.07 |
| Exch_Fe | 1679 | 94.75 | 7.18 | 20.71 | 3.59 | log Fe+1 | 1.469 ns | 2.23 | -.31 | 18.35 |
| Exch_K | 1679 | 71.75 | 4.85 | 3.73 | 1.66 | log K+1 | 2.02 | 2.36 | -.85 | 25.08 |
| Exch_Mg | 1679 | 197.46 | 12.39 | 108.70 | 7.69 | log | 2.83 | .31 | .49 | 48.27 |
| Exch_Mn | 1679 | 267.26 | 14.51 | 182.45 | 10.90 | log | 1.32 | -.22 | .10 | 68.38 |
| Exch_Na | 1679 | 67.41 | 6.52 | 11.28 | 2.22 | log Na+1 | 2.65 | 1.93 | .13 | 19.40 |
| total_N | 749 | 91.95 | 3.81 | 5.85 | 1.61 | log N+1 | 3.27 | 1.05 | 0.75 | 85.01 |
| OM_dept | 3034 | 179.29 | 17.75 | 22.63 | 4.47 | log | 5.85 | 1.03 | .66 | 29.07 |
| Exch_P | 1679 | 152.13 | 10.47 | 8.38 | 2.74 | log | 0.9268 ns | -.41 | -.12 | 66.99 |
| Soil_OM | 750 | 73.36 | 2.72 | 2.46 | 1.35 | sqrt | 1.032 ns | .02 | .30 | 38.58 |
| Soil_pH | 2327 | 12.33 | 4.39 | 3.84 | .83 | sqrt | 3.21 | 0.48 | 0.48 | 7.5 |
| Exch_Zn | 1679 | 118.51 | 9.32 | 28.69 | 4.15 | log Zn+1 | 5.28 | 1.39 | .95 | 60.81 |

Table 4-3: Distribution characteristics of soil variables, before and after non-linear transformation of data of the study area. * ns = non- significant KS test of normality, $p > .05$.

transformation almost always brought the soil data closer to normality, but most of the soil variables still could not pass the quantitative test of normality even after transformation.

Also, some of the soil variables (e.g., Acid, Al and SOM) were less responsive to logarithmic transformation; their non-normality was better corrected with the square root transformation.

The distributions of soil variables and the effect of transformation of data in ECOMAP sections were very similar to those of the study area. However, soil data in some ECOMAP sections showed much greater departure from normality, and more varied distribution shapes. Even soil pH and CEC which were almost normally distributed in the study area, showed marked non-normal distributions for some ECOMAP sections. This might have been due, in part, to the smaller sample sizes in the sub-units than for the study area. Figure 4-3 shows that the transformation that best corrected non-normality in a data set for the entire study area, usually led to better distributions within subunits also. However, a transformation that was effective on data of the study area, indicated by minimum KS factor and/or largest p value did not always produce the same result within each of the areas. The distribution characteristics (degree of skewness and kurtosis) of a data set in the sub-areas were often different from those of the study area, and from one another.

Based on this and other studies, soil pH may be the most normally distributed soil variable, especially in forested environment. In a similar study of the forest soils in north central US, Grigal et al. (1990) also found that pH was almost normally distributed, and they attributed this to the fact that pH (log of $H^+$ concentration) is already a transformed variable. This study also confirms the pedologic truism that soil variables are rarely normally distributed. However, it has also demonstrated that soil variables may not conform to the lognormal distribution as frequently as published literature would lead one to expect.

**Figure 4-2: Histograms before and after transformations of soil data of the study area**

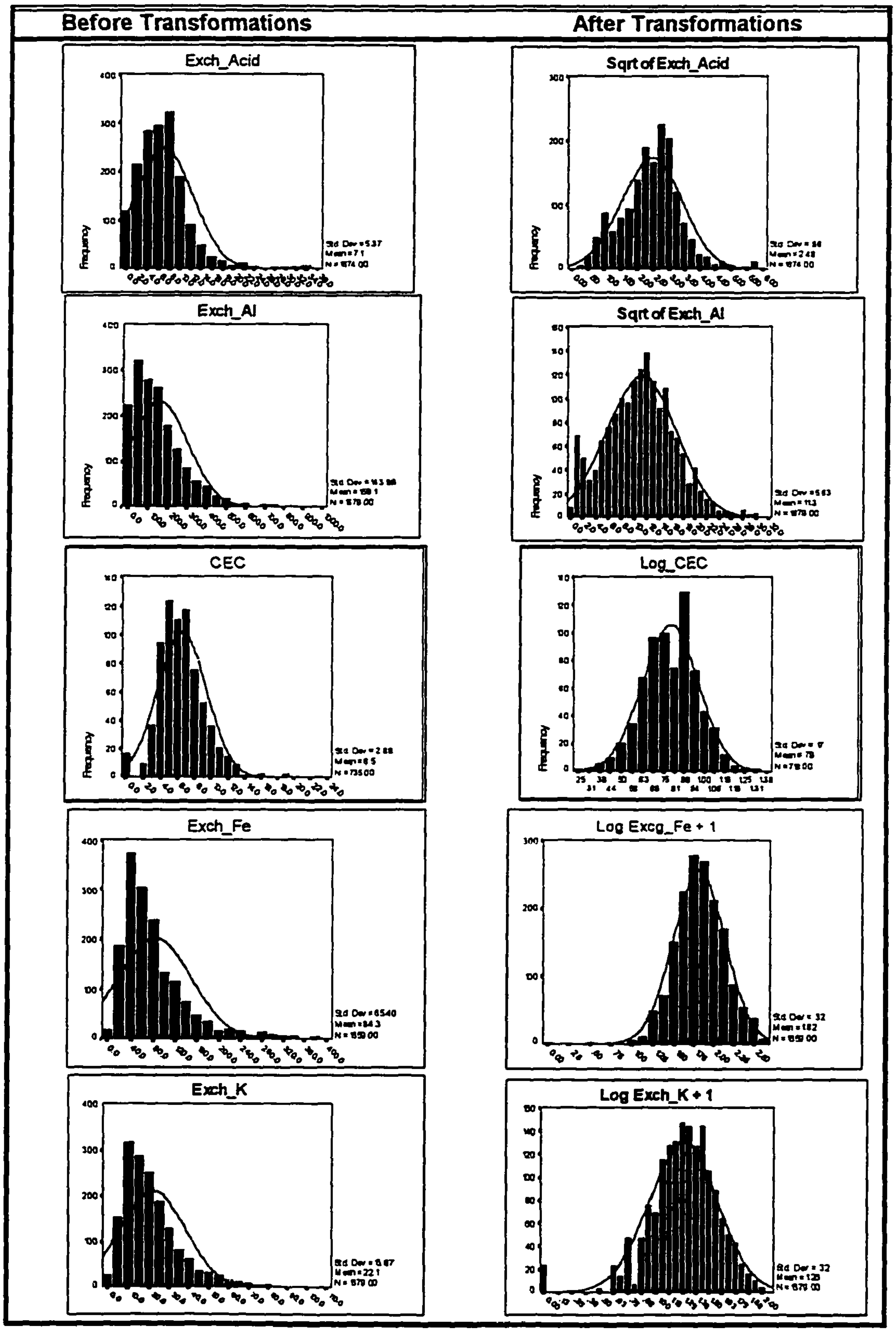


**Figure 4-2 (cont.)**

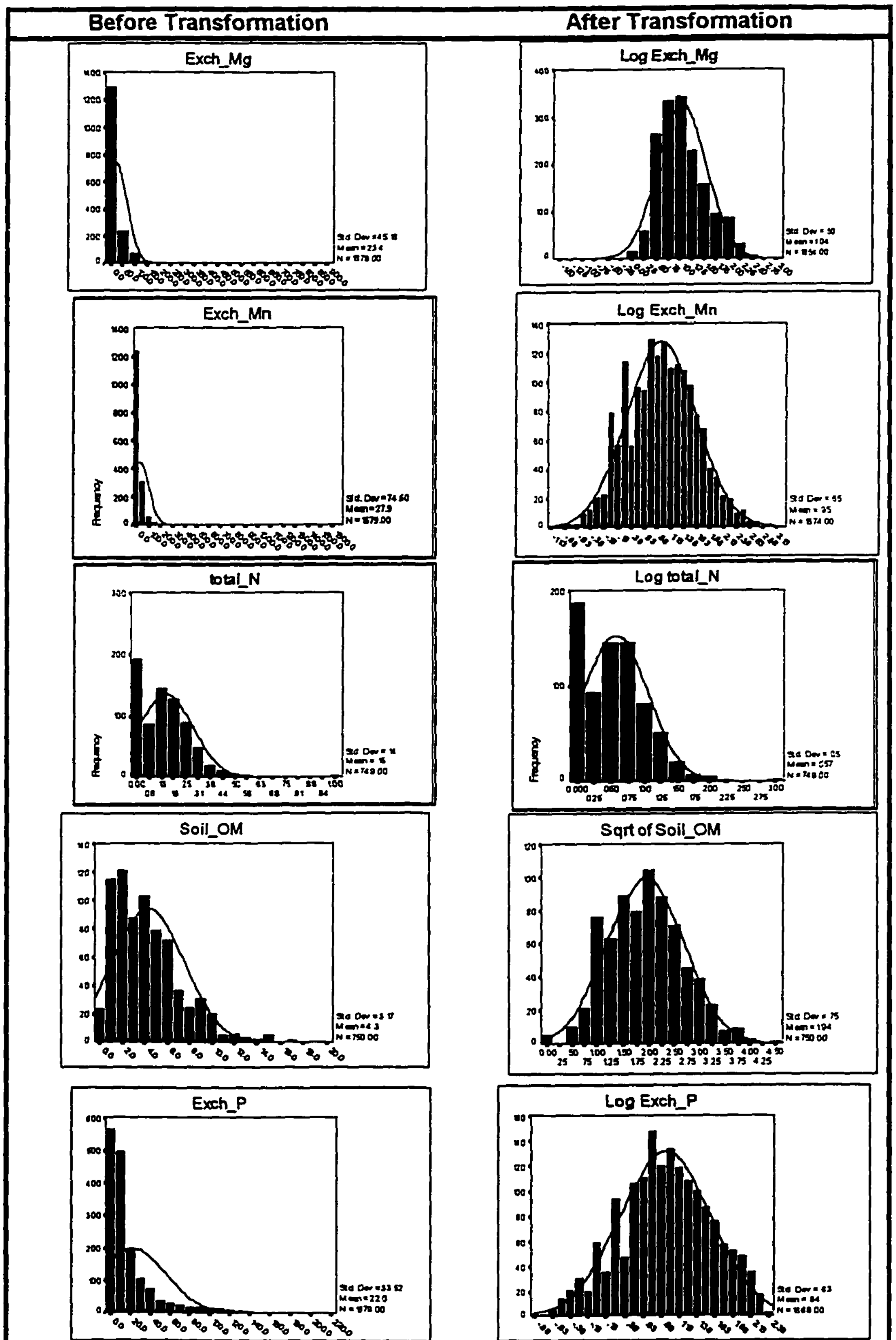


**Figure 4-3: Histograms showing that transformation of data for study area usually resulted in more normally distributed data for subareas (e.g., ECOMAP Section 7)**

**Parent dataset was not transformed** | **Parent dataset was transformed**

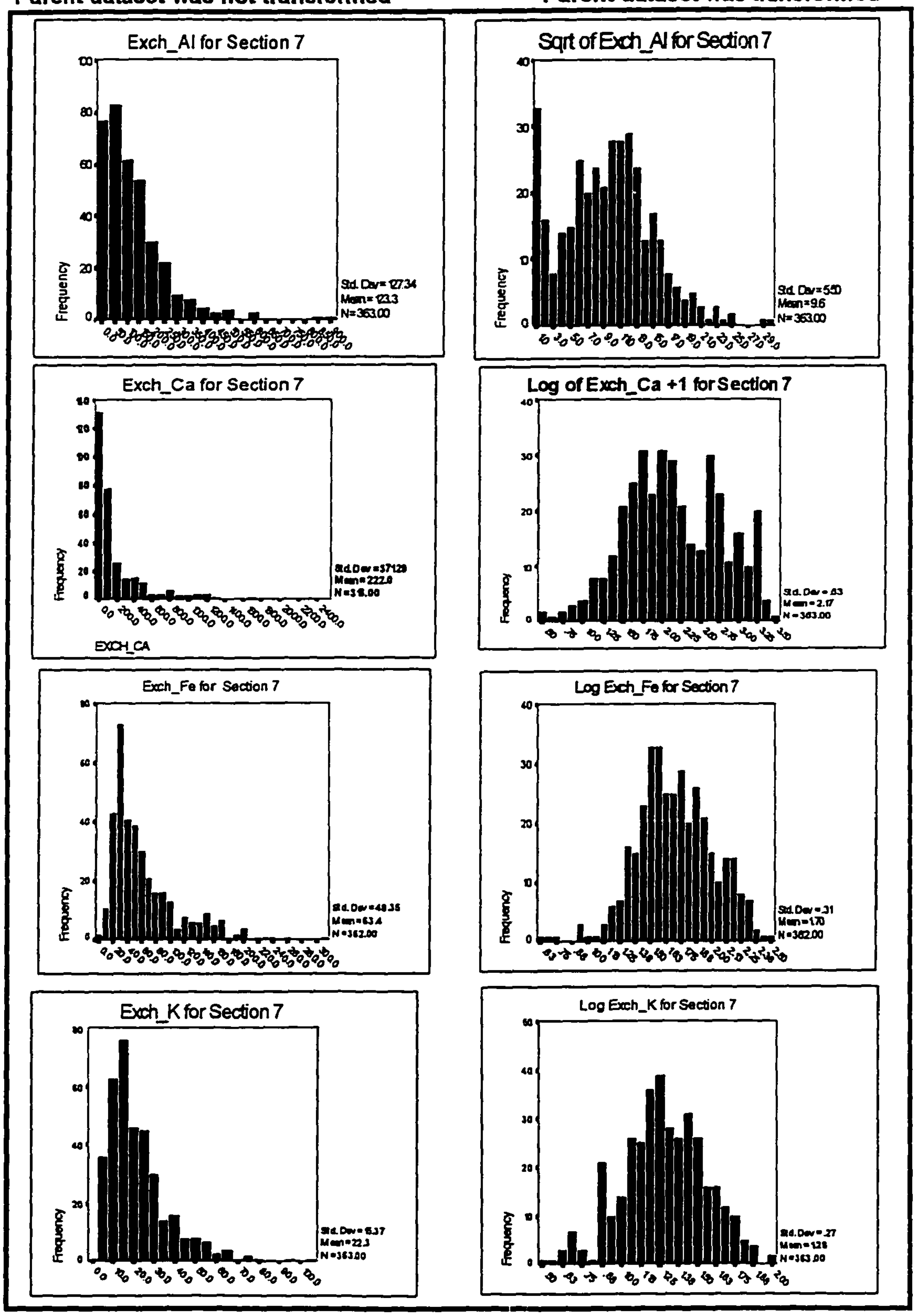
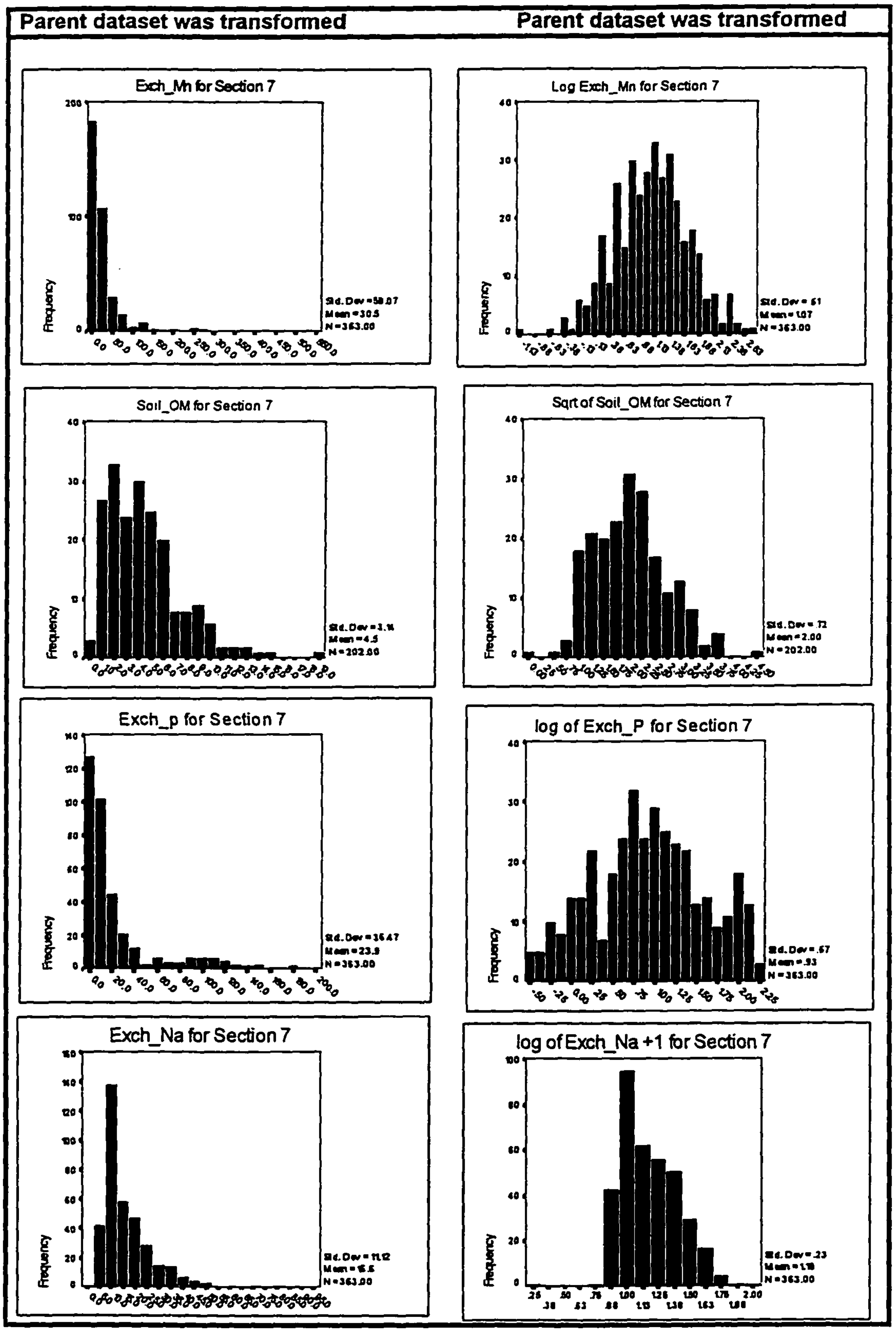


**Figure 4-3 (cont.)**



A data set is said to be lognormally distributed if the logarithm of its value is normally distributed (see Gaddum, 1945). As demonstrated in this study, soil variables show diverse distributions as a function of sample size, the size and heterogeneity of the study area, and perhaps also, the prevalent pedological factors and functions in the study site. It is important therefore that the soil or environmental scientist evaluates actual histograms, and the effect of alternative transformations if the objective is to achieve normality and/or variance stabilization. These tasks are much easier today with readily available, easy and quick-to-use data analysis programs, than they were many years ago.

### 4.3.2 Variability Statistics of Northern New England Forest Soils

Table 4-4a shows the CV, as well as the number of samples (n), maximum, mean, and standard deviation (SD) of the soil variables, for the entire study area. These summary statistics were computed both before and after outlying data values were removed, to evaluate the effect of the presence of outliers on these statistics. The results of statistical analysis performed to evaluate how the CV's and KS' of soil variables are affected by the removal of outliers, is discussed more fully in the next chapter 5, *The Legitimacy of Variability Statistics Computed From Non-normal Soil Data.* Before the removal of outliers, most variables had CV of 90% and above. Soil pH, CEC, Na, K, SOM, and Acid had low to moderate (12 to 75%) CV's before the removal of outliers, and were mostly unaffected by their removal. However, the CV's of many others variables (especially Cu, Zn, Mn, and OM_depth) which were initially above 100% were considerably reduced by the removal of outliers.

Based on published and unpublished data that were available at the time, Wilding & Drees (1983, p. 100) provided mean CV's of several soil variables, and also indicated that chemical properties, especially exchangeable Ca, Mg and K, are among the more variable soil properties, having CV's up to 160%. Table 4-4a shows that among the soil properties examined in this study, Ca, Mg, and Mn, P and OM_depth are the most variable, while pH, CEC, K, Na, and total N are the least variable in the study. For the purposes of comparison, many of the statistics computed for the study area were recomputed for each of the states of Maine, New Hampshire, and Vermont. The results, given on Table 4-4b show that the above observation about the least and most variable soil properties also applies to the states. Table 4-4b shows that in Vermont the mean value of Na (12 mg $kg^{-1}$) was highly significantly different from that of Maine and especially New Hampshire (20 mg $kg^{-1}$), $p < .001$. This is not surprising considering that unlike Vermont, a large portion of Maine, and New Hampshire to less extent, is bordered by the coast---a major source of alkalinity and sodicity in soils. On the other hand, the mean of Ca in Vermont (424 mg $kg^{-1}$) was about three times that of Maine or New Hampshire. Aside from these, Table 4-4b shows that the means and CV's of most soil properties were generally similar among the states. The apparent lack of variability of soil attributes among the states points to the fact that political boundaries are not an effective scale for studying the variability of natural phenomena.

The geographic variation of soil properties was evaluated by computing variability statistics within each of the ECOMAP sections in the study region. Table 4-5 shows the results of the analyses which include the number of soil samples (n) used for the computation, the minimum, maximum and mean values, the standard deviation

| Soil Variables | USING ALL DATA VALUES | | | | | | AFTER SCREENING DATA FOR OUTLIERS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *Max* | *Mean* | *SD* | *CV* | *KS** | *N* | *Max* | *Mean* | *SD* | *CV* | *KS** |
| **Exch_Acid** | 1674 | 36.90 | 7.10 | 5.37 | 75.61 | 4.57 | 1672 | 36.90 | 7.11 | 5.36 | 75.48 | 4.51 |
| **Exch_Al** | 1679 | 998.00 | 159.13 | 143.98 | 90.48 | 5.13 | 1660 | 778.00 | 155.71 | 133.24 | 85.57 | 5.00 |
| **Exch_Cu** | 1679 | 20.71 | .70 | 1.22 | 174.14 | 12.58 | 1355 | 4.20 | .73 | 0.58 | 79.34 | 8.01 |
| **Soil_OM** | 750 | 19.91 | 4.32 | 3.17 | 73.36 | 2.72 | 744 | 19.91 | 4.35 | 3.16 | 72.51 | 2.74 |
| **Exch_Ca** | 1679 | 7205.00 | 291.55 | 504.46 | 173.03 | 11.54 | 1649 | 2297.00 | 265.75 | 397.67 | 149.64 | 10.24 |
| **Exch_Fe** | 1679 | 939.08 | 90.01 | 85.29 | 94.75 | 7.18 | 1663 | 442.00 | 85.84 | 68.86 | 80.22 | 6.35 |
| **Exch_K** | 1679 | 109.85 | 22.12 | 15.87 | 71.75 | 4.85 | 1646 | 84.00 | 22.04 | 14.80 | 67.14 | 4.66 |
| **Exch_Na** | 1679 | 99.90 | 17.22 | 11.61 | 67.41 | 6.52 | 1660 | 62.10 | 16.71 | 9.81 | 58.69 | 5.94 |
| **Exch_Zn** | 1679 | 28.00 | 1.75 | 2.07 | 118.51 | 9.32 | 1561 | 9.84 | 1.69 | 1.50 | 88.76 | 7.55 |
| **Exch_Mg** | 1679 | 909.00 | 23.38 | 46.16 | 197.46 | 12.39 | 1638 | 203.00 | 20.83 | 30.25 | 145.24 | 10.36 |
| **Exch_Mn** | 1679 | 1666.50 | 27.91 | 74.60 | 267.26 | 14.51 | 1666 | 397.00 | 24.32 | 45.31 | 186.31 | 12.09 |
| **OM_dept** | 3034 | 40.00 | 2.88 | 5.16 | 179.29 | 17.75 | 2958 | 25.00 | 2.86 | 3.09 | 108.05 | 13.43 |
| **Exch_P** | 1679 | 216.50 | 22.04 | 33.52 | 152.13 | 10.47 | 1648 | 149.00 | 20.09 | 28.18 | 140.27 | 9.72 |
| **Soil_CEC** | 735 | 24.00 | 6.55 | 2.88 | 43.92 | 2.02 | 718 | 24.15 | 6.70 | 3.16 | 47.09 | 2.25 |
| **Total_N** | 749 | 1.00 | .15 | .14 | 91.95 | 3.81 | 574 | 1.00 | .19 | .13 | 64.73 | 2.77 |
| **Soil_pH** | 2327 | 7.75 | 4.74 | .58 | 12.15 | 4.42 | 2326 | 7.75 | 4.74 | .58 | 12.15 | 4.42 |

**Table 4-4a: Summary statistics and distribution charateristics of soil properties in the study region before, and after the removal of outliers. *** KS is an index of misfit or departure from normal distribution

| Soil Variables | MAINE | | | | NEW HAMPSHIRE | | | | VERMONT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Max | Mean | CV (%) | N | Max | Mean | CV (%) | N | Max | Mean | CV (%) |
| Exch_Acid | 1139 | 33.48 | 7.61 | 50.46 | 533 | 36.90 | 6.03 | 125.37 | n/a | n/a | n/a | n/a |
| Exch_Al | 523 | 765.00 | 155.11 | 72.28 | 549 | 753.00 | 155.54 | 83.03 | 588 | 778.00 | 156.40 | 97.86 |
| Exch_Cu | 524 | 3.95 | .61 | 53.76 | 541 | 4.20 | 0.58 | 104.34 | 290 | 4.00 | 1.24 | 49.52 |
| Soil_OM | 212 | 15.37 | 3.99 | 71.68 | 368 | 19.91 | 4.34 | 76.96 | 164 | 17.00 | 4.87 | 62.83 |
| Exch_Ca | 511 | 2281.25 | 214.85 | 147.94 | 553 | 1891.48 | 144.71 | 156.54 | 585 | 2297.00 | 424.63 | 121.94 |
| Exch_Fe | 522 | 421.60 | 99.89 | 70.21 | 547 | 442.00 | 79.07 | 82.75 | 594 | 419.00 | 79.93 | 86.43 |
| Exch_K | 502 | 83.50 | 22.69 | 68.84 | 547 | 84.00 | 22.99 | 66.25 | 597 | 82.00 | 20.61 | 65.74 |
| Exch_Na | 521 | 48.60 | 17.87 | 46.67 | 543 | 62.10 | 20.36 | 56.34 | 596 | 59.00 | 12.39 | 60.05 |
| Exch_Zn | 523 | 6.50 | 1.08 | 81.20 | 542 | 9.84 | 1.88 | 95.16 | 496 | 9.00 | 2.13 | 68.54 |
| Exch_Mg | 496 | 194.38 | 19.79 | 163.21 | 553 | 153.01 | 13.26 | 147.21 | 589 | 203.00 | 28.82 | 119.88 |
| Exch_Mn | 519 | 384.00 | 24.50 | 196.47 | 552 | 397.00 | 18.40 | 229.29 | 595 | 386.00 | 29.65 | 116.53 |
| OM_dept | 1821 | 25.00 | 3.23 | 103.72 | 567 | 23 | 2.55 | 113.33 | 570 | 24.00 | 2.00 | 111.29 |
| Exch_P | 515 | 142.80 | 20.61 | 127.46 | 546 | 137.83 | 14.96 | 161.23 | 587 | 149.00 | 24.40 | 132.30 |
| Soil_CEC | 200 | 19.13 | 6.28 | 40.61 | 365 | 24.15 | 6.82 | 41.50 | 153 | 19.00 | 6.97 | 38.16 |
| Total_N | 206 | .71 | .17 | 59.14 | 365 | .70 | 0.20 | 57.33 | n/a | n/a | n/a | n/a |
| Soil_pH | 1776 | 7.75 | 4.73 | 13.11 | 550 | 6.40 | 4.75 | 8.45 | n/a | n/a | n/a | n/a |

**Table 4-4b: Some descriptive statistics of soil properties by states in the study area.**

(SD), CV, and 90% confidence interval (CI) of the mean. Also, the sample size required to effectively estimate the mean at the 90% confidence level, and given 10, 20 and 30% marginal errors or deviations from the mean were also computed (Table 4-5). This part of the discussion will focus on general trends in these results, and will highlight important geographic patterns of variation in the means and CV's of the more important soil variables. The mean and CV were selected for the following reasons. Bivariate correlation based on the data in Table 4-5 showed that the means are highly correlated with the SD's, r (df = 79) = .97, so that ECOMAP sections which have relatively high mean values in a given soil property, almost always have relatively high SD's also. This situation is termed heteroscedasticity, and is an evidence that soil samples within ECOMAP sections come from populations with unequal variances (see Zar, 1996, p.204). Heteroscedasticity is caused by non-normality of the variables, or by the fact that one variable is related to some transformation of the others (Tabachnick & Fidell, 1996, p. 80). Both situations are rather common in soil data. The linear correlation coefficient (r) between the maximum and mean values was .88, and between maximum and SD was .95. The positive and very strong association among the mean, maximum and SD implied that it would be redundant to discuss each of these statistics. However, there was a weak relationship between the CV's and maxima, r = .40 (and even weaker relationship between the former and the mean, r = .32). This implies that the relative CV value is independent of the relative mean value, for most soil variables. The CV and mean, therefore, provide information that is similar to that of the SD and maximum, but different from each other.

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(1) Extractable Acid | | (meq $H^{+}$/100g soil) | | | | | | |
| 212A | 50 | 3.04 | 29.51 | 10.70 | 5.03 | 46.98 | 9.51 - 11.89 | 64 | 16 | 8 |
| 212B | 159 | 1.12 | 23.84 | 6.64 | 3.59 | 54.00 | 6.17 - 7.11 | 85 | 22 | 10 |
| 212C | 109 | 1.60 | 23.48 | 7.32 | 4.03 | 55.01 | 6.68 - 7.96 | 88 | 22 | 10 |
| 212D | 426 | 1.09 | 27.02 | 7.61 | 3.47 | 45.65 | 7.33 - 7.89 | 61 | 16 | 7 |
| M212A | 389 | 0.07 | 23.84 | 7.04 | 4.32 | 61.31 | 6.68 - 7.40 | 109 | 28 | 13 |
| M212B | 177 | 0.20 | 26.90 | 6.00 | 8.36 | 139.19 | 4.97 - 7.04 | 560 | 140 | 63 |
| 221A | 361 | 0.05 | 34.80 | 6.75 | 6.94 | 102.74 | 6.15 - 7.36 | 306 | 77 | 34 |
| | | :(2) Extractable Al | | (in ppm of soil) | | | | | | |
| 212D | 316 | 1.30 | 765.00 | 158.19 | 112.62 | 71.19 | 147.74 - 168.65 | 147 | 37 | 17 |
| 212E | 58 | 1.00 | 761.00 | 152.74 | 165.70 | 108.49 | 116.36 - 189.12 | 341 | 86 | 38 |
| M212A | 382 | 1.00 | 753.00 | 184.97 | 135.28 | 73.14 | 173.56 - 196.38 | 155 | 39 | 18 |
| M212B | 361 | 1.00 | 616.00 | 119.23 | 15.31 | 12.84 | 109.22 - 129.24 | 5 | 2 | 1 |
| M212C | 266 | 1.00 | 772.00 | 206.78 | 161.64 | 78.17 | 190.42 - 223.14 | 177 | 45 | 20 |
| 221A | 250 | 1.00 | 467.70 | 107.53 | 92.77 | 86.27 | 97.84 - 117.22 | 216 | 54 | 24 |
| 221B | 24 | 1.00 | 778.00 | 162.25 | 202.68 | 124.92 | 91.59 - 233.41 | 451 | 113 | 51 |
| | | :(3) Extractable Ca | | (in ppm of soil) | | | | | | |
| 212D | 306 | 0.10 | 2281.25 | 221.15 | 352.47 | 159.38 | 187.91 - 254.40 | 735 | 184 | 82 |
| 212E | 57 | 9.00 | 2256.00 | 610.21 | 601.98 | 98.65 | 476.85 - 743.57 | 282 | 71 | 32 |
| M212A | 379 | 0.10 | 1891.48 | 228.28 | 294.59 | 129.05 | 203.33 - 253.23 | 482 | 121 | 54 |
| M212B | 360 | 1.88 | 2240.00 | 354.93 | 495.52 | 139.61 | 331.86 - 398.00 | 564 | 141 | 63 |
| M212C | 270 | 4.00 | 2297.00 | 289.56 | 380.14 | 131.28 | 251.38 - 327.75 | 499 | 125 | 56 |
| 221A | 251 | 0.67 | 1448.13 | 98.97 | 173.10 | 174.89 | 80.94 - 117.012 | 884 | 221 | 99 |
| 221B | 23 | 79.00 | 2297.00 | 750.83 | 704.75 | 93.86 | 98.49 - 1003.1 | 255 | 64 | 29 |

**Table 4-5: Summary and variability statistics of soil attributes in ECOMAP sections in the study area**

Table 4-5 (cont.)

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(4) CEC (meq/100g soil) | | | | | | | | |
| 212D | 92 | 3.90 | 6.40 | 4.75 | 0.39 | 8.20 | 4.71 - 4.78 | 2 | 1 | 1 |
| 212E | 22 | 2.00 | 12.00 | 7.45 | 2.58 | 34.63 | 6.51 - 8.40 | 35 | 9 | 4 |
| M212A | 157 | 1.69 | 24.15 | 7.51 | 3.39 | 45.17 | 7.06 - 7.95 | 59 | 15 | 7 |
| M212B | 201 | 2.81 | 16.00 | 6.98 | 2.43 | 34.76 | 6.70 - 7.27 | 35 | 9 | 4 |
| M212C | 46 | 3.00 | 12.00 | 6.78 | 1.98 | 29.15 | 6.29 - 7.27 | 25 | 7 | 3 |
| 221A | 200 | 1.77 | 15.95 | 5.49 | 2.12 | 38.55 | 5.25 - 5.74 | 43 | 11 | 5 |
| | | :(5) Extractable Cu (in ppm of soil) | | | | | | | | |
| 212D | 316 | 0.15 | 3.95 | 0.62 | 0.35 | 56.18 | 0.59 - 0.66 | 92 | 23 | 11 |
| 212E | 31 | 1.00 | 3.00 | 1.32 | 0.60 | 45.28 | 1.14 - 1.51 | 60 | 15 | 7 |
| M212A | 321 | 0.05 | 4.00 | 0.72 | 0.61 | 84.84 | 0.66 - 0.78 | 209 | 53 | 24 |
| M212B | 247 | 0.02 | 3.24 | 0.61 | 0.47 | 77.09 | 0.56 - 0.66 | 172 | 43 | 20 |
| M212C | 171 | 1.00 | 4.00 | 1.28 | 0.66 | 51.76 | 1.19 - 1.36 | 78 | 20 | 9 |
| 221A | 250 | 1.00 | 4.20 | 0.52 | 0.50 | 96.73 | 0.47 - 0.57 | 271 | 68 | 31 |
| 221B | 17 | 1.00 | 4.00 | 1.35 | 0.79 | 58.09 | 1.02 - 1.69 | 98 | 25 | 11 |
| | | :(6) Extractable Fe (in ppm of soil) | | | | | | | | |
| 212D | 316 | 11.00 | 387.00 | 94.33 | 65.91 | 69.87 | 88.21 - 100.44 | 142 | 36 | 16 |
| 212E | 63 | 2.00 | 249.00 | 65.05 | 48.37 | 74.36 | 54.87 - 75.22 | 160 | 40 | 18 |
| M212A | 378 | 4.00 | 435.30 | 106.28 | 81.34 | 76.53 | 99.38 - 113.18 | 170 | 43 | 19 |
| M212B | 363 | 3.00 | 442.00 | 64.45 | 52.22 | 81.01 | 59.93 - 68.97 | 190 | 48 | 22 |
| M212C | 267 | 1.00 | 419.00 | 105.30 | 81.67 | 77.56 | 97.05 - 113.55 | 174 | 44 | 20 |
| 221A | 249 | 15.65 | 277.00 | 60.36 | 37.83 | 62.68 | 56.40 - 64.31 | 114 | 29 | 13 |
| 221B | 24 | 14.00 | 272.00 | 83.58 | 69.91 | 83.64 | 59.12 - 108.04 | 203 | 51 | 23 |

Table 4-5 (cont.)

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(7) Extractable K | | (in ppm of soil) | | | | | | |
| 212D | 301 | 1.30 | 83.50 | 24.21 | 16.46 | 68.00 | 22.65 - 25.78 | 134 | 34 | 15 |
| 212E | 62 | 4.00 | 72.00 | 23.45 | 16.35 | 69.72 | 19.98 - 26.92 | 141 | 36 | 16 |
| M212A | 364 | 3.00 | 83.50 | 22.09 | 15.39 | 69.68 | 20.78 - 23.40 | 141 | 36 | 16 |
| M212B | 361 | 2.00 | 78.00 | 21.85 | 14.32 | 65.52 | 20.61 - 23.10 | 125 | 32 | 14 |
| M212C | 272 | 4.00 | 82.00 | 20.78 | 13.12 | 63.10 | 19.47 - 22.10 | 116 | 29 | 13 |
| 221A | 250 | 2.00 | 84.00 | 20.57 | 13.96 | 67.90 | 19.10 - 22.02 | 134 | 34 | 15 |
| 221B | 23 | 9.00 | 41.00 | 23.30 | 9.70 | 41.63 | 19.83 - 26.78 | 51 | 13 | 6 |
| | | :(8) Extractable Mg | | (in ppm of soil) | | | | | | |
| 212D | 297 | 0.03 | 194.38 | 22.06 | 35.98 | 163.09 | 18.61 - 25.50 | 769 | 193 | 86 |
| 212E | 57 | 1.00 | 178.00 | 42.32 | 52.88 | 124.96 | 30.60 - 54.03 | 452 | 113 | 51 |
| M212A | 375 | 0.62 | 167.50 | 19.91 | 27.26 | 136.92 | 17.59 - 22.23 | 542 | 136 | 61 |
| M212B | 361 | 1.00 | 189.00 | 23.57 | 29.46 | 125.01 | 21.01 - 26.12 | 452 | 113 | 51 |
| M212C | 271 | 1.00 | 203.00 | 22.82 | 29.43 | 128.93 | 19.87 - 25.77 | 481 | 121 | 54 |
| 221A | 250 | 0.67 | 110.00 | 8.80 | 13.29 | 151.11 | 7.41 - 10.18 | 660 | 165 | 74 |
| 221B | 24 | 3.00 | 103.00 | 30.25 | 29.54 | 97.65 | 19.92 - 40.58 | 276 | 69 | 31 |
| | | :(9) Extractable Mn | | (in ppm of soil) | | | | | | |
| 212D | 313 | 0.20 | 384.00 | 28.39 | 53.75 | 189.35 | 23.37 - 33.40 | 1037 | 260 | 116 |
| 212E | 64 | 1.00 | 154.00 | 35.33 | 37.76 | 106.88 | 27.45 - 43.21 | 331 | 83 | 37 |
| M212A | 379 | 0.15 | 349.60 | 19.58 | 38.67 | 197.47 | 16.31 - 22.86 | 1127 | 282 | 126 |
| M212B | 362 | 0.08 | 397.00 | 29.00 | 51.11 | 176.23 | 24.57 - 33.43 | 898 | 225 | 100 |
| M212C | 270 | 1.00 | 386.00 | 28.59 | 47.63 | 166.60 | 23.80 - 33.37 | 803 | 201 | 90 |
| 221A | 251 | 0.20 | 124.00 | 9.51 | 18.18 | 191.17 | 7.62 - 11.40 | 1057 | 265 | 118 |
| 221B | 24 | 8.00 | 318.00 | 48.25 | 65.73 | 136.24 | 25.25 - 71.25 | 537 | 135 | 60 |

## Table 4-5 (cont.)

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(10) Total N (% of dry soil) | | | | | | | | |
| 212D | 90 | 0.01 | 0.38 | 0.15 | 0.08 | 55.31 | 0.135 - .164 | 89 | 23 | 10 |
| M212A | 136 | 0.02 | 1.00 | 0.22 | 0.14 | 63.53 | .200 - .240 | 117 | 30 | 13 |
| M212B | 147 | 0.01 | 0.70 | 0.21 | 0.12 | 57.99 | .191 - .224 | 98 | 25 | 11 |
| 221A | 199 | 0.01 | 0.71 | 0.18 | 0.10 | 56.98 | .167 - .191 | 94 | 24 | 11 |
| | | :(11) Extractable Na (in ppm of soil) | | | | | | | | |
| 212D | 314 | 7.20 | 48.60 | 20.64 | 8.95 | 43.33 | 19.81 - 21.48 | 55 | 14 | 7 |
| 212E | 62 | 4.00 | 48.00 | 12.39 | 7.79 | 62.86 | 11.19 - 14.49 | 115 | 29 | 13 |
| M212A | 380 | 3.30 | 62.10 | 17.92 | 11.69 | 65.22 | 16.93 - 18.91 | 123 | 31 | 14 |
| M212B | 361 | 1.00 | 62.00 | 16.19 | 9.96 | 61.51 | 15.33 - 17.06 | 110 | 28 | 13 |
| M212C | 271 | 4.00 | 59.00 | 13.99 | 8.57 | 61.25 | 13.13 - 14.85 | 109 | 28 | 13 |
| 221A | 246 | 4.29 | 55.31 | 15.09 | 7.21 | 47.77 | 14.33 - 15.85 | 66 | 17 | 8 |
| 221B | 23 | 5.00 | 32.00 | 12.304 | 6.65 | 54.04 | 9.92 - 14.69 | 85 | 22 | 10 |
| | | :(12) % SOM | | | | | | | | |
| 212D | 94 | 0.21 | 11.16 | 3.11 | 2.18 | 70.23 | 2.732 - 3.479 | 143 | 36 | 16 |
| 212E | 25 | 1.00 | 13.00 | 5.40 | 2.96 | 54.78 | 4.3878 - 6.41 | 87 | 22 | 10 |
| M212A | 179 | 0.42 | 19.91 | 5.91 | 3.84 | 64.92 | 5.435 - 6.384 | 122 | 31 | 14 |
| M212B | 201 | 0.36 | 19.23 | 4.54 | 3.13 | 69.05 | 4.17 - 4.90 | 138 | 35 | 16 |
| M212C | 44 | 1.00 | 9.00 | 4.48 | 2.04 | 45.57 | 3.96 - 4.99 | 61 | 16 | 7 |
| 221A | 201 | 0.26 | 14.50 | 3.21 | 2.36 | 73.54 | 2.94 - 3.489 | 157 | 40 | 18 |

## Table 4-5 (cont.)

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(13) Depth of Organic Matter (cm) | | | | | | | | |
| 212A | 143 | 0.50 | 25.00 | 3.92 | 4.35 | 110.96 | 3.32 - 4.53 | 356 | 89 | 40 |
| 212B | 212 | 0.50 | 25.00 | 3.25 | 3.09 | 95.06 | 2.90 - 3.61 | 262 | 66 | 30 |
| 212C | 116 | 0.50 | 20.00 | 3.26 | 2.94 | 90.06 | 2.81 - 3.71 | 235 | 59 | 27 |
| 212D | 462 | 0.50 | 24.00 | 2.54 | 3.10 | 122.17 | 2.30 - 2.77 | 432 | 108 | 48 |
| 212E | 45 | 0.50 | 24.00 | 2.67 | 3.61 | 135.25 | 1.76 - 3.57 | 529 | 133 | 59 |
| M212A | 951 | 0.50 | 25.00 | 3.43 | 3.28 | 95.71 | 3.25 - 3.60 | 265 | 67 | 30 |
| M212B | 368 | 0.50 | 23.00 | 2.28 | 2.85 | 124.96 | 2.04 - 2.53 | 452 | 113 | 51 |
| M212C | 270 | 0.50 | 16.00 | 1.87 | 1.65 | 88.47 | 1.70 - 2.03 | 227 | 57 | 26 |
| 221A | 371 | 0.50 | 22.00 | 2.41 | 2.61 | 108.30 | 2.19 - 2.63 | 339 | 85 | 38 |
| 221B | 17 | 0.50 | 8.50 | 2.09 | 1.96 | 93.87 | 1.26 - 2.92 | 255 | 64 | 29 |
| | | :(14) Extractable P (in ppm of soil) | | | | | | | | |
| 212D | 309 | 0.30 | 142.80 | 17.18 | 25.12 | 146.19 | 14.83 - 19.54 | 618 | 155 | 69 |
| 212E | 62 | 2.00 | 146.00 | 32.21 | 33.74 | 104.73 | 25.05 - 39.37 | 318 | 80 | 36 |
| M212A | 381 | 0.21 | 137.83 | 18.26 | 24.82 | 135.94 | 16.16 - 20.35 | 535 | 134 | 60 |
| M212B | 360 | 0.31 | 149.00 | 22.51 | 33.36 | 148.21 | 19.61 - 25.40 | 635 | 159 | 71 |
| M212C | 262 | 1.00 | 145.00 | 18.87 | 27.11 | 143.70 | 16.10 - 21.63 | 597 | 150 | 67 |
| 221A | 247 | 0.13 | 126.50 | 20.47 | 27.16 | 132.68 | 17.62 - 23.32 | 509 | 128 | 57 |
| 221B | 24 | 2.00 | 131.00 | 9.73 | 29.57 | 303.90 | 16.99 - 27.68 | 27 | 7 | 3 |
| | | :(15) Soil pH | | | | | | | | |
| 212A | 169 | 3.4 | 7.35 | 4.97 | 0.77 | 15.41 | 4.87 - 5.06 | 7 | 2 | 1 |
| 212B | 209 | 3.65 | 6.75 | 4.65 | 0.57 | 12.28 | 4.58 - 4.54 | 5 | 2 | 1 |
| 212C | 112 | 3.5 | 6.25 | 4.47 | 0.46 | 10.27 | 4.40 - 4.54 | 4 | 1 | 1 |
| 212D | 440 | 3.45 | 6.80 | 4.67 | 0.54 | 11.63 | 4.63 - 4.72 | 4 | 1 | 1 |
| M212A | 837 | 3.40 | 7.75 | 4.78 | 0.65 | 13.51 | 4.74 - 4.81 | 6 | 2 | 1 |
| M212B | 183 | 3.80 | 6.00 | 4.77 | 0.38 | 7.87 | 4.72 - 4.81 | 2 | 1 | 1 |
| 221A | 375 | 3.90 | 6.40 | 3.95 | 0.39 | 10.01 | 4.71 - 4.78 | 4 | 1 | 1 |

Table 4-5 (cont.)

| Section MUID | N | Min | Max | Mean | SD | CV (%) | 90%CI of the Mean | Sample size requirements* 10% error | 20% error | 30%error |
|---|---|---|---|---|---|---|---|---|---|---|
| | | :(16) Extractable Zn | (in ppm of soil) | | | | | | | |
| 212D | 316 | 0.10 | 6.50 | 1.04 | 0.83 | 79.21 | 0.97 - 1.12 | 182 | 46 | 21 |
| 212E | 55 | 1.00 | 6.00 | 2.06 | 1.35 | 65.84 | 1.75 - 2.36 | 126 | 32 | 14 |
| M212A | 364 | 0.18 | 9.84 | 1.90 | 1.83 | 96.32 | 1.74 - 2.06 | 269 | 68 | 30 |
| M212B | 312 | 0.10 | 9.00 | 1.84 | 1.49 | 80.98 | 1.70 - 1.98 | 190 | 48 | 22 |
| M212C | 246 | 1.00 | 8.00 | 2.35 | 1.53 | 65.13 | 2.19 - 2.51 | 123 | 31 | 14 |
| 221A | 249 | 0.07 | 9.16 | 1.24 | 1.17 | 94.26 | 1.12 - 1.36 | 257 | 65 | 29 |
| 221B | 18 | 1.00 | 5.00 | 2.22 | 1.35 | 60.89 | 1.67 - 2.78 | 108 | 27 | 12 |

* Sample Size required for estimating the mean at 90% confidence level given 10, 20 and 30% error margins

The mean and maximum values in Table 4-5 as well as Tables 4-4a & 4-4b are all within limits of published (e.g., Tisdale et al., 1985) concentrations of these nutrients in US' soils. Exchangeable Acid, CEC, pH, total N, and SOM were not sampled or analyzed in ECOMAP section 221B. In many instances, the CV of available data in this section was consistently the least (e.g., Zn, P, K, and Mn), or next to the least (e.g., Mn and Na) among ECOMAP sections. However, because of the limited sample sizes ($n \leq 24$) used for the analyses in Section 221B, much significance may not be attached to the apparent trend in the statistics of this section.

Soil pH, CEC and total N which were identified as least variable in the study area (Tables 4-4a & 4-4b), also showed the least variation in means among ECOMAP sections. Low variation among ECOMAP section means implies that the mean of such a soil property in any section may be well approximated by the mean of the entire study area (Tables 4-4). Moreover, pH, CEC, total N, in addition to K and Na exhibited low to moderate variability or CV's within most sections. Most ECOMAP sections had a mean soil pH around 4.70 except 221A which was more acidic, with mean pH of 3.75. The maximum CV for pH within ECOMAP section was observed in section 212A. The mean of CEC ranged from 4.75 meq/100g in 212D to 7.51 meq/100g in M212A. The CV for CEC was under 50% within all sections, and was the lowest (8.20%) in section 212D. Data on total N were available only in four ECOMAP sections where the CV's ranged from .15 to .22%. Of all the basic cations, K seemed to be the least variable among sub-areas. Mean K concentrations ranged between 20.57 mg $kg^{-1}$ in section 221A to 24.21 mg $kg^{-1}$ in section 212D. The CV's of K were > 60% < 70% within all ECOMAP sections.

As noted in the analysis for the study area (Tables 4-4a & 4-4b), Ca, Mg, Mn. and to a less extent, P and OM_depth, also had the most variable means among ECOMAP sections, and highest CV's within sections. The mean concentrations of Ca range from less than 100 mg $kg^{-1}$ in section 221A to more than 750 mg $kg^{-1}$ in section 221B. The means of Mg were between 8.80 mg $kg^{-1}$ in section 221A to 42.32 mg $kg^{-1}$ in section 212E. There was substantial variability in both Ca and Mg concentrations within ECOMAP sections, indicated by very large SD values and CV's well above 100%. The CV of Ca in section 212E and 221B were exception, and were slightly under 100%. Although the mean concentration of Na varied well among ECOMAP sections, the variability within any section was low to moderate, CV's $\geq$ 65%. Variability of SOM both among and within ECOMAP sections was moderate. The means of SOM ranged from 3.11% in 212D to 5.91% in M212A, and the CV's ranged from 45% in M212C to 70% in 212D. The highest mean depth of organic matter, 3.92 inches, was observed in section 212A, while the highest CV's or within unit variability was in 212E and 212D. Both depth of organic matter and P varied considerably from section to section, and within each section. The means concentrations of P ranged from about 17 mg $kg^{-1}$ in section 212D to 32.21 mg $kg^{-1}$ in 212E, and the CV's were consistently above 100%.

The preceding discussion, and further evaluation of Table 4-5 show that ECOMAP section M212C has the highest or next to the highest means for Al, Fe, and Zn, but lowest mean in organic matter depth, and lowest within-unit variability or CV in organic matter (SOM) and organic matter depth. Section 212D had the highest means for K and Na, and lowest or second to the lowest means for SOM, CEC, total_N and Zn. Section 212A had data for only exchangeable acid, organic matter depth and pH, and had the highest means for all three variables. Notable was section 212E which has the highest means for SOM, CEC, Ca, K, Mg, P, Cu, and Mn, but had the lowest mean for Na. Section 212E also has the maximum CV's for Al and organic matter depth, but the lowest CV's for Cu, Ca, and Mn. Section 212E had markedly small sample sizes for most variables. However, Figure 4-1 shows that this ECOMAP section also has limited aerial extent, perhaps proportionate to the number of samples used. On the other end from section 212E, is 221A which has the lowest or next to the lowest means for SOM, CEC, Ca, Mg, K, Fe, Al, Cu, total_N, Mn and pH. ECOMAP section 221A also has the maximum within unit variability for Acid, Cu, SOM, Na, Zn, Mn, and Mg. Descriptions of the lithology and soil taxa (see McNab & Avers, 1994), for example, about ECOMAP sections would offer explanation for some of the trends observed in these sections. Linear correlation between pairs of soil variables was evaluated, first based on data for the entire study area, and then within states and ECOMAP sections. As expected, inter-correlation among soil attributes provides an aid to the understanding of why highs and/or lows of certain of soil variables were frequently associated as seen above. However, this study showed that the strength of association between pairs of soil variables changes spatially depending on the location and scale of reference or size of the study area. For example, the correlation coefficients between exchangeable acid and other soil variables were all less than .25 in the study area. In New Hampshire, and all ECOMAP sections for which there were data on exchangeable acid (except 212D), it had similar maximum r, but the variable it was most correlated with was usually different. However, in Maine, acid had r = .65 with SOM, .56 with Al, .50 with total_N, and -.42 with pH. And within section 212D, these r values were .74, .69, .66, and -.52 for SOM, Al, total_N, and pH, respectively. The influence of SOM and Al on acidity, the relationship between pH and soil acidity, and between SOM and total_N are well established and even expected in soil science and ecology. But, this study demonstrates the importance of scale and prevailing local conditions in evaluating such global "truisms" about inter-correlations among soil variables. The local pedo-geomorphic status in ECOMAP section 212D (see McNab & Avers, 1994) allow some relationships among soil variables that may not otherwise hold true on a larger scale.

### 4.3.3 Required Sampling Intensity of Soil Variables in ECOMAP Sections

The sample sizes required to estimate the mean of each soil property within each ECOMAP section, at the 90% confidence level, and given 10%, 20, and 30% deviations from the means, are also given in Table 4-5. Optimum sampling intensity is related to the variability of that attribute, and hence to CV. The linear relationship between CV and sample size requirements at the 20% marginal error was evaluated using the data on Table 4-5. The Pearson's correlation coefficient, $r(df = 110) = .96$, $p < .001$. The strong correlation between CV and sample size also held true within each ECOMAP section, $r(13 - 16) = .98$ on average, and $p < .001$ always. The strength of association between CV and sample size was exactly the same for 10% and 30% error margins also. Although the relative variability of individual soil attributes changes spatially, soil pH, CEC and total_N generally required the smallest sample sizes in most ECOMAP sections. For pH and CEC, the sample sizes required even at 10% error margin was usually less than 50. Next to this group of soil variables are Na, K, SOM and Fe; these require sample sizes of 50 or less within most sections at 20 or 30% error margins. As expected, the largest sample sizes were required by Mn > Ca > Mg > P. These highly variable soil properties required sample sizes much greater than 100 at the 30% error, about 300 at the 20% error, and close to 1000 at the 10% error margin in many of the studied ECOMAP sections.

As shown earlier (Figure 4-1 and Table 4-2), there are ten ECOMAP sections in the study area. Soil chemistry data (except exchangeable acid and pH) were not available for 212A, 212B, and 212C---three of the four ECOMAP sections in the New England geomorphic province or northeastern parts of Maine. It was interesting to compare the actual number of FIA soil samples (n), and the sample sizes required to efficiently estimate the means of the selected soil properties within the ECOMAP sections included in this study. Table 4-5 shows that over 90% of the time, the number of soil samples (n) analyzed during the 1983 FIA surveys were more than the sample sizes required to efficiently estimate the means of the selected chemical properties within each ECOMAP section, at the 90% confidence level and given 20% error margin or greater. The small n's available for section 221B were noted earlier, and this section accounts for many of the occasions where n was smaller than the optimum sample size required. Almost 100% of the time, the sample size requirements at the 10% error margin were substantially greater than n's for Mn, Ca, P, and Depth of organic matter. When these highly variable soil properties were excluded, n's were larger than the sample sizes requirements at the 10% error margin over 80% of the time.

Although small marginal errors are desirable, sample sizes required to achieve ambitiously low (10% or less) error margins are often impractical for most soil variables. Allowable error of 20% has been commonly used in soil studies, and error margins of 25% and more are warranted where sample sizes estimated at higher precision levels are too unrealistic or impractical to carry out (see Troedsson & Tamm, 1969; Wilding, 1984; Grigal et al., 1990). Choosing a lower confidence level has been the alternative to low precision or high allowable error margins in sampling schemes of highly variable soil attributes. The traditional 95% level used in many fields is a rarity in soils, and confidence levels of 80 to 90% have been commonly reported in the literature. Wilding (1984) remarked that in soils a confidence level of 70 to 80% is probably more realistic in terms of time and money inputs that are practical to a sampling scheme, and that there are many circumstances under which confidence levels lower than 80% and error margins up to 50% would still permit sufficiently accurate mean estimates (see Grigal et al., 1990).

### 4.3.4 The Nature of Soil Variation Among ECOMAP Map Units

Table 4-6 shows the results of MANOVA, to test the hypothesis that subsections within an ECOMAP section do not differ significantly from one another with respect to selected soil properties. This analysis was done only for ECOMAP sections with three or more subsections, and where these subsections had sufficient samples for the soil variables used. MANOVA (like many other multivariate methods) requires that sample-to-variable ratio exceed certain threshold to ensure that the results of the analysis are not unstable. It is recommended that the number of samples (n) in the smallest group be greater than five times the number of variables (Tabachnick & Fidell, 1996, p. 513), and/or that the total n in the analysis be equal to or greater than ten times the number of variables (Norman & Streiner, 1997, p. 132). When n-to-variable ratios are critically low, the robustness of MANOVA to violations of assumptions of homoscedasticity and multivariate normality diminishes, and the analysis results (significant or not) become unreliable and unlikely to be found if the study were replicated (Tabachnick & Fidell, 1996; Norman & Streiner, 1997). To guarantee favorable ratios, the selected soil variables in this study were rationally divided into two groups (Table 4-6), and MANOVA among the selected sections was done for each of the variable groups.

Within each section, data were evaluated for conformity to the assumption of normality. In most instances, log and square-root transformations where used to achieve univariate normality and/or homoscedasticity. Box's M test for homogeneity of within-group variance-covariance matrices was, in most cases, highly significant, $p < .01$. However, the Box's M test is notably sensitive and has been known to reject the hypothesis of homogeneity of dispersion matrices even when the group covariance matrices are not really dissimilar (Norusis, 1990, p. 104). Heuristically, heterogeneity is perceived if there are variables having ratios > 10:1 between the largest and smallest group variances (see Tabachnick & Fidell, 1996, p. 413), but this was not the case in this study. Pooled within-group correlation matrices indicated modest inter-correlations among soil variables with no evidence of the problem of multicollinearity. For the variables used in each analysis, the highest correlation was always between Ca and Mg, $r = .73$, and .71 in sections 212D and M212A, respectively, and $r < .70$ on all other instances. In all analyses, the different multivariate test criteria (Wilk's Lambda, Hotelling's trace, Pillai's and Roy's gcr criterion) gave similar results, so only one, Wilk's Lambda, was reported.

(a) Section 212D: (i) macronutrient cations, P and OM_Depth

| Subsection | n size | Estimated | | | Means | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Log_Acid* | *Log_Ca* | *Log_Mg* | *Log_Na* | *Log_P* | *Log_OMD* | *Sqrt_K* |
| 212Da | 162 | 0.822 | 1.881 | 1.022 | 1.285 | 0.946 | 0.167 | 4.723 |
| 212Db | 22 | 0.881 | 0.917 | 0.913 | 1.251 | 0.769 | 0.278 | 4.009 |
| 212Dc | 49 | 0.869 | 1.546 | 0.919 | 1.307 | 0.707 | 0.287 | 4.734 |
| **Total** | **233** | **0.857** | **1.448** | **0.951** | **1.281** | **0.807** | **0.244** | **4.489** |
| ***Sig. univ. F (2, 230) test:*** | | 0.165 | <.001 | 0.372 | 0.449 | 0.018 | 0.034 | 0.121 |
| ***Univ. Wilk's Lambda:*** | | 0.984 | 0.908 | 0.991 | 0.993 | 0.966 | 0.971 | 0.982 |

**Sig. Multivariate F (14, 448) test** < .001

**Multiv. Wilk's Lambda** = .833 $R^2$ (variance explained) = apprx. 17%

(a) Section 212D: (ii) micronutrient cations

| Subsection | n size | Estimated | | | Means | |
|---|---|---|---|---|---|---|
| | | *Exch_Al* | *Log_Cu* | *Log_Fe* | *Log_Mn* | *Log_Zn* |
| 212Da | 225 | 146.82 | -0.251 | 1.893 | 1.092 | -0.07 |
| 212Db | 34 | 190.34 | -0.329 | 1.876 | 0.704 | -0.166 |
| 212Dc | 54 | 188.48 | -0.211 | 1.911 | 0.863 | -0.058 |
| **Total** | **313** | **158.73** | **-0.253** | **1.894** | **1.01** | **-0.079** |
| ***Sig. univ. F (2, 310) test:*** | | 0.011 | 0.015 | 0.814 | 0.001 | 0.133 |
| ***Univ. Wilk's Lambda:*** | | 0.971 | 0.973 | 0.999 | 0.953 | 0.987 |

**Sig. Multivariate F (10, 612) test** < .001

**Multiv. Wilk's Lambda** = .90 $R^2$ (variance explained) = apprx. 10%

**Table 4-6: Results of MANOVA of soil properties among subsections within selected ECOMAP sections in the study area**

## Table 4-6 (cont.)

### (b) Section M212A: (i) macronutrient cations, P and OM_Depth

| | | Estimated | | | Means | | | |
|---|---|---|---|---|---|---|---|---|
| Subsection | n size | *Log_Acid* | *Sqrt_Ca* | *Log_Mg* | *Log_Na* | *Log_P* | *Log_OMD* | *Log_K* |
| M212Ad | 73 | 0.413 | 9.642 | 0.972 | 1.354 | 0.441 | 0.404 | 1.359 |
| M212Ae | 92 | 0.713 | 13.153 | 1.027 | 1.223 | 0.87 | 0.237 | 1.285 |
| M212Af | 77 | 0.796 | 13.802 | 1.107 | 1.156 | 1.009 | 0.173 | 1.258 |
| M212Ag | 33 | 0.822 | 13.189 | 0.909 | 1.089 | 1.249 | 0.241 | 1.068 |
| **Total** | **275** | **0.67** | **12.407** | **1.021** | **1.223** | **0.841** | **0.264** | **1.271** |
| ***Sig. univ. F (3, 271) test:*** | | < .001 | 0.003 | 0.115 | < .001 | < .001 | < .001 | < .001 |
| ***Univ. Wilk's Lambda:*** | | 0.858 | 0.951 | 0.978 | 0.873 | 0.828 | 0.916 | 0.913 |

**Sig. Multivariate F (21, 761) test** < .001
**Multiv. Wilk's Lambda** = .562    **$R^2$** (variance explained) = apprx. 44%

### (b) Section M212A: (ii) micronutrient cations

| | | Estimated | | | Means | |
|---|---|---|---|---|---|---|
| Subsection | n size | *Sqrt_Al* | *Log_Cu* | *Sqrt_Fe* | *Log_Mn* | *Log_Zn* |
| M212Ad | 75 | 14.403 | -0.215 | 10.109 | 0.63 | 0.294 |
| M212Ae | 104 | 12.645 | -0.294 | 9.991 | 0.765 | 0.049 |
| M212Af | 78 | 13.152 | -0.182 | 10.173 | 1.046 | 0.12 |
| M212Ag | 46 | 8.828 | -0.271 | 9.211 | 0.666 | -0.179 |
| **Total** | **303** | **12.631** | **-0.242** | **9.949** | **0.789** | **0.093** |
| ***Sig. univ. F (3, 299) test:*** | | < .001 | 0.04 | 0.509 | < .001 | < .001 |
| ***Univ. Wilk's Lambda:*** | | 0.881 | 0.973 | 0.992 | 0.941 | 0.831 |

**Sig. Multivariate F (15, 814) test** < .001
**Multiv. Wilk's Lambda** = .677    **$R^2$** (variance explained) = apprx. 32%

# Table 4-6 (cont.)

## (c) Section M212B: (i) macronutrient cations, P and OM_Depth

| Subsection | n size | Estimated | | | Means | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Acid* | *Log_Ca* | *Log_Mg* | *Sqrt_Na* | *Log_P* | *Log_OMD* | *Sqrt_K* |
| M212Ba | 147 | n/a | 2.417 | 1.292 | 3.3141 | 1.149 | 0.128 | 4.213 |
| M212Bb | 31 | n/a | 2.131 | 1.106 | 3.857 | 1.053 | 0.091 | 4.631 |
| M212Bc | 102 | n/a | 1.887 | 0.935 | 4.592 | 0.523 | 0.288 | 4.618 |
| M212Bd | 56 | n/a | 1.742 | 0.792 | 4.432 | 0.792 | 0.189 | 4.686 |
| **Total** | **336** | n/a | **2.117** | **1.083** | **3.863** | **0.89** | **0.1833** | **4.453** |
| ***Sig. univ. F (3, 332) test:*** | | n/a | < .001 | < .001 | < .001 | < .001 | < .001 | 0.065 |
| ***Univ. Wilk's Lambda:*** | | n/a | 0.799 | 0.822 | 0.649 | 0.826 | 0.939 | 0.978 |

**Sig. Multivariate F (18, 925) test** < .001

**Multiv. Wilk's Lambda** = .436  $R^2$ (variance explained) = apprx. 56%

## (c) Section M212B: (ii) micronutrient cations

| Subsection | n size | Estimated | | | Means | |
|---|---|---|---|---|---|---|
| | | *Sqrt_Al* | *Sqrt_Cu* | *Log_Fe* | *Log_Mn* | *Sqrt_Zn* |
| M212Ba | 40 | 8.062 | 1.02 | 1.688 | 1.34 | 1.416 |
| M212Bb | 30 | 9.406 | 0.867 | 1.687 | 1.429 | 1.183 |
| M212Bc | 104 | 12.619 | 0.631 | 1.819 | 0.797 | 1.315 |
| M212Bd | 57 | 10.516 | 0.559 | 1.751 | 0.932 | 1.12 |
| **Total** | **231** | **10.894** | **0.711** | **1.762** | **1.006** | **1.267** |
| ***Sig. univ. F (3, 227) test:*** | | < .001 | < .001 | 0.063 | < .001 | 0.013 |
| ***Univ. Wilk's Lambda:*** | | 0.883 | 0.661 | 0.968 | 0.856 | 0.954 |

**Sig. Multivariate F (15, 616) test** < .001

**Multiv. Wilk's Lambda** = .532  $R^2$ (variance explained) = apprx. 47%

# Table 4-6 (cont.)

## (d) Section M212C: (i) macronutrient cations, P and OM_Depth

| Subsection | n size | Acid | Log_Ca | Log_Mg | Log_Na | Log_P | Log_OMD | Log_K |
|---|---|---|---|---|---|---|---|---|
| | | Estimated Means | | | | | | |
| M212Ca | 109 | n/a | 2.183 | 1.064 | 1.001 | 1.012 | 0.247 | 1.183 |
| M212Cb | 27 | n/a | 2.495 | 1.354 | 1.108 | 1.036 | -0.002 | 1.243 |
| M212Cc | 51 | n/a | 2.107 | 1.101 | 1.19 | 0.9 | 0.141 | 1.333 |
| M212Cd | 58 | n/a | 2.079 | 1.099 | 1.124 | 0.797 | 0.125 | 1.283 |
| **Total** | **245** | **n/a** | **2.177** | **1.112** | **1.084** | **0.941** | **0.169** | **1.245** |
| ***Sig. univ. F (3, 241) test:*** | | n/a | < .001 | 0.015 | < .001 | 0.062 | < .001 | 0.003 |
| ***Univ. Wilk's Lambda:*** | | n/a | 0.937 | 0.958 | 0.879 | 0.971 | 0.911 | 0.945 |

**Sig. Multivariate F (18, 667) test** < .001

**Multiv. Wilk's Lambda** = .702 $R^2$ (variance explained) = apprx. 30%

## (d) Section M212C: (ii) micronutrient cations

| Subsection | n size | Sqrt_Al | Sqrt_Cu | Sqrt_Fe | Sqrt_Mn | Log_Zn |
|---|---|---|---|---|---|---|
| | | Estimated Means | | | | |
| M212Ca | 106 | 13.361 | n/a | 9.97 | 4.484 | 0.307 |
| M212Cb | 30 | 10.805 | n/a | 8.159 | 6.658 | 0.321 |
| M212Cc | 49 | 12.555 | n/a | 10.002 | 4.909 | 0.283 |
| M212Cd | 51 | 14.594 | n/a | 9.194 | 2.686 | 0.216 |
| **Total** | **236** | **13.135** | **n/a** | **9.579** | **4.46** | **0.284** |
| ***Sig. univ. F (3, 232) test:*** | | 0.039 | n/a | 0.063 | < .001 | 0.159 |
| ***Univ. Wilk's Lambda:*** | | 0.965 | n/a | 0.969 | 0.845 | 0.978 |

**Sig. Multivariate F (12, 606) test** < .001

**Multiv. Wilk's Lambda** = .805 $R^2$ (variance explained) = apprx. 20%

## Table 4-6 (cont.)

### (e) Section 221A: (i) macronutrient cations, P and OM_Depth

| Subsection | n size | Estimated | | | Means | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Sqrt_Acid* | *Log_Ca* | *Exch_Mg* | *Log_Na* | *Log_P* | *Log_OMD* | *Sqrt_K* |
| 221Ai | 84 | 1.909 | 1.413 | 6.943 | 1.131 | 0.882 | 0.127 | 4.114 |
| 221Ak | 20 | 1.804 | 1.633 | 21.65 | 1.183 | 0.154 | 0.203 | 4.809 |
| 221Al | 108 | 2.825 | 1.68 | 7.064 | 1.111 | 0.91 | 0.211 | 4.139 |
| **Total** | **212** | **2.365** | **1.57** | **8.392** | **1.126** | **0.922** | **0.177** | **4.192** |
| ***Sig. univ. F (2, 209) test:*** | | <.001 | <.001 | 0.002 | 0.336 | 0.256 | 0.139 | 0.101 |
| ***Univ. Wilk's Lambda:*** | | 0.886 | 0.943 | 0.893 | 0.989 | 0.987 | 0.981 | 0.978 |

**Sig. Multivariate F (14, 406) test** < .001
**Multiv. Wilk's Lambda** = .687 $R^2$ (variance explained) = apprx. 31%

### (e) Section 221A: (ii) micronutrient cations

| Subsection | n size | Estimated | | | Means | |
|---|---|---|---|---|---|---|
| | | *Log_Al* | *Log_Cu* | *Log_Fe* | *Log_Mn* | *Sqrt_Zn* |
| 221Ai | 94 | 1.857 | -0.464 | 1.69 | 0.568 | 0.952 |
| 221Ak | 26 | 1.93 | -0.426 | 1.805 | 0.836 | 0.84 |
| 221Al | 126 | 1.86 | -0.327 | 1.717 | 0.548 | 1.113 |
| **Total** | **246** | **1.866** | **-0.389** | **1.716** | **0.586** | **1.023** |
| ***Sig. univ. F (2, 243) test:*** | | 0.686 | < .001 | 0.079 | 0.055 | 0.002 |
| ***Univ. Wilk's Lambda:*** | | 0.997 | 0.934 | 0.979 | 0.976 | 0.948 |

**Sig. Multivariate F (10, 478) test** < .001
**Multiv. Wilk's Lambda** = .826 $R^2$ (variance explained) = apprx. 17%

As shown in Table 4-6, differences among subsections were analyzed for ECOMAP sections 212D, M212A, M212B, M212C, and 221A. The soil variables used, the number of observations, and the subsection and section means of these variables arereported in Table 4-6. Note that the means were often on transform scales (log = logarithmic, and sqrt = square root). The p values for both the univariate and multivariate F-tests are also given. The p values that are less than the traditional .05 may be considered significant. Univariate significance means that, at least, one of the subsections was significantly different from the others with respect to the mean concentration of that particular soil variable. Multivariate significance means that the vector of means of the soil variables in one or more subsections are significantly different from other subsections (i.e., the subsections are different on some weighted linear composite of the soil variables). Wilk's Lambda ($\lambda$) is the most commonly reported test statistics in multivariate analyses, and tells the proportion of variances in the dependent variables (soil properties) <u>not</u> explained by the independent variable (subsections). Effect size or the percent variance explained (like $R^2$ in multiple regression), therefore, is computed as 1- $\lambda$. Wilk's $\lambda$'s are given for both the univariate and multivariate tests of significant differences.

Table 4-6 shows that except for section 212D, the subsections within other ECOMAP sections were significantly different from one another both with respect to individual soil properties, and when all soil properties are taken together. However, it seems that spatial variability in soil properties were best captured by the subsectional delineation in ECOMAP sections M212B > M212A > M212C. In these sections, the number of soil properties on which the subsections differ was more, the univariate significance of these variables were higher (p was almost always less than .001), and the percent explained variance ($1-\lambda$) in many of the variables was relatively high (up to 15% and higher), compared to the others. Section 212D did poorly---the pércent variance explained was mostly less than 5% except for Ca which had 10%. No one soil variable was consistently well delineated from section to section, by subsectional map units. There were occasional highs in percent explained variances in M212B which showed 35%, 34% and 20% for Na, Cu and Ca, respectively.

Six ECOMAP sections: 212D, M212A, M212B, M212C, 221A, and 212E had sufficient data to allow a similar analysis of the differences in soil properties among ECOMAP sections. This analysis was done for soil attributes that are important in most soil mapping, namely, texture of the B-horizon, depth to bedrock, soil drainage, elevation, parent material type, depth to organic matter, and rooting depth. The results showed that the ECOMAP sections were highly significant different from one another with respect to each of these variables, $p < .001$. The results also revealed that ECOMAP section map units could explain 53% of the variation in elevation in the study area, about 9% of the variances in each of drainage and depth to organic matter, and 5% or less for all others. The multivariate F-test was also highly significant, $p < .001$, and showed that about 63% ( or $\lambda = .368$) of the variation in the composite of these soil variables in the study area could be explained by ECOMAP sections. The analysis was repeated with only soil chemical properties, namely, Al, Ca, Fe, K, Mg, Mn, Na, P, and Zn. Again, the univariate $F(3, 1311)$ was highly significant, $p < .001$ for all variables, except K and P whch had p = about .04.

Percent variance in the study area that was explain for the individual soil properties was about 16% for Zn (the highest), about 10%for each of Al, Ca, Fe, Mg, Mn, and Na, and lowest for P (< 1%). The multivariate $F(50,5941)$-test was highly significant, and 50% ($\lambda$ = .507) of the variance of the chemical properties in the study area was predictable from knowledge of ECOMAP sections.

## 4.4 Summary and Conclusion

The initial objective of this study was to provide variability statistics of several chemical soil properties measured during the 1983 FIA of the states of Maine, New Hampshire, and Vermont. These statistics, and other information provided in this study will be useful for resource management on a regional scale, and may provide important aids to future studies requiring field sampling and/or analyses of geographic variation of soil properties in the study area. To increase precision of estimates and sampling efficiency of the selected soil properties, the study area was stratified by ECOMAP sections.

Although, soil properties exhibited a clear case of heteroscedasticity, Table 4-7 shows that about 80% of the time, the CV's (variability) of soil properties within ECOMAP sections were lower than those for the entire study area. The relatively low within-section variability may also explain why over 90% of the time, the available FIA soil sample sizes in these sections were greater than the sample sizes required to efficiently estimate the means of the selected chemical properties, at the 90% confidence level and given 20% or more margins of error. Perhaps not surprisingly, the optimum required sample sizes were very highly correlated with CV's, $r = .96$ or higher.

| | Study Area | ECOMAP Sections in the Study Area | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 212D | 212E | M212A | M212B | M212C | 221A | 221B |
| **Soil Props.** | | **% Coefficient of Variation** | | | | | | |
| **Acid** | 75.48 | 45.65 | n/a | 61.31 | 139.19* | n/a | 102.74* | n/a |
| **Al** | 85.57 | 71.19 | 108.49* | 73.14 | 12.84 | 78.17 | 86.27 | 124.92* |
| **Ca** | 149.64 | 159.38* | 98.65 | 129.05 | 139.61 | 131.28 | 174.89* | 93.86 |
| **CEC** | 47.09 | 8.20 | 34.63 | 45.17 | 34.76 | 29.15 | 38.55 | n/a |
| **Cu** | 79.34 | 56.18 | 45.28 | 84.84* | 77.09 | 51.76 | 96.73* | 58.09 |
| **Fe** | 80.22 | 69.87 | 74.36 | 76.53 | 81.01 | 77.56 | 62.68 | 83.64* |
| **K** | 67.14 | 68.00 | 69.72 | 69.68 | 65.52 | 63.10 | 67.90 | 41.63 |
| **Mg** | 145.24 | 163.09* | 124.96 | 136.92 | 125.01 | 128.93 | 151.11* | 97.65 |
| **Mn** | 186.31 | 189.35 | 106.88 | 197.47* | 176.23 | 166.60 | 191.17* | 136.24 |
| **Na** | 58.69 | 43.33 | 62.86* | 65.22* | 61.51* | 61.25* | 47.77 | 54.04 |
| **SOM** | 72.51 | 70.23 | 54.78 | 64.92 | 69.05 | 45.57 | 73.54 | n/a |
| **OM_dept** | 108.05 | 122.17* | 135.25* | 95.71 | 124.96* | 88.47 | 108.30 | 93.87 |
| **P** | 140.27 | 146.19* | 104.73 | 135.94 | 148.21* | 143.70 | 132.68 | 30.38 |
| **total N** | 64.73 | 55.31 | n/a | 63.53 | 57.99 | n/a | 56.98 | n/a |
| **Zn** | 88.76 | 79.21 | 65.84 | 96.32* | 80.98 | 65.13 | 94.26* | 60.89 |

**Table 4-7: Comparison of variation (CV) of specific soil properties in ECOMAP sections, and also the entire study area.** * CV of soil property in these sections were greater than that of the study area.

Generally speaking, Ca, Mg, and Mn, P and OM_depth were most variable among the soil properties examined in this study. Soil pH, CEC, and total N were the least variable, while Na, K, SOM and Fe were intermediate between this and the first group of variables. Optimum sampling intensity is a function of the degree of variability, hence the relative sample sizes required to estimate the means of these soil properties, would more or less follow the same pattern. Some ECOMAP sections showed notable trends in either the mean concentrations or degree of variability of soil properties (data not shown ). Section 212E had the highest means for SOM, CEC, Ca, K, Mg, P, Cu, and Mn, the maximum CV's for Al and organic matter depth, and the lowest CV's for Cu, Ca, and Mn. On the other hand, section 221A had the lowest or next to the lowest means for

SOM, CEC, Ca, Mg, K, Fe, Al, Cu, total_N, Mn and pH, and the maximum variability for Acid, Cu, SOM, Na, Zn, Mn, and Mg. Further investigations and/or explanation of why these and other observed trends hold in the ECOMAP sections are certainly appropriate, but could not be accommodated in this study. The large data sets (large number of samples and soil attributes) available for this study, made it possible for some conclusions about the distribution characteristics of natural soil populations to be empirically evaluated. This study confirmed the reports in other studies that soil variables are rarely normally distributed, but they are not always lognormal either.

Another major objective of the study was to test the hypotheses that ECOMAP sections in the study area, and subsections of a given section, are not significantly different from one another, and/or to evaluate the nature of the differences among ECOMAP ecological map units, in terms of specific soil properties. Table 4-8 provides a synoptic view and summary of MANOVA tests of differences among subsections within selected ECOMAP sections in the study area. The univariate F-tests of differences among subsections were highly significant for most soil variables in the sections. The multivariate F-tests were all highly significant, $p < .001$. The proportion of explained variance in the composite of soil variables within ECOMAP sections, ranged from 10% in 212D to 56% in M212B. The percent explained within-section variances of individual soil properties were also evaluated (Table 4-6), but these were always much smaller than the multivariate counterpart. This demonstrates one of the superiorities of multivariate analysis to multiple univariate tests in soils and other studies involving many variables with reasonably high inter- variable correlation. It is not uncommon for univariate tests of soil variables that were all non-significant to have significant multivariate effect. No one soil variable was consistently well delineated by the subsection map units. The notably high percent explained variances occurred in section M212B where 35%, 34% and 20% of the variances in Na, Cu and Ca, respectively, were captured by the subsectional delineations.

| SOIL VARIABLES | ECOMAP SECTIONS 212D$^{ss=3}$ | M212A | M212B | M212C | 221A$^{ss=3}$ |
|---|---|---|---|---|---|
| Exch_Acid | 0.165 | <.001*** | n/a | n/a | <.001*** |
| Exch_Al | .011** | <.001*** | <.001*** | .039** | 0.686 |
| Exch_Cu | .015** | .04** | <.001*** | n/a | <.001*** |
| Exch_Ca | <.001*** | <.001*** | <.001*** | <.001*** | <.001*** |
| Exch_Fe | 0.818 | 0.509 | .063* | 0.063* | .079* |
| Exch_K | 0.121 | <.001*** | .065* | <.001*** | 0.101 |
| Exch_Na | 0.449 | <.001*** | <.001*** | <.001*** | 0.336 |
| Exch_Zn | 0.133 | <.001*** | .013** | 0.159 | <.001*** |
| Exch_Mg | 0.372 | 0.115 | <.001*** | .015** | <.001*** |
| Exch_Mn | <.001*** | <.001*** | <.001*** | <.001*** | .055* |
| OM_dept | .034** | <.001*** | <.001*** | <.001*** | 0.139 |
| Exch_P | .018** | <.001*** | <.001*** | .062* | 0.256 |
| Multivariate Signif. | <.001*** | <.001*** | <.001*** | <.001*** | <.001*** |
| Explained Variance | 10 - 17% | 32 - 44% | 47 - 56% | 20 - 30% | 17 - 31% |

**Table 4-8: p values of tests of among-subsection differences in soil attributes, within selected ECOMAP sections in the study area.** ss = # of subsections; when not indicated = 4 (out of 7 for M212A) ; n/a = not used in the MANOVA study.

Six out of the essentially nine sections in the study area had enough data to allow similar analysis of the differences in soil attributes among these sections. All univariate, and multivariate tests were highly significant, with K and P at the bottom of the list. The percent univariate and multivariate variances in the study area explained by ECOMAP sections were much higher than those explained by the subsections in any ECOMAP section. Elevation topped the list with 53% of its variation in the study area explained. Others include Zinc with 16%, about 10% for each of Al, Ca, Fe, Mg, Mn, Na, drainage and depth to organic matter. The percent multivariate variance in the soil variables in the study area that was accounted for or explainable by the ECOMAP section map units was as high as 63% ($\lambda$ = .368) for a group of field variables, and 50% for chemical properties. Although ECOMAP was not primarily a soil-mapping endeavor, this study provides some assessment of the effectiveness of the ecological land classification system in delineating targeted soil-site conditions in this forested study region, and perhaps, similar areas.

# CHAPTER 5

# THE LEGITIMACY OF VARIABILITY STATISTICS COMPUTED FROM NON-NORMAL SOIL DATA

## 5.1 Introduction

### 5.1.1 Statement of Problem

Preliminary analysis of data, and a complete characterization of the population of a data set often include the examination of the frequency distributions (see Warrick & Nielsen, 1980). The normal distribution is a requirement of most parametric statistical analyses and procedures, and frequency distribution is usually examined to see if this requirement is met. When normality or other assumptions of parametric tests are violated, the validity of test statistics may be questionable, and severe violations can lead to spurious conclusions (see Thoni, 1967; Wilding & Drees, 1983; Zar, 1996, p.278).

One of the consistent conclusions from soil variability studies is that the frequency distributions of soil attributes are rarely normal or symmetrically bell-shaped about the mean. Instead, the distributions of natural populations of many soil properties have been found to be positively skewed, often in a lognormal fashion (Wilding & Drees, 1983, p. 91; Parkin et al., 1988; Grigal et al., 1990; Webster & Oliver, 1990, p.24; Parkin & Robinson, 1992). The frequency distribution of a positively skewed data set is both asymmetrical, and has high and infrequent scores to the right of the data values that are considered typical or representative scores (i.e., the mean, median, or mode) of the distribution. If the logarithm of scores of a positively skewed distribution yields normally distributed scores, the initial data set is said to be lognormal (see Gaddum, 1945). Non-linear transformation is often used to correct non-normality, i.e., to make a skewed data set more normally distributed and better suited for parametric analyses. Transformation changes the scores of the original variable, *Y*, into a new variable, *Z*, by a single-valued monotone function, $X = T(Y)$, such that Z may fulfil one or more of the basic requirements of parametric statistics (Thoni, 1967) namely, normality, homoscedasticity and linearity.

Transformation is a subject of many statistical texts and much research literature including those cited in this study. Clear answers exist in these sources about the use of nonlinear transformations for standard statistical analyses (e.g., analysis of variance) used in soil studies. However, much confusion still exists about whether or not to use transformations prior to the computation of traditional variability statistics, namely, the mean, standard deviation, coefficient of variation, confidence intervals of the mean, and optimum sample sizes for the estimation of the mean. A review of existing literature reveals conflicting recommendations, and the publication of studies advocating incongruous approaches to the problem of estimating these statistics from non-normally distributed soil data. The more common practice in soil variability studies has been not to use transformation but to assume normal distributions (see Beckett & Webster, 1971; Wilding & Drees, 1983, p. 90). Wilding & Drees observed that the normal distribution is hastily assumed in these soil studies because this eases calculation, analysis and interpretation of results. Webster & Oliver (1990) also noted that results from transformations are not as readily understood as are those from data that do not need transformation. However, there have been a number of published research papers that decried the observed disinclination to the use of transformation, and questioned the validity of variability statistics computed from non-normally distributed soil data.

Wilding & Drees (1983) stated that results of statistical analyses may be misleading if they are based on invalid assumptions of normality. In a study involving pore water infiltration velocity, Warrick & Nielsen (1980) showed that the mean (21.52 cm/day) of a data set of 20 samples computed with assumption of logarithmic distribution, was higher than the arithmetic mean (19.19 cm/day) calculated by assuming normal distribution. Grigal et al. (1990) showed that a smaller mean, smaller width for a given confidence interval, and much smaller sample size requirement at a given margin of error and confidence, were computed when lognormal transformation was used, compared to when normality was assumed for soil calcium. They also noted that the discrepancy between statistics from transformed and untransformed soil pH data was smaller because the original data set was not grossly non-normal. The ostensible conclusion from these and other studies (Warrick & Nielsen, 1980; Parkins et al., 1988; and Grigal et al., 1990) is that, not recognizing the actual skewed distribution of a data set leads to invalid variability statistics, and spurious conclusions. Most young soil scientists would find the above situation confusing, and many older colleagues might be abashed at the apparent contradictions in our research literature. Parkin et al. (1988) observed that there is a knowledge and communication gap between statistics and soil science, and that some of the work done in the statistical sciences simply do not find its way into soil science.

### 5.1.2 Objectives and Justification

The Large data and several soil properties analyzed in Chapter 4 confirmed the pedological truism that soil data are almost always non-normally distributed. The goals of this study were based posteriorly on this and other findings from Chapter 4. The major objectives in this study were to 1) review some of the findings in Chapter 4 about the distribution characteristics of natural soil populations; 2) review pertinent statistical and soil science literature in order to answer questions about whether or not to transform non-normally distributed soil data prior to the computation of traditional variability statistics; 3) empirically determine the relative impact of kurtosis (peakedness) and skewness (asymmetry) on the failure of soil properties to pass normality tests; and 4) evaluate the use of the coefficients of variation (CV) as a "semi-quantitative" index of non-normality in soil variables.

Non-normality is caused by skewness or kurtosis or both. Severe skewness and kurtosis each affects variability statistics in a unique and predictable fashion. If the relative impact of skewness and kurtosis on non-normality of soil data can be determined, this information can be used to better understand and more validly interpret variability statistics computed from soil data with skewness and/or kurtosis that are significantly greater than zero. In other words, knowledge about the nature and degree of the non-normality is needed in order to correctly interpret variability statistics computed from non-normally distributed and untransformed soil data. The qualitative (graphical) methods and commonly used quantitative indices of non-normality have major limitations and are impractical for use in soils. On the other hand, there are reasons to think that CV---an easy to compute and already familiar statistics to soil scientists, is also a measure of non-normality in data. If CV is highly and reliably correlated with skewness or kurtosis or both, this would mean that CV can be used to provide more practical and easier-to-interpret information about non-normality in soil data than the presently used quantitative indices of the departure of a distribution from the normal.

## 5.2 Materials and Methods

The frequency histograms and other distributive characteristics (e.g., skewness, kurtosis, and CV) of several soil variables were evaluated (Chapter 4) to determine the nature of the distributions of these soil variables. Kolgomorov-Smirnov test of normality was then used to ascertain the extent of the departure of distributions of the soil data from normality. The Kolgomorov-Smirnov test yields a KS factor which is based on the largest absolute difference between the distributions (i.e., the cumulative density functions) being compared, in this case, the soil variable and a hypothetical normal distribution. A large KS factor implies that the distribution being evaluated is significantly different from the normal. The distributive statistics and KS factors were determined before and after non-linear transformations were used on the soil variables, and also before and after the removal of outlying values from the soil data. These investigations were carried out at different spatial scales in the study area, and using more than one plausible alternative transformation. These studies were performed in Chapter 4, but the pertinent results from them are reviewed shortly.

Extensive correlation analyses were used to evaluate the nature and strength of association between each pair of KS factor, CV, and measures of skewness and kurtosis.

Statistical analysis was also used to see if the variability statistics computed after the use of alternative transformations are significantly different from each other and from those of untransformed soil data. Pertinent literature in applied statistics and soil science were reviewed to 1) show that it is desirable but not necessary to achieve normality in data to validly compute traditional variability statistics; 2) show major limitations in the use non-linear transformations on soil data prior to the computation of variability statistics; 3) review how peakedness and asymmetry differentially impact variability statistics, and the limitations of the qualitative and commonly used quantitative tests of normality; and 4) to provide practical guide on how the CV and other distributive characteristics of soil can be used to adequately handle the problem of non-normality in soil variability studies.

## 5.3 Results and Discussion

### 5.3.1 Distribution Characteristics of Natural Soil Populations

Most of the soil properties examined in Chapter 4 were positively skewed, and none of them passed the Kolmogorov-Smirnov test of normality ($p < .001$). Based on that and other studies (e.g., Grigal et al., 1990), soil pH may be the most normally distributed soil variable, especially in forested environment. This is apparently due to the fact that pH (log of $H^+$ concentration) is already a transformed variable. These studies also showed that soil variables have diverse distributions which may be a function of the size and heterogeneity of the study area, the sample size being analyzed, and perhaps, the pedological factors and functions that are prevalent in the study site.

Most of the soil variables examined in these studies still could not pass quantitative tests of normality even after the use of non-linear transformation (see Section 4.3.1), and some of the soil variables (e.g., Acid, Al and SOM) were less responsive to logarithmic transformation than to the square root transformation. Some of the conclusions derivable from these and other studies are that soil variables are rarely normally distributed, and they may not conform to lognormal distribution as frequently as published literature would lead one to expect. This means that uncritical assumption of either the normal or lognormal distribution in analyses where knowledge of the distribution of a soil variable is important, may lead to spurious results. This fact underscores the need for the soil or environmental scientist to evaluate actual histograms, and the effect of alternative transformations if the objective is to achieve normality and/or variance stabilization.

### 5.3.2 Major Problems With the Use of Non-linear Transformations

Obviously, the assumption of a normal distribution in parametric procedures is the theoretical basis for recommending the use of non-linear transformation in variability studies. Parametric statistics are techniques for estimating population parameters, and for testing hypotheses and making inferences about features of a population, from a sample set. Most of the statistical procedures undertaken in a standard soil variability study are parametric. However, the assumption of normality applies differently to different parametric statistics. In analyses that are based on the Pearson's product-moment correlation such as the different forms of regression, the assumption of normality is critical, and applies to the distribution of the observed scores themselves or to the residuals of the analyses (see Tabachnick & Fidell, 1996, p. 70). But in analyses such as the Z test, t tests, and the different forms of ANOVA where the hypotheses are tested about the means, the distribution representing hypothetical states of reality are the distributions of means rather than distribution of individual scores (Tabachnick & Fidell, 1996, p. 34). The central limit theorem guarantees that the sampling distributions of means differ systematically from the distribution of individual scores, and are normal irrespective of the distribution of the sampled population. The subject of our inquiry in variability studies are the distributions of the means of natural soil populations, not necessarily of the samples at hand, and traditional variability studies usually do not even involve formal hypotheses testing. Hence, the requirement of normality, though desirable, is not critically important nor necessary (see Zar, 1996, p. 325) in order to validly estimate most variability statistics.

Chapter 4 and other studies (e.g., Grigal et al., 1990; Cambardella et al., 1994) have shown that often normality can not be achieved even after non-linear transformation of soil data. Even if normality is achieved, the use of transformation for variability studies is fraught with serious and hard-to-deal with problems, and practically prohibits the comparison of studies done at different times and places. Webster & Oliver (1990) noted that the results from transformations are not as readily understood as are those from data that do not need transformation. In addition, variability statistics computed from transformed data are said to be biased, and therefore require to be further corrected in a certain fashion (see Thoni, 1967, p. 16; Parkin et al., 1988; Zar, 1996, p.281).

Chapter 4 (section 4.3.1) discussed the use of a family of power transformation for normalization among which the logarithmic and square root transformations are the most commonly used members. Perhaps, the most serious problem with transforming soil data in variability studies is the vagaries associated with the choice of transformation from the suite of commonly used members of the family of power transformation discussed in Chapter 4 (section 4.3.1), the magnitude of differences among statistics from alternative transformations, and between these and those of the untransformed data. Below are actual data, computed from the 511 available soil Ca records in the state of Maine. Exch_Ca is in original unit while log_Ca and sqrt_Ca are back-transforms of the logarithmic and square root transformations, respectively, of the same data. These data represent the rather common situation where there is more than one plausible transformation for a data set, and no clear choice between the alternative options of transformation. Notice the chasm between corresponding back-transformed statistics from the logarithmic and square root

| | Exch_Ca | Log_Ca | Sqrt_Ca |
|---|---|---|---|
| **Mean** | 214.85 | 49.55 | 130.41 |
| **SD** | 317.84 | 10.93 | 84.64 |
| **90% CI mean** | 191.68 - 238.02 | 41.62 - 58.99 | 115.50 - 146.12 |
| **CV (%)** | 147.93 | 61.28 | 80.57 |
| **KS Z** | 5.64(p. < .001) | 3.04(p. < .001) | 2.57(p. < .001) |

transformation of the data. The means were different by an order of almost 300%, and the standard deviations (SD), about 800%. Without a universal agreement to always transform a particular soil variable in a specified way (as the soil science community has successfully and consistently done for pH), sporadic use of transformation in soil variability studies would prohibit the comparison of results among different studies.

### 5.3.3 The Use of CV's as an Index of Non-normality in Soil Variables

The conclusion from the preceding discussion is that transformation is unnecessary or even inadvisable in variability analyses of soil data. However, the extant problem is that the standard deviation and standard error of a skewed data set are not symmetrical about the mean. Therefore, probability statements about an observation (e.g., that there is 68% chance that an observed value will be within the mean $\pm$ one SD), and confidence intervals of the mean, are suspect and must be interpreted with caution. A more valid interpretation of these and other variability statistics would require some knowledge of the degree of non-normality in the (untransformed) soil data set.

While the qualitative ("it looks good") test of normality is too vague and subjective, most quantitative tests are either too sensitive (i.e., they always reject the null hypothesis with slightest departure from normality) especially with a reasonable sample size (see Tabachick & Fidell, 1996, p. 73), or they have major limitations and are cumbersome to apply (D'Agostino, 1986; Zar, 1996, p. 89). On the other hand, Beckett & Webster (1971) in their classic review paper on soil variability had indicated that CV values greater than 100% may be symptomatic of skew distributions, implying that CV may also be an expression of the departure of soil data from normality. Extensive statistical analyses of large data sets and several soil variables were done in this study to see if CV and KS Z (from Kolmogorov-Smirnov test of normality) are multicollinear, and therefore, provide equivalent information. The Kolmogorov-Smirnov test is a quantitative test commonly used to evaluate if a distribution is significantly different from the normal. The KS Z is based on the largest absolute difference between a distribution and hypothetical normal distribution, and KS values are indices of relative degree of non-normality.

Non-normality is caused by skewness or kurtosis or both. Cambardella et al. (1994) had observed that some soil properties could not pass normality tests after log transformation, more because of a failure to reduce kurtosis than a failure to reduce skewness. Knowledge of the differential effect of skewness and kurtosis on the non-normality of soil data has significant implication (as discussed below) in the quest for a practical solution to the problems of non-normality in soil variability studies. This study, therefore, also investigated the interrelations among KS factor, kurtosis and skewness, in addition to evaluating if CV is reliably correlated with KS factor enough to serve as an alternative index of non-normality.

Table 5-1 shows Pearson's correlation coefficients between pairs of CV, KS, kurtosis and skewness, based on the data in Table 4-3 (Chapter 4). The analysis was first done with all 32 observations, and then repeated for the before transformation, and after transformation subsets of data, to see if results were consistent. Because of the similarity between the first two results (Table 5-1), only one will be discussed, and where this is different from the third analysis (of transformed data) will be highlighted. This table shows that KS was highly correlated with each of skewness and kurtosis (as expected), but had stronger association, $r(df=32) = .86$ or $r^2 = .74$, with skewness than with kurtosis, $r = .70$ or $r^2 = .49$. The $r^2$ (coefficient of determination) is equivalent to the percent variation in KS that is explained or accounted for by skewness or kurtosis. The table also shows a very strong correlation between kurtosis and skewness (r as high as .96), implying that their individual correlation with KS may be an artifact of their relationship with each other. Appropriate multiple regression procedures were then performed to assess the percent variance in KS that is uniquely attributable to each of skewness and kurtosis, and to the two combined.

**(a): All data in Table 4-3 were used, so n =32**

| | KS | Kurtosis | Skewness |
|---|---|---|---|
| CV | 0.876 | 0.825 | 0.915 |
| KS | | 0.701 | 0.856 |
| Kurtosis | | | 0.944 |

**(b): With untransformed data (Table 4-3), n=16**

| | KS | Kurtosis | Skewness |
|---|---|---|---|
| CV | 0.883 | 0.83 | 0.905 |
| KS | | 0.642* | 0.786 |
| Kurtosis | | | 0.963 |

**(c): With transformed data (Table 4-3) only, n=16**

| | KS | Kurtosis | Skewness |
|---|---|---|---|
| CV | 0.196~~ns~~ | 0.15~~ns~~ | 0.33~~ns~~ |
| KS | | 0.641* | 0.811 |
| Kurtosis | | | 0.785 |

**Table 5-1: Bivariate correlation coefficients among distribution characteristics of soil properties.** $p < 001$ when not indicated; $= .006$ for *; and $>.10$ for ns.

Semi partial correlation (sr) computed from the regression analysis shows the strength of association between the dependent variable, DV (e.g., KS) and each of the independent variables, IV (e.g., skewness and kurtosis) after statistically controlling the effect of the other. The multiple regression $R^2$ is the variance in DV that is explained by all the IV's combined, and can be partitioned into percentages that are uniquely accounted for by each IV ($sr^2$), and that which is shared or common to them.

The multiple regression based on all 32 observations in Table 4-3 had R = .92 and $R^2$ = .85 (adjusted $R^2$ = .84) which was significantly different from zero, p < .001. The squared semi partials ($sr^2$) for skewness and kurtosis were .37 and .12, respectively. Hence based on this analysis, about 85% of the variance in KS was predictable from measures of skewness and kurtosis together. Of this 85%, 37% was uniquely attributed to skewness alone, 12% to kurtosis, and about 36% was shared variance between the two. The above analysis was repeated with only the untransformed, and then the transformed data in Table 4-3. The result in the former was almost identical with the one described above. With transformed data set, $R^2$ went down to .66 (adjusted $R^2$ = .60), and $sr^2$ for skewness and kurtosis were .25 and .007 respectively, while share variance went up to about 40%.

These analyses have shown, rather consistently, that KS (index of departure from normality) is more responsive, or affected more by changes in skewness than in kurtosis, and that the discrepancy in the response is further accentuated after the use of non-linear transformation, and/or in data that approach normality. Except for when data were transformed, Table 5-1 also shows $r \geq .88$ ($r^2 \geq .77$) between CV and KS. To validate the this result, Pearson's correlation between CV and KS was also computed using the 32 sample data on Table 4-4 (Chapter 4). Again, the r(32) = .85 or $r^2 \geq .72$, confirming that on the average, CV explains about 75% of the same information as KS or kurtosis (see Table 5-1), and often higher percentage (r as high as .92 ) of the information in skewness.

### 5.3.4 Practical Guide to Handling the Problem of Non-normality in Soil Data

To recap, there is rarely the question of whether or not soil data are non-normally distributed. The preceding sections showed that it is unnecessary and probably inadvisable to transform soil data for the purpose of computing traditional variability statistics. But it also showed that knowledge of the degree of non-normality in the untransformed soil data is required for a more valid interpretation of variability statistics computed from them. The limitations of presently used qualitative and quantitative indices of non-normality were highlighted. Thus, the extant and pandemic problem is how to quantitatively express the degree of non-normality in soil data, when it is clearly inexpedient to use non-linear transformation. It was then shown that CV is highly and significantly correlated with KS factor, and thus can be reliably used as an index of non-normality. Some of the advantages of this proposition include the facts that CV (i.e., $SD^2$/mean, expressed in percentage) is readily computed from summary statistics, and it is already a familiar statistic to soil scientists. If used as a "semi-quantitative" index, CV would provide a more practical and easier-to-interpret measure of non-normality in soil data than the KS factor or other commonly used quantitative indices of the departure of a distribution from the normal. In this section, practical guidelines or "rules of thumb" are provided on how CV and other distributive characteristics of soil data can be used to adequately handle the problem of non-normality in soil variability studies.

The following suggested rules of thumb in interpreting CV values of soil variables are based on the examination of the frequency histograms (Figure 4-2), and tables of the distribution characteristics of several soil variables (Tables 4-3, 4-4a, and 4-4b) shown in Chapter 4. It seems that CV values less than 50% shows soil variables that are exceptionally (by "soils' standard") close to normal. CV values between 50 and 70% can be regarded as satisfactory. Soil variables with CV values above 70% but less than 100% are skewed with probable presence of outliers. CV's above 100% is definitely skewed and/or kurtosis that is, perhaps, severely non-zero. Recall that it was shown that the differential influence of skewness on non-normality is greater than that of kurtosis, but also, that CV differentially captures more of the variance in skewness than in kurtosis. However, knowledge of the differential impacts of severe kurtosis or skewness on variability statistics, and how to mitigate each of these separately were also investigated, and presented below.

When a data set has non-zero kurtosis, the variance is underestimated especially if sample size is small (see Tabachnick & Fidell, 1996, p. 73). One solution to the use of soil data with non-zero kurtosis in variability studies is to increase the number of samples used to compute statistics. If sample size can not be increased, awareness that the expected value of the variance may be higher could be used to subjectively interpret the statistical results. For instance, a higher variance would imply that the actual range of confidence interval of the mean may be wider, and sample size requirement at a given confidence and marginal error would be higher, than those computed from the untransformed dataset with either a positive or negative kurtosis.

Skewed soil data can be remedied by the exclusion of outliers. Using the data in Table 4-4a (Chapter 4), one-tail Paired t-test ($df = 15$, $t = 3.35$) showed a highly significant difference, $p = .002$, between both the CV's and KS' of soil variables before and after the removal of outliers. Based on that study, the CV's of soil variables were on the average 28% lower after outliers were removed. Note that most soil studies do not have the luxury of the very large sample sizes used in Table 4-4a. With smaller sample sizes, the disproportionate influence of outliers on variability statistics would become more pronounced, and the mean CV difference after the removal of outliers will, expectedly, be higher.

I would recommend that the cutoff point for outliers in soil data be the mean plus four standard deviations (SD). However, the mean $\pm$ 3 SD is more commonly recommended, especially in the social sciences where data are less likely to be as expensive and harder to come by as in soil science. Apparently also, the use of three standard deviations assumes that the data are near-normally distributed, so that the mean $\pm$ 3 SD covers about 99.7% of the population. But we know that soil variables are mostly skewed, and although data outside three standard deviations may be less common, they may still be within expected range of values in the natural soil populations.

## 5.4 Summary and Conclusion

This study provides the basis to put an end to the apparent confusion and contradiction in the soil science literature about the need to transform soil data before computing variability statistics. The probable problems associated with variability statistics computed from non-normal or skewed data were recognized and discussed. But unlike those emanating from the use of transformation, these problems are systematic and could be solved in the way put forward in this study. Moreover, variability statistics from untransformed data represent natural states of reality in the field, and are facts of the discipline of pedology, but what exactly are we measuring with transformed soil variables?. On occasions I agree with those who describe statistics as social conventions. And it is clear, therefore, that the desire to follow statistical convention cannot but be often tampered by practical considerations. This is as true in soil science as it is in most specific traditional fields of studies.

# CHAPTER 6

# MULTIVARIATE ANALYSIS OF MAP UNIT VARIABILITY IN NRCS-STATSGO: A CASE STUDY IN NORTHERN NEW ENGLAND

## 6.1 Introduction

### 6.1.1 State Soil Geographic Database (STATSGO).

The US Natural Resources Conservation Service (NRCS) formerly known as Soil Conservation Service (SCS) has the Federal leadership in a national effort to provide digital soil data for use in geographic information systems (GIS). NRCS has established three soil geographic databases representing kinds of soil maps at differing levels of detail, namely, Soil Survey Geographic Database (SSURGO), the most detailed; State Soil Geographic Database (STATSGO); and the National Soil Geographic Database (NATSGO), the least detailed of these digital soil databases. SSURGO is made from NRCS standard county soil surveys at scales typically between 1:15,000 to 1:24,000. Soil maps for STATSGO are compiled by generalizing the more detailed SSURGO maps. Where SSURGO maps are not available, data on geology, topography, vegetation, and climate are assembled and used, together with remotely sensed satellite images. Soils of like areas are studied, and the probable classification and extent of the soils is determined. Map unit composition for STATSGO is determined by sampling areas on the more detailed maps and expanding the data statistically to characterize the whole map unit. Then, using the US Geological Surveys 1:250,000 quadrangle series as a map base, the soil data are digitized to comply with national guidelines and standards (see SCS 1991, p. 2). Data for STATSGO are distributed as complete coverage for a state, and are available for most states of the US (see Bliss & Reymond, 1989; Lytle, 1993).

Soil survey has traditionally been the most practical method for partitioning field variation or grouping similar and separating different soils on a regional scale (Trangmar et al., 1985). The importance of soil surveys as a source of detailed information about the landscape, has been recognized in environmental and earth sciences, natural resource management, and land use planning (Lytle, 1993). Bliss & Reymond (1989) discussed the importance of the small-scaled digital soil data in regional, state and multistate-level resource management and planning, and noted that STATSGO was developed in recognition of the many values of small-scaled soil maps. In many soil-based regional and national analyses, STATSGO is the only source of soil information that is appropriate and/or available. STATSGO is of particular interest to regional ecosystem modeling community (Lathrop et al., 1995) because of its wide availability and digital format or readiness for use in geographic information systems GIS. As Hammer et al. (1991) observed, soil surveys have become a frequent component of GIS applications in natural resources planning, landuse planning, and environmental protection. And, such soil-based applications of GIS technology have continued to produce new users, and significantly increase the uses of soil survey information.

### 6.1.2 Problem Statements

Within the last two decades, concerns about the reliability of soil survey or accuracy of soil map information have gained increased importance among scientists and users of soil surveys and land evaluation data. The literature is replete with documentation of the causes of these concerns (e.g., Butler, 1980; Holmgren, 1988; Nash & Daugherty, 1990; Netttleton et al., 1991; Moore et al, 1993; Rogowski & Wolf, 1994). A soil survey is a predictive study to identify bodies of soils that can be recognized as natural units, predict and delineate their areas on maps, and describe the delineated areas in terms of kinds and properties of soils. One of the shortcomings of traditional soil surveys include the fact that the reliability of the predictions obtained from soil survey varies widely as a function of a number of factors (Webster, 1985; Hartung et al., 1991; Oberthur et al., 1996). Also, the inferred homogeneity in conventional soil maps does not exist for many soil physical and chemical attributes, and ranges given for some attributes often vary by an order of magnitude (Moore et al., 1993; Wilding, 1984). Other concerns of surveys include uncertainty regarding the placement of soil boundaries, presence of inclusions or map units containing *dissimilar* soils, and absence of quantitative expressions of map unit variability with respect to specific soil attributes. Digital soil maps such as the NRCS soil geographic databases are not only subject to all the problems of traditional soil survey procedures, but the process of digitization or automation is a source of other potential errors and uncertainties. Jordon et. al., (1986) as cited in Day et al. (1988) stated that in the US, approximately 80% of published soil surveys and 50% of soil surveys in progress are on spatially distorted base maps that do not meet National Map Accuracy Standards (NMAS). The kinds and sources of serious errors in a geographic information system are discussed by Lunetta et al. (1991), Heuvelink et al. (1989) and Burrough (1987). As Aronoff (1993) notes, error is introduced and propagated at every step in the process of generating and using geographic information.

The need to evaluate soil map quality or characterize the variability present within soil map units (Nordt et al., 1991; Brown & Huddleston, 1991) and of individual soil properties (Lammers & Johnson, 1991) is well documented. However, evaluating soil map reliability requires intensive field sampling and actual measurement of many properties of soil---a luxury that is often rare and practically non-existent for large areas. Hammer et al. (1991) warned of the existence of databases of unknown accuracy and precision, and emphasized the need to obtain ground-truth measurements to verify the precision of computer-generated soil maps. Lathrop et al. (1995) used STATSGO data to estimate soil water holding capacity needed in a regional ecosystem modeling study, and found greater within-unit variability than between map unit variability. They discussed other practical limitations of STATSGO, and concluded that estimates of the spatial variability of soil properties in STATSGO need to be better quantified and communicated to the prospective users, if the utility of STATSGO for modeling purposes is to be improved. To our knowledge, no systematic study has been done to evaluate the reliability of STATSGO or the degree of variation of specific soil properties within STATSGO map units, and between spatially-, and/or pedologically-related cartographic units. The ostensible reason for this is the unavailability of measured soil data, especially on regional scales, required for such a study.

### 6.1.3 Study Objectives

The objectives of this study were to 1) assess the reliability of STATSGO in the northern New England states; 2) quantitatively assess the variability of individual soil properties within selected STATSGO map units; and 3) evaluate the relative efficiency with which a number of edaphologically important soil chemical properties were mapped in STATSGO of the study region. The goal was to provide the users of the readily available STATSGO soil data, information on when and for what soil properties STATSGO is adequate, or the degree of variation in specific soil properties to expect within a given map unit and between related map units. The reference data used for this study were collected during the 1983 USDA Forest Inventory and Analysis (FIA) survey of the states of Maine, New Hampshire, and Vermont. The data include actual field and laboratory measurements of many taxonomically and edaphologically important soil variables, made from about 2,000 geo-referenced soil profiles in the study region. Using geographic information systems and multivariate statistical techniques, these data were analyzed to answer the following specific questions:

> 1) Are delineations of the same STATSGO taxonomical units significantly similar with respect to soil attributes that are important in soil classification and mapping?
>
> 2) Are STATSGO map units significantly different from one another, i.e., having smaller within-unit than between-unit variability, with respect to specific soil attributes?

3) What are the relative efficiencies with which the spatial variabilities of specific soil attributes are mapped in STATSGO of the study area?

4) Based on the taxonomically relevant soil attributes used in this study, how accurate was STATSGO in assigning soil profiles to the most probable map units, and what soil properties are most effective predictors of map unit membership?

## 6.2 Materials and Methods

### 6.2.1 Available FIA soil data

The FIA soil variables used in this study include the following soil chemical properties measured from the B-horizons of soil profiles dug at the FIA plots in the study area: calcium (Ca), potassium (K), sodium (Na) and magnesium (Mg), soil pH, phosphorus (P), aluminum (Al), iron (Fe), manganese (Mn), and zinc (Zn). In addition to these, field determinations of parent material types, texture of the B horizon, elevation, percent slope, and drainage condition, of the plot sites were also used. The chemical soil properties selected for this study have established and well documented edaphological importance (influence on agricultural and forestry plant growth). The field variables were selected because they are soil-forming factors (Jenny, 1941), criteria in Soil Taxonomy (Soil Survey Staff, 1975)---the US soil classification system, and/or are commonly used in soil mapping as indicators of change in soil types in the field. The field sampling and laboratory analyses of the selected soil attributes were as described in Chapters 4 of this dissertation.

### 6.2.2 GIS and Sample Selection Procedures

The processes of adapting FIA data to a relational database management system (Chapter 3), and creating a *point coverage* in a geographic information system (Chapter 4) were described earlier. STATSGO data for each state in the study area were received from the New Hampshire Geographically Referenced Analysis and Information Transfer Systems (GRANIT), as polygon coverage in ArcInfo (by ESRI, Inc.) format. STATSGO spatial data were in *Albers Conical* projection, and were accompanied by a number of attribute data files in a relational database format. The objectives of the GIS procedures in this study are similar to those described in the Materials and Methods section of Chapter 4. GIS techniques were used to spatially combine FIA and STATSGO maps, so that the STATSGO map units within which each FIA plot falls will be identified. The GIS operations required to achieve this objective were also performed in ArcInfo (ESRI, Inc.) just as described in Chapter 4. The result of the overlay analysis was brought up in *Paradox for Windows*, and again, relationally joined to the tables of FIA soil variables.

Each STATSGO polygon or delineation had a unique, four-digit polygon identification (PID) number. A polygon is a parcel within which the land is considered to be of the same kind or of a few kinds of soils that can be listed and described. For example, there were about 1250 of these polygons in Maine. Polygons or delineations that are similar in nature form a soil class called map unit. STATSGO map units are uniquely identified within each state by MUID---a concatenation of two-character State FIPS code (e.g., ME for Maine) and a three-digit Arabic number. There are about 70 STATSGO map units in Maine and 45 in New Hampshire, for example.

Although there were almost 3900 georeferenced FIA plot sites, only a fraction of these had any soil data, and much fewer had measurements for all the soil variables needed for this study. Convenience and especially ease of interpretation of statistical results dictated that *list-wise* method be used in the selection of FIA plots to include in many of the analyses. List-wise selection means that only FIA plots or soil profiles with data for all or most soil attributes of interest were included for analysis. Although, most STATSGO map units encompassed some FIA plot samples, only few map units had large enough samples to allow the statistical analyses in this study to be done with a satisfactory degree of confidence. Consideration about the adequacy of available sample-size, therefore, had a major influence on the way many of these statistical analyses were designed and eventually carried out in this study. Elaborate and prolonged data exploration and pre-analysis procedures were performed in *Paradox for Windows* (a relational database management program), *Microsoft Excel* (a spreadsheet program), and *SPSS for Windows* (a major statistical analysis, data management and display program), in order to gain familiarity with the content and structure of STATSGO data, and determine if and how the STATSGO and available FIA soil data could be used achieve the desired study objectives.

### 6.2.3 Statistical Analyses

The first hypothesis was that different delineations of the same STATSGO map unit are reasonably similar or internally homogeneous with respect to many soil attributes. To test this hypothesis, STATSGO map units (MUIDs) with multiple delineations (PIDs) having sufficiently large samples were selected. If a STATSGO soil class or map unit occurs in two or more states, it is usually given a different MUID in each of the states. Hence, PIDs were used to represent delineations of the same map unit within a state, while MUIDs were for delineations of the same map unit across states. The SPSS Independent t-tests and one-way analysis of variance (ANOVA) procedures were used to test if the means of specific soil attributes were significantly different among the delineations of same map units. Hotelling's $T^2$ (the multivariate equivalents of the Student's t-test) and multivariate analysis of variance (MANOVA) would have been more appropriate statistical procedures than the t-test and univariate ANOVA since this study involved multiple and correlated variables. However, the selected STATSGO delineations had insufficient number of samples to allow the analysis of more than a few variables at a time. The constraint of sample size-to-number of variable ratios in multivariate analyses was discussed in detail in Chapter 4.

Next, multivariate analysis of variance (MANOVA) and discriminant function analysis (DFA) procedures were used to test if STATSGO map units were distinct with respect to selected soil attributes. Seven STATSGO map units had sufficiently large sample sizes (ranging from 32 to 64) to allow their inclusion in the analysis. The analysis was two-fold, one for a group of soil chemical properties, and also for a group of soil variables considered important for classification. As in Chapter 4, the analyses were used to see if the STATSGO map units were statistically different from one another based on individual soil properties, and on all members of the group of soil properties considered together. The percent explained variance in individual soil variables or the relative efficiency with which the variation in each soil property was separated by STATSGO map units, was also assessed. Finally, the relative effect of the selected "soil mapping variables" in discriminating among STATSGO map units, and the classification accuracy of STATSGO based on these soil attributes, were also evaluated. The analyses were performed in *SPSS for Windows* using the MANOVA and DISCIMINANT procedures, and entering all the variables on one step. The seven selected map units were separated into two groups of five map units each, to allow the analysis of each group of soil variables (mentioned above) to be replicated. The two groups of map unit were formed by splitting the seven map units in half after ordering them by sample size, and including the median map units in both groups.

Table 6-1 provides limited taxonomic information about the STATSGO map units used in one or more of the analyses in this study. It lists the MUID's and major component soil types in the map units, and also, the Soil Taxonomic (Soil Survey Staff, 1975) descriptive names of the components. More detailed information about STATSGO data structure and/or nature of data it provides will be found in SCS (1991) or Bliss & Reymond (1989).

## 6.3 Results and Discussion

### 6.3.1 Variation Among Delineations of Same STATSGO Map Units

STATSGO MUID's ME059 and NH027, and ME005 and NH037 are each a map unit pair in which each member occurs in a different state. The Student t-test was used to test if members of each pair were significantly different from each other in terms of soil chemical properties. In addition, MUID's NH031, NH023, ME053 and ME064

| STATSGO Muid | Major Component Soils |
|---|---|
| ME005 | Becket-Monadnock-Tunbridge |
| ME019 | Dixfield-Brayton-Colonel |
| ME053 | Scantic-Buxton-Lamoine |
| ME059 | Skerry-Hermon-Monadnock |
| ME064 | Swanville-Boothbay-Lyman |
| NH012 | Canton-Hollis-Chatfield |
| NH017 | Colton-Adams-Monadnock |
| NH022 | Marlow-Peru-Monadnock |
| NH023 | Monadnock-Lyman-Tunbridge |
| NH026 | Hermon-Lyman-Berkshire |
| NH027 | Skerry-Hermon-Monadnock |
| NH031 | Marlow-Lyman-Berkshire |
| NH037 | Becket-Monadnock-Tunbridge |
| **Compname** | **Classification** |
| Adams | Typic Haplorthods, Sandy, Mixed, Frigid |
| Becket | Typic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Berkshire | Typic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Boothbay | Aquic Dystric Eutrochrepts, Fine-Silty, Mixed, Frigid |
| Brayton | Aeric Haplaquepts, Coarse-Loamy, Mixed, Nonacid, Frigid |
| Buxton | Aquic Dystric Eutrochrepts, Fine, Illitic, Frigid |
| Canton | Typic Dystrochrepts, Coarse-Loamy Over Sandy, Or -Sandy-Skeletal, Mixed, Mesic |
| Chatfield | Typic Dystrochrepts, Coarse-Loamy, Mixed, Mesic |
| Colonel | Aquic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Colton | Typic Haplorthods, Sandy-Skeletal, Mixed, Frigid |
| Dixfield | Typic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Hermon | Typic Haplorthods, Sandy-Skeletal, Mixed, Frigid |
| Hollis | Lithic Dystrochrepts, Loamy, Mixed, Mesic |
| Lamoine | Aeric Haplaquepts, Fine, Illitic, Nonacid, Frigid |
| Lyman | Lithic Haplorthods, Loamy, Mixed, Frigid |
| Marlow | Typic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Monadnock | Typic Haplorthods, Coarse-Loamy Over Sandy, Or -Sandy-Skeletal, Mixed, Frigid |
| Peru | Aquic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Scantic | Typic Haplaquepts, Fine, Illitic, Nonacid, Frigid |
| Skerry | Aquic Haplorthods, Coarse-Loamy, Mixed, Frigid |
| Swanville | Aeric Haplaquepts, Fine-Silty, Mixed, Nonacid, Frigid |
| Tunbridge | Typic Haplorthods, Coarse-Loamy, Mixed, Frigid |

**Table 6-1: Soil components and their classification, of STATSGO map units used in this study.**

had multiple delineations (PID's) and sufficient data to allow similar test of homogeneity of soil properties within each of the map units. Univariate ANOVA instead of Independent t-test was used to evaluate NH031 because this map unit had three delineations. The results of these tests are reported in Table 6-2, and include the soil variables and number of samples used, their means and standard deviations in each delineation, and the p values of the tests of significance. In the ANOVA test (for NH031), the mean total, and F-ratio were added, but the standard deviation was not reported. Prior to analysis, the data were evaluated to ensure that they were within reasonable limits of test assumptions. Non-linear transformation was used when necessary to correct gross departure from normality, and a preceding test of homogeneity of within-group variances was used to decide which t-test results to report.

As shown in Table 6-2, there were about 60 tests in all, that is, ten soil chemical properties were each analyzed within each of the six STATSGO map units. In 17 or about 30% of these tests, the delineations of the map units were significantly different from one another at the 95% confidence level. The delineations of ME053 (Table 6-2(b)iii) were not significantly different on any of the soil properties, while the delineations of NH023 Table (12(b)ii) were significantly different on only two soil properties (exchangeable acid and K). Probably contrary to expectation, delineations occurring in separate states (Table 6-2(a)'s) did not appear to be more heterogeneous with respect to the selected chemical properties than delineations occurring within a

(a) i

| Soil Vars. | Transf. | df | MEANS ME059 | MEANS NH027 | STD DEVIATION ME059 | STD DEVIATION NH027 | P for 2-tail t-test |
|---|---|---|---|---|---|---|---|
| Exch_Ca | sqrt | 52 | 12.00 | 9.52 | 7.26 | 5.49 | 0.127 |
| Exch_Fe | non | 67 | 93.43 | 91.81 | 72.66 | 111.41 | 0.945 |
| Exch_K | sqrt | 67 | 4.06 | 4.38 | 1.43 | 1.49 | 0.79 |
| Exch_Na | log | 67 | 1.08 | 1.42 | 0.13 | 0.20 | <.001* |
| Exch_Mg | non | 67 | 11.07 | 10.08 | 1.91 | 1.54 | 0.688 |
| Exch_P | log | 66 | 1.23 | 0.51 | 0.52 | 0.64 | <.001* |
| Soil CEC | non | 27 | 4.63 | 6.92 | 1.56 | 1.88 | 0.002* |
| Org. Matter | non | 28 | 4.32 | 4.53 | 3.01 | 4.53 | 0.871 |
| OM_dept | non | 75 | 3.58 | 3.43 | 6.80 | 3.53 | 0.903 |
| Soil pH | non | 67 | 4.55 | 4.67 | 0.26 | 0.43 | 0.012* |

(a) ii

| Soil Vars. | Transf | df | MEANS ME005 | MEANS NH037 | STD DEVIATION ME005 | STD DEVIATION NH037 | P for 2-tail t-test |
|---|---|---|---|---|---|---|---|
| Exch_Ca | sqrt | 34 | 20.96 | 14.29 | 7.32 | 8.45 | 0.042* |
| Exch_Fe | sqrt | 33 | 10.30 | 10.07 | 3.68 | 3.60 | 0.876 |
| Exch_K | log | 33 | 1.23 | 1.28 | 0.18 | 0.27 | 0.523 |
| Exch_Na | non | 32 | 13.96 | 24.21 | 5.49 | 13.89 | 0.003* |
| Exch_Mg | log | 34 | 1.26 | 1.10 | 0.41 | 0.48 | 0.371 |
| Exch_P | log | 21 | 1.16 | 0.76 | 0.36 | 0.64 | <.001* |
| Soil CEC | non | 18 | 5.02 | 8.99 | 0.94 | 1.42 | 0.032* |
| Org. Matter | non | 18 | 3.08 | 4.54 | 2.30 | 4.37 | 0.354 |
| OM_dept | log | 34 | 0.45 | 0.24 | -0.75 | -0.69 | 0.09 |
| Soil pH | non | 35 | 4.69 | 4.65 | 0.31 | 0.46 | 0.163 |

(b)i

| Soil Vars. | Transf | MEANS PID1957 | MEANS PID1862 | MEANS PID1926 | Total mean (df) | F-ratio | P. for ANOVA |
|---|---|---|---|---|---|---|---|
| Exch_Acid | log | 0.68 | 0.62 | 1.11 | 0 .77 (27) | 4.37 | .024* |
| Exch_Ca | log | 1.66 | 2.01 | 2.03 | 1.89 (33) | 3.27 | 0.052 |
| Exch_Fe | sqrt | 6.50 | 11.62 | 11.15 | 9.65 (33) | 9.97 | <.001* |
| Exch_K | log | 1.30 | 1.42 | 1.45 | 1.39 (33) | 1.50 | 0.238 |
| Exch_Na | non | 10.58 | 20.33 | 23.33 | 17.95 (33) | 10.39 | <.001* |
| Exch_Mg | log | 0.69 | 1.14 | 1.24 | 1.02 (33) | 11.22 | <.001* |
| Exch_P | log | 0.25 | 0.38 | 0.70 | 0.45 (33) | 2.82 | 0.075 |
| Soil CEC | non | 8.74 | 9.50 | 7.33 | 8.13 (24) | 1.68 | 0.21 |
| Org. Matter | non | 5.23 | 7.15 | 5.02 | 5.28 (24) | 0.424 | 0.659 |
| OM_dept | sqrt | 1.40 | 1.32 | 1.28 | 1.33 (33) | 0.25 | 0.782 |
| Soil pH | non | 4.69 | 4.47 | 4.61 | 4.60 (28) | 1.45 | 0.254 |

**Table 6-2: Mean, standard deviations, and results of tests of differences in soil properties among delineations of the same STATSGO map units.** (a)i & ii show the same map units occuring in different states, while (b)i - iv represent multiple delineations of NH031, NH023, ME053, and ME064, respectively. * = test was significant at the 95% confidence level.

Table 6-2 (cont.)

(b)ii

| Soil Vars. | Transf | df | MEANS PID 1813 | MEANS PID 1939 | STD DEVIATION PID 1813 | STD DEVIATION PID 1939 | P for 2-tail t-test |
|---|---|---|---|---|---|---|---|
| Exch_Acid | sqrt | 33 | 3.19 | 2.02 | 1.37 | 0.62 | .001* |
| Exch_Ca | log | 42 | 1.71 | 1.94 | 0.54 | 0.54 | 0.162 |
| Exch_Fe | log | 42 | 1.89 | 1.76 | 0.31 | 0.33 | 0.192 |
| Exch_K | log | 42 | 1.13 | 1.38 | 0.31 | 0.25 | .006* |
| Exch_Na | log | 42 | 1.12 | 1.13 | 0.13 | 0.15 | 0.735 |
| Exch_Mg | log | 42 | 0.71 | 0.90 | 0.28 | 0.36 | 0.065 |
| Exch_P | log | 41 | 0.75 | 0.39 | 0.67 | 0.73 | 0.100 |
| Soil CEC | non | 39 | 6.91 | 7.68 | 2.98 | 2.65 | 0.394 |
| Org. Matter | non | 38 | 4.29 | 5.90 | 3.20 | 3.93 | 0.162 |
| OM_dept | log | 44 | 0.37 | 0.29 | -0.66 | -0.56 | 0.500 |
| Soil pH | non | 43 | 4.91 | 4.80 | 0.39 | 0.34 | 0.111 |

(b) iii

| Soil Vars. | Transf | df | MEANS PID 751 | MEANS PID 1134 | STD DEVIATION PID 751 | STD DEVIATION PID 1134 | P for 2-tail t-test |
|---|---|---|---|---|---|---|---|
| Exch_Acid | non | 39 | 6.04 | 7.20 | 2.63 | 2.74 | 0.173 |
| Exch_Ca | log | 28 | 2.17 | 1.78 | 0.62 | 0.48 | 0.101 |
| Exch_Fe | sqrt | 28 | 8.68 | 8.17 | 2.45 | 1.41 | 0.48 |
| Exch_K | log | 28 | 1.48 | 1.54 | 0.23 | 0.16 | 0.51 |
| Exch_Na | sqrt | 28 | 4.50 | 4.77 | 0.93 | 0.62 | 0.425 |
| Exch_Mg | log | 27 | 1.24 | 1.03 | 0.58 | 0.44 | 0.35 |
| Exch_P | log | 26 | 1.08 | 0.89 | 0.57 | 0.17 | 0.181 |
| Soil CEC | non | 28 | 7.41 | 6.22 | 3.33 | 2.21 | 0.335 |
| Org. Matter | non | 28 | 1.89 | 2.90 | 1.66 | 1.58 | 0.133 |
| OM_dept | log | 40 | 0.07 | 0.19 | -0.70 | -0.65 | 0.26 |
| Soil pH | non | 41 | 4.89 | 4.75 | 0.60 | 0.55 | 0.445 |

(b)iv

| Soil Vars. | Transf | df | MEANS PID 553 | MEANS PID 868 | STD DEVIATION PID 553 | STD DEVIATION PID 868 | P for 2-tail t-test |
|---|---|---|---|---|---|---|---|
| Exch_Acid | sqrt | 41 | 2.41 | 2.92 | 0.51 | 0.56 | .004* |
| Exch_Ca | log | 32 | 2.12 | 1.23 | 0.80 | 1.05 | .023* |
| Exch_Fe | log | 34 | 1.95 | 1.89 | 0.21 | 0.24 | 0.519 |
| Exch_K | sqrt | 32 | 5.98 | 4.20 | 1.25 | 1.52 | .003* |
| Exch_Na | non | 34 | 23.20 | 21.65 | 6.46 | 8.17 | 0.595 |
| Exch_Mg | log | 30 | 1.25 | 0.89 | 0.70 | 0.38 | 0.171 |
| Exch_P | log | 31 | 0.74 | 0.71 | 0.51 | 0.50 | 0.193 |
| Soil CEC | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| Org. Matter | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| OM_dept | log | 40 | 0.29 | 0.39 | -0.54 | -0.74 | 0.362 |
| Soil pH | non | 22 | 4.89 | 4.58 | 0.70 | 0.39 | 0.111 |

state (Table 6-2(b)i & iv). In three out of the four map units in which data on exchangeable acid were available, the delineations were significantly different from one another. Test for Na was significant three times out of six. Aside from these, no chemical property showed consistent heterogeneity within the map units analyzed in this study. It seems, however, that the macronutrient cations and hence, cation exchange capacity were clearly more variable (Na > CEC = Ca = K < Mg) among delineations, compared to micronutrient cations.

### 6.3.2 The Variation of Soil Chemical Properties in STATSGO Map Units

Preliminary data screening showed that non-linear tranformation was needed, for most of the soil variables, in order to achieve univariate normality. Inspection of pooled within-group correlation matrices (Table 6-3) showed modest inter- correlations among the soil variables in the MANOVA and DFA tests. The highest inter- variable correlation was 0.75, hence there was no evidence of the problem of multicollinearity. However, the Box's M multivariate test of homogeneity of within-group variance-covariance matrices was found to be highly significant, $p < .001$. Bartlett-Box F test of univariate homogeneity of variance was significant for Ca > Mg > Na > Fe, $p > .005$, indicating that these variables were responsible for the significant Box's M test, most likely because of their non-normal distributions. Again, the sensitivity of Box's M test was recognized, and secondary or confirmatory diagnostics (as explained in Chapter 4) suggested that there was no real problem of heterogeneity of dispersion matrices. No problem was also found when the stability of the MANOVA and DFA results was evaluated by the addition and/or deletion of variables, and by repeating the analyses on a fresh batch of data.

| | Exch_Al | Exch_Ca | Exch_Fe | Exch_K | Exch_Mg | Exch_Mn | Exch_P | Exch_Na | Exch_Zn |
|---|---|---|---|---|---|---|---|---|---|
| Exch_Al | | | | | | | | | |
| Exch_Ca | -0.296 | | | | | | | | |
| Exch_Fe | 0.380 | 0.172 | | | | | | | |
| Exch_K | 0.265 | 0.371 | 0.321 | | | | | | |
| Exch_Mg | -0.025 | 0.758 | 0.347 | 0.591 | | | | | |
| Exch_Mn | -0.157 | 0.426 | 0.024 | 0.400 | 0.485 | | | | |
| Exch_P | -0.385 | 0.311 | 0.050 | -0.043 | 0.075 | 0.179 | | | |
| Exch_Na | 0.041 | 0.260 | 0.191 | 0.334 | 0.343 | 0.266 | 0.086 | | |
| Exch_Zn | 0.377 | 0.150 | 0.207 | 0.394 | 0.333 | 0.326 | -0.130 | 0.032 | |
| Soil pH | -0.668 | 0.470 | -0.076 | -0.021 | 0.228 | 0.253 | 0.318 | 0.090 | -0.223 |

(a)

| | Prnt. Mtrl. | B-text. | Elevatn. | Exch_Ca | Exch_Acid | Slope |
|---|---|---|---|---|---|---|
| Prnt. Mtrl. | | | | | | |
| B-texture | 0.042 | | | | | |
| Elevation | -0.138 | 0.135 | | | | |
| Exch_Ca | -0.060 | 0.242 | 0.073 | | | |
| Exch_Acid | 0.008 | -0.017 | -0.075 | -0.187 | | |
| Slope | -0.084 | 0.078 | 0.241 | -0.105 | -0.022 | |
| Drainage | -0.119 | 0.079 | 0.138 | 0.249 | -0.013 | -0.302 |

(b)

**Table 6-3: Pooled within-group correlation matrices of soil attributes used in MANOVA and DFA of STATSGO map units.**

The results of MANOVA tests of significant differences in selected chemical properties among STATSGO map units, are shown in Table 6-4. The table shows the seven selected map units, the soil variables used, their sample sizes, the group and total means. It also shows the p values of univariate and multivariate tests of significance, as well as the univariate and multivariate Wilk's Lambda's and effect-size (1 - Lambda) for the analysis. The selected STATSGO map units were significantly different from one another, both with respect to individual soil chemical properties, and when all the

| Map Units | n size | Sqrt_Al | Sqrt_Ca | Log_Fe | Log_K | Log_Mg | Log_Mn | Log_P | Log_Zn | Log_Na | Soil pH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Group Means | | | | | |
| ME059 | 30 | 10.871 | 11.990 | 1.860 | 1.167 | 0.878 | 0.765 | 1.233 | -0.064 | 1.081 | 4.546 |
| NH027 | 32 | 12.813 | 9.191 | 1.781 | 1.240 | 0.875 | 0.443 | 0.533 | 0.150 | 1.422 | 4.676 |
| ME064 | 32 | 12.276 | 10.533 | 1.897 | 1.325 | 1.029 | 0.950 | 0.795 | -0.113 | 1.349 | 4.716 |
| ME019 | 37 | 10.627 | 13.711 | 1.802 | 1.304 | 1.028 | 1.249 | 0.990 | -0.149 | 1.331 | 4.805 |
| NH017 | 31 | 9.405 | 8.187 | 1.646 | 1.146 | 0.750 | 0.658 | 0.831 | 0.051 | 1.246 | 4.834 |
| NH012 | 49 | 9.401 | 5.796 | 1.715 | 1.176 | 0.672 | 0.522 | 0.790 | -0.071 | 1.089 | 4.848 |
| NH022 | 61 | 10.400 | 10.658 | 1.697 | 1.230 | 0.879 | 0.660 | 0.765 | 0.111 | 1.341 | 4.843 |
| NH023 | 66 | 11.626 | 9.068 | 1.782 | 1.254 | 0.833 | 0.674 | 0.610 | 0.077 | 1.181 | 4.829 |
| **Total** | 338 | 10.876 | 9.718 | 1.764 | 1.232 | 0.858 | 0.723 | 0.791 | 0.011 | 1.249 | 4.782 |
| **Sig. univ. F (7, 330) test** | | 0.003 | <.001 | 0.004 | 0.106 | <.001 | <.001 | <.001 | <.001 | <.001 | 0.006 |
| **Univ. $R^2$ (1 - Lambda)** | | 0.062 | 0.103 | 0.061 | 0.035 | 0.079 | 0.120 | 0.084 | 0.084 | 0.248 | 0.057 |

**Multivariate Test of Significance**

| | |
|---|---|
| **F (70, 1878) = 5.41** | **Sig. of F = < .001** |
| **Wilk's Lambda = .343** | **R2** (variance explained) **= apprx. 65.70%** |

**Table 6-4: Results of MANOVA of soil chemical properties in selected STATSGO map units in the study area**

soil properties were considered simultaneously. Table 6-4 shows highly significant univariate $F(7, 33)$, $p < .001$ or close for all variables except K for which the p value was 0.106. The variance of each of the soil variables that was explained by the selected map units ranged from about 6% for pH $\simeq$ Fe = Al, to 25% for Na. The multivariate $F(70, 1878)$ was 5.41, and was highly significant, $p < .001$. The multivariate Wilk's Lambda was .343, indicating that about 65% of the variance in the composite of the chemical properties could be explained by, or predicted from the map units. The observed univariate and multivariate proportions of explained variances in the chemical properties seem rather impressive considering that 1) these variables are not field-observable or mappable soil properties, 2) capturing the variability of many of them (e.g., the macronutrient cations) is, usually, not the primary objective focus of a general soil survey like STATSGO, and 3) except for a few of them (e.g., Ca) these soil variables are not major criteria in Soil Taxonomy (Soil Survey Staff, 1975) on which STATSGO is based.

### 6.3.3 Statistical Evaluation of the Reliability of Soil Classification in STATSGO

Soil classification is usually polythetic, that is, soil class membership is based on observations of several variables, no one of which is either [absolutely] necessary or sufficient to define the class (Webster & Burrough, 1974). As Edmonds et al. (1981) observed, soil characteristics can be categorized into those observable [in the field] by the senses and those observable only by laboratory procedures. Soil mapping usually involves the evaluation of the spatial variation in field-observable soil characteristics, and the assumption that the observable characteristics are correlated with those measurable only by laboratory procedures. It is not possible for a soil map to efficiently reflect spatial variation in all soil variables simultaneously. However, to be reliable, a soil map would have relatively high within-map unit homogeneity and between-map unit variances in soil properties relevant to the purpose of the survey, and/or used in making the soil map. The reliability of a soil map, or the effectiveness of its soil classification can be statistically evaluated by comparing the within and between variances of the map-unit differentiating characteristics and characteristics strongly correlated with them (see Webster & Oliver, 1990; Burrough, 1993). Higher between-unit than within-unit variance (i.e., a high F-ratio or significant F-test) in a soil property implies that the map units are distinct, and hence that the classification is effective with respect to that soil property (see Leenhardt et al., 1994; Oberthur et al. 1996). MANOVA and discriminant function analysis (DFA) are the statistical procedures for evaluating the reliability of soil maps involving multiple variables simultaneously.

DFA and MANOVA were performed to assess how well selected STATSGO map units could be predicted from parent material type, texture of B-horizon, elevation, percent slope, drainage status, exchangeable Ca, and exchangeable acid. These soil variables were selected because of their known importance in soil classification and mapping systems. DFA also allowed the assessment of the relative importance of these variables in the assignment of soils to STATSGO map units. The STATSGO map units selected for this study, and some of the results of the analyses are shown Table 6-5 . The number of soil samples in the selected map units are as follows: NH017 (n = 31), ME059 (n = 31), NH027 (n = 32), ME064 (n = 38), ME019 (n = 39), NH012 (n = 52), NH023 (n = 64), and NH022 (n = 66). The analyses were performed twice to see if the results would be consistent, and they were. Table 6-5(i) shows results based on five map units with relatively smaller sample sizes (total n = 170), while Table 6-5(ii)'s results were based on map units with larger sample sizes (total n = 259). Pre-analysis evaluation of data for conformity to the assumptions of MANOVA and DFA was done as in Chapter 4, and the results were satisfactory.

Univariate F-tests showed that STATSGO map units in each of the two groups were highly significantly different from one another with respect to each of the selected soil-mapping variables. All of the univariate tests had $p < .001$ except for parent material which was .016 and .005 in the first and second tests, respectively. Recall that Wilk's Lambda expresses the percent of variance in a dependent variable or set of variables not explained by the independent variable. Wilk's Lambda values for the univariate tests (Table 6-5) show that parent material had the least proportion of variance about 6 - 8%) explained by STATSGO map units in the study area. On the other hand, about 55 - 60% of the variation in elevation among STATSGO map units could be explained in each of the tests. In both tests, parent material and elevation were the least and most effective discriminator of STATSGO map units, respectively. According to this study, STATSGO map units membership is weighted on the selected variables in the following order: elevation > exchangeable acid > drainage > B texture > %slope > Exchangeable Ca > parent material. The variance explained in the variables other than elevation and parent material was always greater than 10% and sometimes as high as 30%.

(i)

| | $D_1$ | $D_2$ | $D_3$ | |
|---|---|---|---|---|
| | Group Centroids for Disc. Functions ($D_i$) | | | |
| ME059 | 0.016 | 0.999 | -0.115 | |
| NH017 | -1.542 | -0.326 | -0.650 | |
| NH027 | -2.417 | -0.146 | 0.524 | |
| ME064 | 1.911 | -0.322 | 0.019 | |
| ME019 | 1.333 | -0.076 | 0.157 | |
| | Std. Canonical Disc. Function Coefficients | | | Univ. W's. L. |
| Parent Material | -0.254 | 0.315 | -0.314 | 0.929 |
| B_horiz. Texture | 0.362 | -0.496 | 0.486 | 0.779 |
| Elevation | -0.815 | 0.370 | 0.160 | 0.412 |
| Sqrt_Ca | 0.033 | 0.341 | 0.108 | 0.919 |
| Log_Acid | 0.371 | 0.664 | -0.175 | 0.636 |
| %Slope | -0.035 | 0.179 | 0.501 | 0.803 |
| Drainage | 0.397 | 0.463 | 0.383 | 0.698 |
| Can. Correlation | 0.860 | 0.433 | 0.350 | |
| Eigenvalue | 2.840 | 0.231 | 0.141 | |
| Wilk's Lambda for significant Disc. Functions (D1, 2 & 3) =.177 | | | | |

(ii)

| | $D_1$ | $D_2$ | $D_3$ | |
|---|---|---|---|---|
| | Group Centroids for Disc. Functions ($D_i$) | | | |
| ME064 | -2.057 | 0.265 | <.001 | |
| ME019 | -1.480 | 0.497 | -0.352 | |
| NH012 | -0.440 | -1.123 | 0.208 | |
| NH023 | 0.911 | 0.52 | 0.511 | |
| NH022 | 1.522 | -0.066 | -0.452 | |
| | Std. Canonical Disc. Function Coefficients | | | Univ. W's. L. |
| Parent Material | 0.318 | 0.010 | 0.281 | 0.943 |
| B_horiz. Texture | -0.438 | 0.070 | -0.126 | 0.865 |
| Elevation | 0.912 | 0.630 | -0.134 | 0.452 |
| Sqrt_Ca | -0.026 | 0.470 | -0.338 | 0.883 |
| Log_Acid | -0.301 | 0.681 | 0.499 | 0.791 |
| %Slope | 0.038 | 0.342 | 0.487 | 0.889 |
| Drainage | -0.153 | 0.540 | -0.280 | 0.863 |
| Can. Correlation | 0.803 | 0.523 | 0.358 | |
| Eigenvalue | 1.820 | 0.376 | 0.147 | |
| Wilk's Lambda for significant Disc. Functions (D1, 2 & 3) =220 | | | | |

**Table 6-5: Results of discriminant function analysis of taxonomic soil variables on selected STATSGO map units in the study area.**

Table 6-5 also shows that three discriminant functions, D1, D2 and D3 were derived in each of the two tests, implying that there are three orthogonal or non-overlapping dimensions along which the group of STATSGO map units can be separated based on the selected variables. In both tests, the discriminant functions were highly significant, $p < .001$. This means that the optimal weighted linear combination of the variables (D's) differ significantly across the map units, and hence can be used to predict map unit membership at significantly better than chance levels of accuracy. With five map units, up to four discriminant functions are theoretically possible but usually only the first few are important. The first discriminant function provides the best separation among the groups, followed by the second discriminant function, and so on. And since the discriminant functions are orthogonal to one another, the second function, for example, separates groups only on the basis of association not used in the first. Table 6-5 also shows the canonical correlations and eigenvalues for the discriminant functions (D1 to D3), the within-group centroids (vector of means) for these functions, Wilk's Lambda for the analyses, and the standardized canonical discriminant function coefficients. Canonical correlation indicates the strength of relationship between each discriminant function and group membership, while the standard canonical discriminant function coefficient (CDFC) shows the strength of relationship or correlation between group membership and the discriminating variables on each of the functions. The larger the absolute CDFC values, the more the variable contributes to the discriminating power of a particular discriminant function, while weights close to zero indicate variables that do not add much to discriminating among groups. Eigenvalues ($\lambda$) convey equivalent information as the canonical correlation (R); $\lambda / (\lambda + 1) = R^2$.

In the two tests shown in Table 6-5, D1 had R = .86 and .80, indicating that it is highly related to STATSGO map units. In both tests also, D1 was most highly loaded on elevation, and much more weakly on any other variable. The within-group centroids shows rather high distinction among map units on this discriminant function (D1), in both tests. The second discriminant function, D2, was moderately but also significantly related to map unit membership, R = .43 and .52 for test 1 and 2, respectively. In test 2, D2 was still moderately loaded on elevation, underscoring the importance of elevation in predicting STATSO map units. In both tests, D2 showed relatively high CDFC with exch_Acid, drainage, and B-texture in test 1 but exch_Ca in test 2. The within group centroids also show good separability among the map units on D2. The multivariate Wilk's Lambda was .177 and .220 in the two tests, implying that on the average, about 80% of the variation among the selected STATSGO map units could be explained, accounted for, or predicted from the variables used in this study.

Finally, the discriminant functions developed from the selected variables (discussed above) were used to classify the sampled soil profiles into the selected STATSGO map units. The resulting statistical classification was then used as a reference to assess how accurate STATSGO soil map was in predicting map unit membership. The result of accuracy assessment of soil classification by STATSGO are shown on Table 7-2, and are more fully discussed in the next chapter (*The Use of Error Matrix in Evaluating Classification Accuracy and Soil Map Quality*). In summary, the results showed that based on soil parent materials, texture of B-horizon. elevation exchangeable Ca, exchangeable acid, percent slope and Drainage information (Table 6-5), the overall accuracies of STATSGO in classifying the map units shown on Tables 6-5a and 6-5b were 61.76% and 63.32%, respectively. This is about 300% better than the expected rate of accuracy by random chance alone (i.e., 20% or 100% divided by number of map categories). The results are also impressive because the soil variables used in this study are neither the only ones used in soil mapping, nor are they necessarily the most important or even the primary classification criteria used by STATSGO.

## 6.4 Summary and Conclusion

The importance of STATSGO as the only source of small-scaled digital soil data appropriate and/or available for regional soil-based studies was discussed. The literature reveals concerns about the reliability of STATSGO, and the need to evaluate the degree of variation of specific soil properties within and between STATSGO map units. Using the soil component of the 1983 FIA data as reference, this study provides the only known systematic regional analysis of the reliability of the classification in STATSGO soil map. A variety of statistical analyses were used to provide specific answers about map unit variabilities of individual soil properties in, and about the reliability of soil classification and soil map quality of STATSGO.

The study showed that 60% or more of the time, the delineations of a STATSGO map unit occurring across states are homogeneous with respect to specific soil properties. This percentage increases (up to 100% in Table 6-2b(iii)) for delineations in the same state. Table 6-2 shows that 70% of the 60 univariate tests of differences in soil properties among multiple delineations was not significant at the 95% confidence level. Exchangeable acid, P, Na, Ca, and hence CEC are the soil properties on which the delineations of a STATSGO map units may be statistically different from one another.

DFA and MANOVA were used to see if selected STATSGO map units were significantly distinct from one another, both with respect to individual soil chemical properties, and when all the soil properties were considered simultaneously. The univariate tests were highly significant for almost all soil properties. The multivariate F-test was also highly significant, and indicated that about 65% of the variance in the composite of the chemical properties could be explained by, or predicted from STATSGO map units.

Finally, DFA was used to assess the reliability of soil classification in STATSGO, and evaluate the relative predictive efficiency of selected soil properties that are important in soil classification and mapping. The DFA test was replicated with five STATSGO map units each. Both tests were consistent, and showed that there were optimal weighted linear combinations of the variables on which the selected map units differ significantly; that on the average, about 80% of the variation among the selected STATSGO map units could be explained, accounted for, or predicted from the selected soil variables; and that elevation is the most effective discriminator (among the selected variables) of STATSGO map units, while parent material is the least effective. Only about 8% or less of the variation in parent material among the STATSGO map units could be explained, while as much as 60% of the variation in elevation was captured by the map units. The assessment of classification accuracy based on DFA showed that STATSGO has, on the average, an overall accuracy of about 62%---about 300% better than the expected rate of accuracy by random chance alone.

The results of this study are significantly better than some of the literature cited in this study would cause one to expect. Perhaps, many of the critics of soil survey data are not aware of the nature of soil variation and/or the practical and invariable limitations of soil survey methodology. A general and small-scale soil map like STATSGO can not be reasonably expected to capture the vertical and horizontal variabilities of all soil and soil-related properties simultaneously and with great precision. Soil water holding capacity (SWHC) on which Lathrop et al (1995) based their critique of STATSGO is well known to be difficult to determine even by direct methods (Nielsen et. Al., 1973; Peck et al., 1977). Lathrop et al. (1995) acknowledged that "field studies indicate that soil-water properties are particularly spatially heterogeneous, even for study areas that were fairly uniform in soil classification." The high within map unit variability of SWHC reported in that study should not have been surprising, and must have been aggravated by the scale at which the analysis was carried out and level of spatial interpolation reported in that study. This study shows that STATSGO map units in the northern New England area are distinct; that soil classification in STATSGO may be surprisingly accurate; and that the within-map unit variability of many individual soil properties are within very reasonable limits, especially for a small-scale generalized soil map.

# CHAPTER 7

# THE USE OF ERROR MATRIX IN EVALUATING CLASSIFICATION ACCURACY AND SOIL MAP QUALITY

## 7.1 Introduction

The last one-half of the sixty or more years of soil survey has witnessed ever increasing concerns about the reliability of soil survey or accuracy of soil maps. The literature is replete with documentation of the causes of these concerns (e.g., Moore et al., 1993; Rogowski & Wolf, 1994). Published research has continued to reveal the need to find an effective quantitative method of evaluating and expressing the reliability of soil classification and soil map quality. Prior to the mid-1960's, soil maps were simply presumed to have met the theoretical standard of 85% map unit "purity" or better (Soil Survey Staff, 1951). This means that the cartographic units of a soil map were expected, 85% of the time, to be composed of soils that are in the taxonomic units they purport to represent. However, later studies (e.g., Wilding et al., 1965; Powell & Springer, 1965; McCormack & Wilding, 1969; Amos and Whiteside, 1975) showed that the impurity of soil survey mapping units was much higher than the theoretically expected 15%. In fact, these studies showed that up to 50% or more of the soils included in a soil survey map units may be taxonomically different from the named soil. These studies spawned concerns about quality of soil maps, and clearly established the need to evaluate and document soil map unit composition.

Determination of map unit composition involves using transects or stratified random sampling to sample delineations of map units to be studied. These sampled soils are then evaluated to see if they are the same as the taxonomic unit (soil class) the map unit represents. The objective is to determine the proportion of soils within a map unit that is in the same taxonomic class as the named soil. Confidence intervals are then calculated using either the Student's t-distribution or a binomial method (see Wilding & Drees, 1983; Upchurch et al., 1988; and Burrough, 1991).

By late 1980's (e.g., Edmonds & Lentner, 1986; and Hopkins et al., 1987), it was well understood that map unit purity of 85% was impossible and that 50% or less was more practical, unless the taxonomic purity was examined at higher levels of soil taxa, or interpretive (instead taxonomic) purity was examined (West et al., 1981; Nordt et al., 1991). In interpretive purity, soils that were taxonomically dissimilar but had similar interpretations are allowed to be included in the map unit. The problem is that the definitions of similar and dissimilar soils (Soil Survey Staff, 1983) are subjective, user-biased and dependent on intended land use (Nordt et al., 1991). According to Miller et al. (1979), and Wilding & Drees (1983), taxonomic purity of map units is not a proper measure of quality or precision of soil survey. The alternative and "better" approach to evaluating soil map "quality" has been to assess the variability of individual soil properties in the map units. This method uses parametric or nonparametric statistics to analyze the between and within map units variances of selected soil properties, and to compute summary statistics including coefficients of variation for these soil properties within map units.

The major advantage of the soil-property-variability method is that it allows "precise quantification" of some sort. Quantification gives soil survey the glamour of "real science", but more importantly, quantitative evaluation has become an absolute necessity in the information age where there is "···an increasing need for measured data and hard conclusions" (Bouma, 1988). And as Bouma also asserted, users of soil survey data have become more professional and sophisticated, and "descriptive and qualitative recommendations are often not adequate anymore: they don't standup in court" (see also (Miller, 1978; Wilding, 1988; Brubaker & Hallmark, 1991). However, the apparent advantages of the soil-property-variability method overshadow and preclude the consideration of major limitations of this approach to assessing soil map accuracy. First, we know that soil classes are usually polythetic---class membership is based on observations of several variables, no one of which is either [absolutely] necessary or sufficient to define the class (Webster & Burrough, 1974). And map units cannot be expected to efficiently separate the variations in all important soil properties simultaneously. Hence, the assessment of soil map quality by the soil-property-variability method involves the sampling and laboratory measurements of a plethora of soil properties. This (as we know) is a costly venture, and the major reason why soil surveys are often based primarily or even entirely on field-observable soil properties and soil-related factors. On the other hand, the taxonomic purity method is in consonance with the soil survey methodology or the art and science process with which the soil map under evaluation was made. What the taxonomic purity method lacks at present is a more efficient way to describe "map unit purity", and to further quantify the extent of "similar" and "dissimilar" soils included in the map units. This paper presents a technique---the use of error matrices and associated descriptive statistics, which will allow these objectives and more to be achieved. These techniques have become the standard for providing comprehensive, quantitative accuracy assessment of maps or classifications of remotely sensed data. The objective of this section is to show that the error matrix techniques can be readily adapted for use in assessing the reliability of soil classification and soil map quality.

## 7.2 Advances in Classification Accuracy Assessment: the Use of Error Matrices

The error matrix and related techniques have gained much popularity in remote sensing where they have become the standard form for expressing classification accuracies, and reporting site-specific errors of commission and omission (see Congalton 1991; Lillesand and Kiefer 1994, p 612; Jensen, 1996, p. 247). The art and science of classifying the landscape into specifically defined map categories based on satellite remotely sensed data, are markedly similar to those of soil classification and mapping. Both remote sensing and soil mapping are interpolative, and are based on the indirect use of surrogate or correlated data to make judgement about the nature of map units. As a result, the need to evaluate and effectively express the level of "correctness" in the results of classification and mapping is equally critical in both remote sensing and soil mapping products. Campbell (1987) writes that accuracy assessment of remotely sensed data affects the legal standing of maps and reports, the operational usefulness of such data for land management, and their validity as a basis for scientific research. Within the last two decades, the remote sensing community has made significant advancement in the area of accuracy assessment of classification through the use of error matrix and discrete multivariate statistical analyses. Hay (1979) showed that accuracy assessment of map accuracy involves answering the following essential questions:

1). What proportion of the classification decision is correct?

2). What proportion of assignments to a given category is correct?

3). What proportion of a given category is correctly classified?

4). Is a given category overestimated or underestimated?

5). Are the errors of classification randomly distributed?

Since their introduction (Congalton et al., 1983), error matrix and associated analytical statistical techniques have been used by the remote sensing community to effectively answer these classification accuracy questions. The error matrix has been declared ..."essential for any serious study of accuracy" (Campbell, 1987), and a "starting point for a series of descriptive and analytical statistical techniques" (Congalton 1991) that provide comprehensive information on the accuracy of a classification or reliability of a map. Clearly, these techniques can also be adapted for use in evaluating the accuracy of soil classification and soil map quality.

The thrust of this paper is that the use of error matrix and the related statistics would significantly improve the present methods of assessing soil map quality. The following discussion will briefly introduce the concepts of error matrix and the complementary statistics; and show that adapting these techniques for use in soils is potentially the solution to the age-long search for an effective method of evaluating and communicating soil map quality. The application of these techniques is demonstrated with real data of STATSGO soil classification.

## 7.3 Description of the Error Matrix

An error matrix is a square array of numbers set out in rows and columns that represents the number of information classes or classification categories, used to compare on a category-by-category basis, the relationship between known reference data and the corresponding results of a classification. Table 7-2 shows an error matrix developed to assess the classification accuracy of a map involving five map units (MU's). The columns and rows of an error matrix show the number of sample units assigned to a particular map category (i.e., soil map unit) relative to the number that actually belong to that category (e.g., soil taxon) as verified in the field. The

| | | Soil Map Data | | | | | |
|---|---|---|---|---|---|---|---|
| | | **MU_1** | **MU_2** | **MU_3** | **MU_4** | **MU_5** | Ref. total |
| Reference Data | **MU_1** | 28 | 15 | 7 | 0 | 0 | 50 |
| | **MU_2** | 9 | 21 | 6 | 2 | 3 | 41 |
| | **MU_3** | 1 | 2 | 36 | 10 | 7 | 56 |
| | **MU_4** | 0 | 0 | 0 | 43 | 18 | 61 |
| | **MU_5** | 0 | 1 | 3 | 11 | 36 | 51 |
| | Map total | 38 | 39 | 52 | 66 | 64 | **259** |
| | **Overall Classification Accuracy = 63.32%** | | | | | | |

**Table 7-1: Example of an error matrix of a soil classification involving five map units.**

effectiveness of the use of error matrix stems from the fact that the accuracy of classification of each category or class is well described, along with both the errors of inclusion or commission and errors of exclusion or omission (Congalton, 1991 Jensen, 1996).

Three types of accuracy are typically determined from an error matrix. The first of these accuracies is the *overall accuracy*. In soil science, this will be the percentage of the reference soil units that were correctly classified through mapping. The overall accuracy is averaged over all map units identified in the mapping procedure. It does not indicate how the accuracy is distributed across the individual map categories [i.e., soil classes] (Story and Congalton, 1986). Fortunately, it is possible also to compute the accuracies of specific map units or soil classes from the error matrix of any classification scheme. These are termed the *producer's accuracy* and *user's accuracy*.

The producer's accuracy indicates how well members of a particular map category are classified. It is a measure of the errors of omission in a specific map unit, and/or indicates the propensity with which members of a particular soil class in the field were misclassified or placed in inappropriate map units. On the other hand, the user's accuracy indicates the probability that samples (e.g., pedons and polypedons) assigned to a particular map category (i.e., map unit) on the [soil]map actually represents the appropriate and expected category [i.e., soil type] in the reference data or on the ground. The user's accuracy is termed a measure of reliability and/or a measure of the errors of commission in specific map categories. The user's accuracy, more or less, measures the probability of encountering the inclusion of "dissimilar" soil types in a particular map unit. However, the use of error matrix also allows the confusion among soil classes to be further analyzed by showing the relative contributions of each class to the confusion (errors of omission or commission) found in a specific soil class.

Congalton (1991) and Jensen (1996) explain why it is important to report all three accuracies---the overall accuracy, producer's accuracy, and user's accuracy. As revealed in the preceding discussion, each of these accuracies conveys unique, yet complementary information about the reliability of the classification and mapping project. Campbell (1987) remarked that the overall accuracy may suggest the relative effectiveness of a classification, but does not form convincing evidence of the accuracy of the classification. The overall accuracy may be unduly high or low due to the ease or difficulty, respectively, of correctly identifying members of one or only a few specific map categories. The producer's accuracies of individual map units will show the relative tendencies for members of each of these map categories to be correctly identified, classified or mapped by the mapping methodology. However, the producer's accuracy alone will tell incomplete and perhaps misleading story about the effectiveness with which the classification scheme or mapping methodology can identify map categories. In soil survey and mapping for instance, a soil scientist who thinks that a certain soil type is "typical" in an area may, in the field, classify far more pedons into this expectedly predominant taxon than there actually are in reality. Similarly, it may be less likely that pedons which actually belong to such a typical taxon will be misclassified into other less "popular" or populous soil classes. In this instance, the producer's accuracy for the typical soil taxon will be high, but the user's accuracy will be low. This implies that although the soil survey person and mapping methods did a great job at correctly identifying soils in this soil class, they also had significant number of soils in other classes wrongly identified as belonging to this soil class or map unit.

In addition to the overall, producer's, and user's accuracies, more advanced statistical techniques can be used to further analyze the information contained in the error matrix. The advanced techniques include discrete multivariate statistics traditionally used (in social sciences) to analyze contingency tables. Congalton et al. (1983) introduced the use of these discrete multivariate methods in remote sensing for analysis of the accuracy of classification derived from satellite remotely sensed data. Since that time, these techniques have become adopted as the standard accuracy assessment tool (see Rosenfield & Fitzpatrick-Lins, 1986; Hudson & Ramm, 1987; Campbell, 1987; Congalton, 1991; Jensen, 1996). The reasons discrete multivariate methods are appropriate, and are preferred over parametric or normal theory statistics (e.g., analysis of variance) for the analysis of remotely sensed data are discussed in Congalton et al. (1983) and Congalton (1991). These include the facts that remotely sensed data are discrete rather than continuous, and are binomially or multinomially distributed rather than being normally distributed. Since these statements about the nature and distribution of remotely sensed data are also true for soils, it is reasonable to posit that these discrete multivariate techniques will also be suitable for analysis of the accuracies in soil classification and mapping projects. The application of these advanced analyses to an error matrix, yields two additional measures of the accuracy of classification called *normalized* accuracy and $K_{hat}$ statistics.

The main difference between the overall accuracy and the normalized accuracy stems from the way the two are computed. As shown in the succeeding section, the overall accuracy is computed by summing up the major diagonal cells of the matrix (this equals the total correctly classified samples), and dividing this sum by the total number of samples in the error matrix. Thus, the overall accuracy does not reflect information from the off-diagonal cells or the levels of the errors of omission and commission in the matrix. On the other hand, the normalized accuracy is computed after an iterative proportional fitting procedure called normalization or standardization, which forces each row and column in the matrix to sum to a unit or one. The normalization process involves the iterative balancing of the row and column cells, and the summation of these to form column and row totals or marginals. This changes the cell values along the major diagonal of the matrix in a way that forces these diagonal cell values to reflect the off-diagonal cell values also.

Normalized accuracy is nothing more than overall accuracy computed from a normalized or standardized error matrix. But unlike the overall accuracy, the normalized accuracy also incorporates the errors of omission and commission all together (Congalton, 1991). The normalized accuracy, it is argued, is a better representation of accuracy than is the overall accuracy computed from the original matrix (Jensen, 1996). Standardized or normalized error matrices have another advantage which will be of value in soil classification and mapping. Normalization provides a convenient way of comparing individual cell values between error matrices regardless of differences in the number of samples used to derive the matrices. Consider a situation where we want to evaluate the soil mapping skills of two trainee soil scientists, or we want to evaluate the relative performance of two soil mapping methodologies or techniques. After normalizing the error matrices generated in each pair of situations, each individual cell can be readily converted to a percentage by multiplying by 100, hence producing a single index by which corresponding cells could be compared. This is a better, and certainly a simpler alternative to comparing the producer's and user's accuracies of the corresponding cells from two or more matrices.

KAPPA analysis and the $K_{hat}$ statistic are not discussed fully in this research but interested readers should see one or more the references in this work. It may suffice to say that KAPPA analysis is a discrete multivariate technique, used to get another measure of the degree of agreement or accuracy ($K_{hat}$ ) in a classification matrix. The computation of the $K_{hat}$ statistic incorporates the off-diagonal elements (just as in normalized accuracy) as a product of the row and column marginals. The $K_{hat}$ statistic, therefore, is usually different from the overall accuracy, and the magnitude of the discrepancy would depend on the amount of errors of omission and commission included in the matrix. The $K_{hat}$ statistic is useful for 1) determining whether the results presented in an error matrix are significantly better than the result of randomly assigning samples to map categories (i.e., a null hypothesis of $K_{hat}$ = zero), and/or 2) comparing two matrices consisting of identical categories to see if they are significantly different from each other. My present position is that the rigor with which a soil sample is assigned to a soil class is much more than the effort required to assign a remote sensing pixel to a map category. Hence, it may rarely be necessary to test that a soil mapping and classification procedure produces results that are better than that of random assignment. And if such a hypothesis is tested, it is almost certain that it will be rejected. Also, although $K_{hat}$ can be used to statistically compare two matrices, an almost equally adequate but easier process may be to compare the normalized accuracies of such matrices. It is still not clear (even in remote sensing) if $K_{hart}$ statistic contains more information about the accuracy of classification than the normalized accuracy, and how the discrepancies between the $K_{hart}$ statistic and the overall and normalized accuracies should be interpreted. For these (and other) reasons, it may be expedient in this initial stage of adapting and applying the error matrix techniques to the evaluation of soil map quality, for attention not to be dissipated on the more challenging issues of KAPPA analysis and its attendant statistics.

## 7.4 Computing Soil Map Accuracies: An Example with STATSGO Data

The computation of the overall, producer's and user's accuracies is demonstrated by using the error matrices of the STATSGO data shown in Table 7-2. The accuracies and associated errors (commission and omission) in this data set are then used as the basis for inferring the probable quality or reliability of the soil classification in the STATSGO data of the northern New England area. Recall that Table 7-2 were developed from the discriminant function analysis discussed in Chapter 6. In that study, soil parent materials, texture of B-horizon, elevation exchangeable Ca, exchangeable acid, percent slope and Drainage information were used to [statistically]

(i)

| DFA Predicted Classification | STATSGO Classification | | | | | |
|---|---|---|---|---|---|---|
| | ME059 | NH017 | NH027 | ME064 | ME019 | DFA total |
| ME059 | 20 | 8 | 4 | 1 | 4 | 37 |
| NH017 | 2 | 16 | 8 | 0 | 0 | 26 |
| NH027 | 1 | 7 | 20 | 0 | 0 | 28 |
| ME064 | 3 | 0 | 0 | 29 | 15 | 47 |
| ME019 | 4 | 0 | 0 | 8 | 20 | 32 |
| STATSGO total | 30 | 31 | 32 | 38 | 39 | Grand Total 170 |
| Overall Classification Accuracy = 61.76% | | | | | | |

(ii)

| DFA Predicted Classification | STATSGO Classification | | | | | |
|---|---|---|---|---|---|---|
| | ME064 | ME019 | NH012 | NH022 | NH023 | DFA total |
| ME064 | 28 | 15 | 7 | 0 | 0 | 50 |
| ME019 | 9 | 21 | 6 | 2 | 3 | 41 |
| NH012 | 1 | 2 | 36 | 10 | 7 | 56 |
| NH022 | 0 | 0 | 0 | 43 | 18 | 61 |
| NH023 | 0 | 1 | 3 | 11 | 36 | 51 |
| STATSGO total | 38 | 39 | 52 | 66 | 64 | Grand Total 259 |
| Overall Classification Accuracy = 63.32% | | | | | | |

**Table 7-2: DFA error matrix of the classification accuracy of selected STATSGO map units in the study area.** (i) and (ii) were for tests involving map units with smaller and larger n sizes, respectively.

classify the FIA sampled soil units into the selected STATSGO map units. The discriminant function analysis procedure then used the resulting statistical classification as reference to asses how accurate STATSGO soil map was in predicting map unit membership. Hence, Table 7-2 is the result of accuracy assessment of soil classification by STATSGO of the study area, based on the selected soil variables (listed above), and the groups of five STATSGO map units.

To illustrate, let us think of the STATSGO soil map as a completed mapping project with only five map units. To evaluate the quality of this soil mapping project, the soil scientist would need to sample selected spots in each of the map units. The samples are selected by one of the various methods (see Wilding & Drees, 1983; Upchurch et al., 1988; Brown & Huddleston, 1991; Burrough, 1991) used in traditional studies of map unit composition. Table 7-2 shows that a total of 170 and 259 of such soil units were used for the first and second parts of the table. The columns in these error matrices represent how the sampled soil units were classified by the soil mapping project. Hence, the column totals indicate the number of observations the soil scientist made in each of the selected map units that (s)he intended to evaluate. The rows represent the actual, true classifications of the sampled soil units. The error matrix techniques presuppose that mutually exclusive and totally exhaustive system of classification was used in the mapping project. This means, therefore, that each of the sampled soil units must belong to one and only one of the map units identified during the classification. If this is true, then the row cells represent the way the reference data or the sampled soil units are distributed among the identified map units.

Table 7-2 (i) for instance shows that of the 30 soil units sampled from the STATSGO map unit ME059, only 20 of these could be verified by the soil scientist (using actual field or laboratory data or both) as actually belonging to this map unit. The numbers of the "correctly classified" samples in each of the map units make up the major diagonal cells (indicated also in boxes. Dividing each diagonal cell by the column total (sample size observed in that map unit) represents the user's accuracy for that map unit (20/30 = 66.7% for ME059). In soil science parlance, the user's accuracy is equivalent to percent map unit purity. It is a measure of reliability (Congalton, 1991) and/or indicates the probability that soils in a map represent the intended soil class on the ground. The difference between a major diagonal cell (or the user's accuracy) and the column total (or 100%) represent error of commission or rate of inclusion. Hence for ME059, the commission error is 30 minus 20, divided by 30 = about 33% or 100% minus 66.7%.

One of the advantages of the error matrix techniques is that they allow the commission error to be further analyzed to gain better understanding of the sources of this error. The off-diagonal column cells show how the inclusions are distributed among the other map units. Table 7-2 shows that the inclusion in ME059 are almost evenly distributed among the other map units. However, the two matrices consistently reveal that most of the inclusions in ME064 are soils that should actually be classified as ME019. With this type of information, the user of the soil survey can decide, given the intended use, if ME064 has sufficiently high user's accuracy, and whether or not ME064 and ME019 are significantly dissimilar to warrant concerns. This is a significant improvement over the traditional one-value map unit composition method.

The row total shows the number of sampled soil units that actually belong to a particular map unit. Hence, of the 170 soil units examined for the first error matrix, 37 of them were actually verified as ME059. Dividing the major diagonal by this row total (i.e., 20/37) gives the producer's accuracy (54.05% for ME059). The producer's accuracy tells the ease or difficulty with which members of a particular map category can be correctly identified or classified, and 100% minus the producer's accuracy is a measure of the omission error. The off-diagonal row cells show how the omitted portions of a map unit are distributed among other map units. This type of information is vital to a better understanding in pedology, and especially to the improvement of soil survey and mapping methodologies. Table 7-2 shows that almost all of the omissions in NH027 are put in NH017, and vice versa. This is not surprising given what is known about these two map units (see Table 6-1). But more importantly, this observation suggests that NH027 and NH017 should be re-examined to see if they are really distinct (theoretically or practically, or both). The user's and producer's accuracies, as well as the errors of omission and commission of the map units in Table 7-2 are thus:

**Table 7-2 (i) with 170 samples**

| Map units | User's Accuracy | Commission Error | Producer's Accuracy | Omission Error |
|---|---|---|---|---|
| ME059 | 20/30 = 66.7% | 33.30% | 20/37 = 54.1% | 45.90% |
| NH017 | 16/31 = 51.6% | 48.40% | 16/26 = 61.5% | 38.50% |
| NH027 | 20/32 = 62.5% | 37.50% | 20/28 = 71.4% | 28.60% |
| ME064 | 29/38 = 76.3% | 23.70% | 29/47 = 61.7% | 38.30% |
| ME019 | 20/39 = 51.3% | 48.70% | 20/32 = 62.5% | 37.50% |

**Table 7-2 (ii) with 259 samples**

| Map units | User's Accuracy | Commission Error | Producer's Accuracy | Omission Error |
|---|---|---|---|---|
| ME064 | 28/38 = 73.7% | 26.30% | 28/50 = 56.0% | 44.00% |
| ME019 | 21/39 = 53.8% | 46.20% | 21/41 = 51.2% | 48.80% |
| NH012 | 36/52 = 69.2% | 39.80% | 36/56 = 64.3% | 35.70% |
| NH022 | 43/66 = 65.2% | 34.80% | 43/61 = 70.5% | 29.50% |
| NH023 | 36/64 = 56.3% | 43.70% | 36/51 = 70.6% | 29.40% |

Finally, the overall accuracy is computed by summing all the major diagonal cells, and dividing by the total number of samples in the matrix. Hence for Table 7-2 (i), this will be 20 (for ME059) + 16 (for NH017) + 20 (for NH027) + 29 (for ME064) + 20 (for ME019) = 105, divided by 170 (the grand total) = 61.76% overall accuracy. Table 7-3 shows the results of normalizing the error matrices in Table 7-2. Recall that the normalization process forces the rows and columns to sum to one, and that the normalized accuracy is computed by dividing the sum of the major diagonals by the number of rows or columns. The normalized accuracies for the two matrices in

**(i)**

| | STATSGO Classification | | | | |
|---|---|---|---|---|---|
| | **ME059** | **NH017** | **NH027** | **ME064** | **ME019** |
| **ME059** | .590 | .170 | .083 | .044 | .113 |
| **NH017** | .123 | .562 | .268 | .025 | .021 |
| **NH027** | .072 | .250 | .632 | .025 | .021 |
| **ME064** | .073 | .007 | .007 | .632 | .282 |
| **ME019** | .142 | .011 | .010 | .275 | .563 |
| **Normalized Classification Accuracy = 59.57%** | | | | | |

**(ii)**

| | STATSGO Classification | | | | |
|---|---|---|---|---|---|
| | **ME064** | **ME019** | **NH012** | **NH022** | **NH023** |
| **ME064** | .616 | .288 | .091 | .004 | .003 |
| **ME019** | .284 | .552 | .108 | .024 | .033 |
| **NH012** | .050 | .072 | .685 | .113 | .079 |
| **NH022** | .024 | .020 | .013 | .664 | .277 |
| **NH023** | .026 | .068 | .103 | .195 | .607 |
| **Normalized Classification Accuracy = 62.47%** | | | | | |

**Table 7-3: Results of normalizing the matrices and classification accuracies in Table 7-2**

Table 7-2 are given in Table 7-3 as 59.57% and 62.49%. These are essentially the same as the overall accuracies of (61.76% and 63.32%, respectively) given on Table 7-2. Congalton (1991) explained that the overall and normalized accuracies tend to disagree when the original matrix has great many off-diagonal cell values of zero---a situation which happens when the matrix is constructed with insufficient sample size or the classification is exceptionally good. Since none of these situations was true for Table 7-2, the equivalence between the overall and normalized accuracies in this study was not surprising. It is tedious to carry out normalization by hand, but Congalton (1983) has written an easy to use computer program (available on request) for this purpose.

## 7.5 Summary and Conclusion

The error matrix and associate statistics are state-of-the-art techniques for comprehensive assessment of classification or map accuracies. These techniques have been in use in remote sensing. The main idea of this paper is that the use these techniques can also be adapted to improve the map unit composition method of assessing soil map quality. The primary objective of the paper was to introduce the concepts of error matrix techniques to the soil science community, and demonstrate how and why these new techniques can be applied to soil survey and mapping. The different types of accuracies derivable from the error matrix were discussed, and their practical implications in soil mapping were demonstrated with real data.

Some of the advantages the use of error matrix in the analysis of soil map quality were highlighted. First, the new techniques are simple to understand and use.

Second, the procedure parallels those which the soil scientists have traditionally used in assessing map unit composition. However, the new techniques overcome major limitations of the present method of assessing map-unit purity. For instance, map-unit purity is usually analyzed for one map unit at a time, and if two or more map units are involved, the present method does not provide a means for comparing them. But with the use of an error matrix, as many map units as are present in the soil map or as resource availability allows could be simultaneously analyzed. In addition to the overall accuracy of classification, the producer's accuracy and errors of commission, and the user's accuracy and errors of omission can also be computed. When there is confusion in discriminating among soil classes (either high omission or commission errors), error matrix can be used to effectively show what specific soil classes are confused, and the relative proportion of that soil class that is incorrectly assigned to each of the other map units. This kind of information enables the users to better interpret soil survey data, and allows soil classification and mapping methodology to be improved with time. The use of the error matrix is also more practical and much less costly than the method that requires laboratory measurements of a plethora of individual soil properties.

It seems clear that these new techniques are potentially the solution to the age-long need for an effective, quantitative method of evaluating and communicating soil map quality. It is my hope, therefore, that this research (upon publication) will stimulate the interests of more experienced soil scientists, and generate further investigation and discussion on the necessary considerations that will enable these novel techniques to be optimally applied in soil classification and mapping. An ample bibliography on accuracy assessment in remote sensing, the error matrix and related concepts (such as appropriate sampling schemes) is included in the study for this reason.

# BIBLIOGRAPHY

Aber, J. D., C. Driscoll, C.A. Federer, R. Lathrop, G. Lovett, J. M. Melillo, P. Steudler and J. Vogelmann. 1993. A Strategy for the Regional Analysis of the Effects of Physical and Chemical Climate change on Biogeochemical cycles in Northeastern (US) Forests. *Ecological Modeling*, 7:37-47.

Aber, J.D. and J. M. Melillo. 1991. *Terrestrial Ecosystems.* Saunders College Publishing, Philadelphia.

Amos, D. F., and E. P. Whiteside. 1975. Mapping Accuracy of a Contemporary Soil Survey in An Urbanizing Area. *Soil Sci. Soc. Am. Proc.* 39:937-942.

Arkley J. Rodney. 1976. Statistical Methods in Soil Classification Research. *Advances in Agronomy* **28**:37-70

Arnold, W. R. 1983. Concepts of Soils and Pedology. p.1-21. *In* L. P. Wilding et al. (ed) *Pedogenesis and Soil Taxonomy. I. Concepts and Interactions.* Elsevier Sci. Pub., New York.

Arnold, R. W., and L. P. Wilding. 1991. The Need to Quantify Spatial Variability. p. 1-8. *In* M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms.* SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Aronoff, Stan. 1993. *Geographic Information Systems: A Management Perspective.* WDL Publications. Ontario, Canada.

Avers P.E., Cleland, D.T., McNab, W.H., Jensen, M.E, Bailey, R.G., King, T., Goudey, C.B. and Russell, W.E, 1994. *National Hierarchical Framework of Ecological units.* USDA-Forest Service, Washington, DC, 15 pp.

Barrett, J. P. and M. E. Nutt. 1979. *Survey Sampling in the Environmental Sciences: A Computer Approach.* Compress, Inc., Wentworth, NH (USA)

Bascomb, C. L., and M. G. Jarvis. 1976. Variability in three areas of the Denchworth soil map unit. 1. Purity of the map unit and property variability within it. *Journal of Soil Sci.* 27:420-437

Beckett, P.H.T. and R. Webster. 1971. Soil Variability: A Review. *Soils and Fertilizers*, 34:1-15.

Bell, C. J., L. R. Cunningham, and M. W. Havens. 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. *Soil Sci. Soc. Am. J.* 56:1860-1866

Bie, W. S. and P. H. T. Becket. 1971. Quality Control in Soil Survey Introduction: I. The Choice of Mapping Unit. *Journal of Soil Science*, Vol. 22, No. 1, 1971

Bigler, R. J., and K. J. Liudahl. 1984. Estimating Map Unit Composition. *Soil Survey Horizon.* 25:21-25

Birdsey, R. A. and H. T. Schreuder. 1992. An Overview of Forest Inventory and Analysis Estimation and Procedures in the Eastern United States- with an Emphasis on the components of Change. USDA Forest Service, Rocky Mtn. Forest and Range Expt. Station, Fort Collins, Colorado. General Tech. Report RM-214.

Bliss, N.B. and W.U. Reybold. 1989. Small-scale Digital Soil maps for Interpreting Natural resources. Journal of Soil and Water Conservation. 44(1):30-34

Blyth, J. F., and D. A. MacLeod. 1978. The Significance of Soil Variability for Forest Soil Studies in North East Scotland. *J. Soil Sci.* 29:419-430.

Borland International, 1994. Paradox For Windows, Version 5.x, Users Guide. Borland International, CA.

Buol, S. W, F. D. Hole and R. J. McCracken, 1989. Soil Genesis and Classification, 3rd edition, Iowa State University Press, Ames, 446pp.

Bouma, J. 1988. When the mapping is over, then what?. p. 3-12. In Proc. Int. Interactive Workshop on Soil Resources: Their Inventory, Analysis, and Interpretation for Use in the 1990's, St. Paul, MN. 22-24 Mar. 1988. Univ. of Minnesota Coop. Ext. Serv., St. Paul, MN.

Bowers, S. D. 1988. *From Data to Database*. Van Nostrand Reinhold, UK.

Brown, R. B. and J. H. Huddleston. 1991. Presentation of Statistical data on Map Units to the uses. p. 127-147. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms.* SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Brubaker, C. S., and C. T. Hallmark. 1991. A comparison of Statistical Methods for Evaluating Map Unit Composition. p.73-87. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms.* SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Burgess, T. M. and R. Webster. 1980a. Optimal Interpolation and Isarithmic Mapping of Soil Properties: 1. The Semi-variogram and Punctual Kriging. *J. Soil Science*, 31:315-331

Burgess, T. M. and R. Webster. 1980b. Optimal Interpolation and Isarithmic Mapping of Soil Properties: II. Block kriging. *J. Soil Science*, 31:333-341.

Burrough, P.A. 1987. *Principles of Geographical Information Systems for Land Resources Assessment.* Oxford Press, N. Y.

Burrough, P.A. 1991. Sampling Designs for Quantifying Map Unit Composition. p. 89-125. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI

Burrough, P. A., and R. Webster. 1976. Improving A Reconnaissance Classification By Multivariate Methods. *J. Soil Science*, 27:554-571

Butler, B. E. 1980. *Soil Classification for Soil survey.* Oxford Univ. Press, Ney York

Cambardella, C.A., T. B. Moorman, J. M. Novak, T.B. Parkin, D. L. Karlen, R.F. Turco, and A. E. Konopka. 1994. Field-Scale Variability of Soil Properties in Central Iowa Soils. *Soil Sci. Soc. Am. J.* 58:1501-1511

Campbell, J. B. 1978. Spatial Variation of sand Content and pH within Single Continuous Delineation of Two Soil Mapping Units. *Soil Sci.Soc. Am. J.* 42:460-464.

Campbell, J. B. 1987. *Introduction To Remote Sensing*, Guilford, New York, 1987

Chen, J., J. W. Hopmans, and G. E. Fogg. 1995. Sampling Design for Soil Moisture Measurements in Large Field Trials. *Soil Science*, 159 (3):155-161.

Cipra, J.E., O.W. Bidwell and F.J. Rohlf. 1970. Numerical Taxonomy of Soils from Nine Orders by Cluster and Centroid-Component Analyses. Soil Sci. Soc. Am. Proc. 34:281-287

Congalton, R. G. 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 54(5):593-600

Congalton, R. G. 1991. A review of assessing the accuracy of classification of remotely sensed data. *Remote Sens. Environ.* 37:35-46

Congalton, R. G. 1993. Accuracy assessment of remotely sensed data: future needs and directions. *Proceedings of Pecora 12*, Sioux Falls, SD August 24-26. 1993.

Congalton, R., and Green, K. 1993. A practical look at sources of confusion in error matrix generation. *Photogrammetic Engineering and Remote Sensing*, 59:641-644

Congalton, R., and Mead, R. A. 1983. A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetic Engineering and Remote Sensing*, 49(1):69-74

Congalton, R. G., Oderwald, R. G., and Mead, R. A. 1983. Assessing landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetic Engineering and Remote Sensing* 49(12):1671-1678

Crepin, J. and R. L. Jothnson. 1993. Soil Sampling for Environmental Assessment. *In* M.R. Carter (ed.) *Soil Sampling and Methods of Analysis*. Canadian Society of Soil Science. Lewis Publishers.

Corns, I.G.W., and R. M. Annas. 1986. *Field Guide to Forestry Ecosystems of West-central Alberta*. Can. For. Serv., Nortn. For. Cent., Edmonton, Alberta. 250p.

D'Agostino, R. B. 1986. Tests for the Normal Distribution, p. 367-419. In: R. B. D'Agostino and M. A. Stephens (eds.), *Goodness-of-fit Techniques*. Marce Dekker, New York.

Day, R. L., G. W. Petersen, and D. A. Miller. 1988. The Base Map Dilemma: Solution Through Digital Cartography and Remote Sensing. p. 40-50. In Proc. Int. Interactive Workshop Soil Resources: Their inventory, analysis and interpretation for use in the 1990's. 22-24 Mar. 1988. Edu. Dev. System, Minnesota Ext. Serv., Univ. Minnesota, St. Paul.

Diekkruger, B., D. Sondgerath, K.C. Kersebaum and C.W. McVoy. 1995. Validity of agroecosystem models: A comparison of results of different models applied to the same data set. *Ecological Modelling* 81:3-29

Duning, X, R.H. Rust and J.R. Crum. 1986. Numerical Classification of Forested Soils in the High Mountain Region of Southwestern China. Soil Science 141 (2):127-137

ECOMAP. 1993. *National Hierarchical Framework of Ecological units*. Unpublished administrative paper. USDA-Forest Service, Washington, DC, 20 p.

Edmonds, W. J., and M. Lentner. 1985. Statistical evaluation of the taxonomic composition of three soil map units in Virginia. *Soil Sci. Soc. Am. J.* 50:997-1001

Fitzgerald, R. W. And Lee, B. G. 1994. Assessing the classification accuracy of multisource remote sensing data. *Remote Sens. Environ.* 47:362-368.

Gaddum, J. H. 1945. Lognormal Distributions. *Nature* (London) 156: 463-466

Geary, R. C. 1947. Testing for normality. *Biometrica* 34: 209-242.

Grigal, D. F., R. E. McRoberts and L. F. Ohmann. 1991. Spatial variation in chemical properties of forest floor and surface mineral soil in the North Central United States. *Soil Science*, Vol. 151, No. 4, pp. 282-290.

Hammer, D. R., J. H. Astroth, Jr., G. S. Henderson, and Fred J. Young. 1991. Geographic Information Systems for Soil Survey and Land-use Planning. p. 243-270 *In* M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms.* SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Hansen, M. H., T. Frieswyk, J. F. Glover, and J. F. Kelly. 1992. The Eastwide Forest Inventory Data Base: Users Manual. USDA Forest Service, North Central Forest Expt. Station General Technical Report NC-151.

Hartung, S. L., S. A. Scheinost, and R. J. Ahrens. 1991. Scientific Methodology of The National Cooperative Soil Survey. p.39-48. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms.* SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Hay, G. H. 1979. Sampling designs to test land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*, 45(4):529-533

Heuvelink, G .B. M., P. A. Burrough, and A. Stein. 1989. Propagation of errors in spatial modeling with GIS. *Int. J. Geogr. Inf. Syst.* 3(4):303-322

Holmgren, G. G. 1988. The point representation of soil. *Soil Sci. Soc. Am. J.* 52:112-716

Hopkins, D. L., J. L. Richardson, and M. D. Sweeney. 1987. Composition Comparisons in Sodic Map Unit Delineations on the Dickinson Experiment Station Ranch Headquarter, North Dakota. *Soil Survey Horizon.* 28:46-50

Horton, I. F., Russell, J. S. And Moore, A. W. 1968. Multivariate-covariance and canonical analysis: a method for selecting the most effective discriminators in a multivariate situation. *Biometrics* 24:845-858

Hudson, B. D. 1980. Ranges of characteristics---How valid are they. *Soil Survey Horizon.* 21:3, 7-11

Hudson, B. D. 1990. Concepts of soil mapping and interpretation. *Soil Survey Horizon*, 31:63-73

Hudson, W. D. And Ramm, C. W. 1987. Correct formation of the Kappa Coefficient of Agreement. *Photogrammetric Engineering and Remote Sensing*, 53(4):421-422.

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

Jenny, H., 1941. Factors of Soil Formation -- a system of quantitative pedology. McGraw-Hill, New York, N. Y., 281pp.

Jensen, R. John. 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective*, 2nd edition. Prentice Hall, Inc. Upper Saddle River, New Jersey (USA).

John, R.A. and D. W. Wichern. 1982. Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, NJ. 594 p.

Jordon, S. K., Liston, R. L. and G. W. Tesselle. 1986. Recompiling Published Soil Survey Maps on Planimetric Base Maps Using Digital Techniques. Internal Report #45. USDA Soil Conservation Service, Cartographic and Geographic Information Systems Division, Washington, D.C.

Keys, Jr., J., Carpenter, C., Hooks, S., Koenig, F., McNab, W.H., Russell, W., Smith, M.L. 1995. Ecological Units of the Eastern United States - First Approximation 1995 (Subsections Map, booklet of Map Unit Tables, and CD-ROM of GIS coverage in ARCINFO format). Atlanta, GA: USDA Forest Service

Klecka R. William. 1980. Discriminant Analysis. 71 pp. Sage university paper # 19. Sage Publications, Newbury Park.

Kosaki Takashi and A. S. R. Juo. 1989. Multivariate Approach to Grouping Soils in Small Fields. I. Extraction of factors causing soil variation by Principal Component Analysis. *Soil Science and Plant Nutrition*, 35 (3):469-477

Lammers A. Duane, and Mark G. Johnson. 1991. Soil Mapping Concepts for Environmental Assessment. p. 149-159. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Lanfear, J. Kenneth. 1989. Geographic Information Systems and Water Resources Application (Guest Editorial), *Water Resources Bulletin*, Vol. 25 (No. 3): v-vi.

Lathrop, G.R., J.D. Aber, and J.A. Bognar. 1995. Spatial Variability of Digital Soil Maps and its Impact on Regional Ecosystem Modeling. Ecological Modeling 82:1-10

Leenhardt, D., M. Voltz, M. Bornand and R. Webster. 1994. Evaluating Soil Maps for Prediction of Soil Water Properties. European Journal of Soil science 45:293-301.

Lillesand, T. M. and R. W. Kiefer. 1994. *Remote Sensing and Image Interpretation*, third Edition. John Wiley and Sons. New York, New York.

Little, I. P., Horton, I. F., Haydock, K. P. And Paton, T. R. 1968. The use of canonical analysis of chemical data to distinguish among materials of the valley fill of the Bremer River, South-eastern Queensland. *Australian Journal of Science* 31:86-88

Lunetta, R., R. Congalton, L. Fenstermaker, J. Jensen, K. McGwire, and L. Tinney. 1991. Remote Sensing and Geographic Information system data integration: Error sources and research issues. *Photogrm. Engineering and Remote Sensing.* Vol 57 ( No. 6) 677-687

Lytle J. Dennis. 1993. Digital Soils Databases for the United States, p 386 - 391. *In*: Goodchild, F.M., B.O. Parks and L.T. Steyaert (eds.) Environmental Modeling with GIS. Oxford University Press, Ney York. p. 488.

Mader, D. L.. 1963. Soil Variability---A Serious Problem in Soil-Site Studies in the Northeast *Soil Sci. Soc. Am. Proc.* 27:707-709

Markewich, W. H., and Cooper C. S. 1991. One Perspective on Geologic Mapping. p.25-37. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

McBratney, A. B., and R. Webster. 1983. How Many Observations are Needed for Regional Estimation of Soil Properties?. *J. Soil Sc.* 135:177-183

McBratney, A. B., R. Webster and T. M. Burgess. 1981. The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables. I. Theory and Method. *Computer and Geosciences* 7:331-443.

McBratney, A. B., R. Webster, R. G. McLaren, and R. B. Spiers. 1982. Regional variation of extractable copper and cobalt in the top soil of south-east Scotland. *Agronomie* 2:969-982.

McBratney, A. B., G.A. Hart and D. McGarry. 1991. The Use of Region Partitioning to Improve the Representation of Geostatistically mapped Soil Attributes. *Journal of Soil Science,* 42:513-532.

McCormack, D. E., and L. P. Wilding. 1969. Variation of soil properties within mapping units of soils with contrasting substrata in northwestern Ohio. *Soil Sci. Soc. Am. Proc.* 33:587-593

McNab, W.H and Avers, P.E. (Eds.). 1994. *Ecological Subregions of the United States: Section Description*. USDA Forest Service WO-WSA-5. Washington, DC. 337 pp.

Miller, F. P. 1978. Soil survey under pressure: The Maryland experience. *J. Soil Water Conserv.* 33:104-111

Miller, F. P., D. E. McCormack, and J. R. Talbot. 1979. Soil Surveys: Review of data collection methodologies, confidence limits, and uses. Transp. Res. Rec. 733. NAS, Trans. Res. REc. Board, Washington, p. 57-65

Moore, I. D., P. E. Gessler, G. A. Nielsen, and G. A. Peterson. 1993. Soil Attribute Prediction Using Terrain Analysis. *Soil Sci. Soc. Am. J.* 57:443-452

Nash, M. H., and L. A. Daugherty. 1990. Statistical Comparison of soil Map-units Boundaries. *Soil Sci. Soc Am. J.* 54:1677-1681

Nettleton, W. D., B. R. Brasher, and G. Borst. 1991 The Taxajunct Problem. *Soil Sci. Soc Am. J.* 55:421-427

Nielsen, D.R., J. W. Biggar and K. T. Erh. 1973. Spatial variability of field-measured soil –water properties. *Higardia*, 42:215-259

Nordt, C.L., John S. Jacob, and L. P. Wilding. 1991. Quantifying Map Unit Composition for Quality Control in Soil Survey. p. 183-197. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI

Norman, R. G. and D. L. Streiner. 1997. *PDQ Statistics* (2nd Edition). Mosby, Baltimore.

Norris, M. J. 1970. Multivariate Methods in the Study of Soils. *Soils and Fertilizers* Vol. 33:313-319

Norris, M. J., and J. Loveday. 1971. The application of multivariate analysis to soil studies. II. The allocation of soil profiles to groups: A comparison of soil survey and computer methods. *J. Soil Sci.* 22:395-340

Norusis, J. M. 1990. *SPSS Advance Statistics Student Guide*. SPSS Inc., Chicago.

Oberthur, T., A. Dobermann and H.U. Neue. 1996. How Good is a Reconnaissance Soil Map for Agronomic Purposes?. Soil use and Management 12:33-43.

Odeh, I. O. A., D. J. Chittleborrough and A. B. McBratney. 1991. Elucidation of soil-landform Interrelationships by Canonical Ordination Analysis. *Geoderma* 49:1-32

Odeh, I.O.A., A. B. McBratney and D. J. Chittleborough. 1990. Design of optimal spacings for mapping soil using fuzzy-k-means and regionalized variable theory. *Geoderma* 47:93-122

Parkin, T.B. and J. A. Robinson. 1992. Analysis of Lognormal data. *Advances in Soil Science* 20:193-235.

Parkin, T.B., J.J. Meisinger, S.T. Chester, J. T. Starr, and J.A. Robinson. 1988. Evaluation of Statistical Estimation Methods for Lognormally Distributed Variables. *Soil Sci. Soc. Am.* J.52:323-329

Pavlick, H. F., and F. D. Hole. 1977. Soilscape analysis of slightly contrasting terrains in southeastern Wisconsin. *Soil Sci. Soc. Am. J.* 41:407-413.

Peck, J. A. 1983. Field Variability of Soil Physical Properties. *Advances in Irrigation*, Volume 2: 189-221.

Peck, J. A., R.J. Luxmoore, and J.L. Stolzy. 1977. Effects of spatial variability of soil hydraulic properties in water budget modeling. *Water Resources Research*, 13:348-354

Powell, J.C. and M.E. Springr. 1965. Composition and precision of classification of several mapping units of the Appling, Cecil and Llyold series in Walton County, Georgia. *Soil Sci. Soc. Am. Proc.* 29: 454-458

Pregitzer, K.S., and B.V. Barnes. 1984. Classification and Comparison of Upland Hardwood and Conifer Ecosystems of the Cyrus H. McCormick Experimental Forest, Upper Michigan. *Can. J. For. Res.* 14:362-375

Protz, R., E. W. Presant, and R. W. Arnold. 1968. Establishment of the modal profile and measurement of variability within a soil landform unit. *Canadian J. of Soil Sci.* 48:7-19

Ragg, M. J., and R. Henderson. 1980. A reappraisal of soil mapping in an area of southern Scottland. Part I The reliability of four soil mapping units and the morphological variability of other dominant taxa. *Journal of Soil Science* 31:559-572

Reybold, U.W and G.W. TeSelle. 1989. Soil Geographic Data Bases. Journal of Soil and Water Conservation. 44(1):28-29.

Rogowski, A. S., and J. K. Wolf. 1994. Incorporating Variability into Soil Map Unit Delineations. *Soil Sci. Soc. Am. J.* 58:163-174

Rosenfield, G. H., and Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetic Engineering and Remote Sensing*, 52(2)223-227.

Royle, A. G. 1980. *Geostatistics*. McGraw-Hill, New York.

Seeling, D. B., J. L. Richardson and R. E. Knighton. 1991. Comparison of Statistical and Standard Techniques to Classify and Delineate Sodic Soils. *Soil Sci. Soc. Am. J.* 55:1042-1048

Smith, M.L., and Carpenter, C. 1996. Application of USDA Forest Service National Hierarchical Framework of Ecological Units at the Subregional Level: The New England-New York Example. *Environmental Monitoring and Assessment*. 39:187-198

Smith, M. L., and R. A. Hallett. 1993. Regional Ecological Analysis Using RDBMS and GIS Technology: An example from Northern New England (Poster). National Center for Geographic Information and Analysis Biennial Symposium, Breckenridge, Co.

Soil Conservation Service. 1991. State Soil Geographic Data Base (STATSGO): Data Users Guide. USDA SCS Misc. Pub. Number 1492

Soil Science Society of America (SSSA). 1987. *Glossary of Soil Science* Terms. 44pp. Soil Science Society of America, Madison, WI

Soil Survey Staff. 1951. *Soil Survey Manual*. USDA-SCS Handbook 18. US Gov. Print. Office, Washington, D.C.

Soil Survey Staff. 1975. Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys. USDA-SCS Agric. Handbook No. 436, US Govt. Print. Office, Washington, D.C.

Soil Survey Staff. 1980a. *Soil Survey Manual* (Revised). USDA-SCS Handbook 18. US Gov. Print. Office, Washington, D.C.

Soil Survey Staff. 1980b. USDA-SCS. National Soils Handbook 430-VI. US Gov. Print. Office, Washington, D.C.

Soil Survey Staff. 1983. USDA-SCS. National Soils Handbook . US Gov. Print. Office, Washington, D.C.

Steers, C. A., and B. F. Hajek. 1979. Determination of Map Unit Composition by Random Selection of Transects. *Soil Sci. Soc. Am. J.* 43:156-160.

Stein, A., M. Hoogerwert, and J. Bouma. 1988. Use of Soil Map Delineations to Improve (co)Kriging of Point Data on Moisture Deficits. *Geoderma* 43:163-177

Stehman, S. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetic Engineering and Remote Sensing*, 58:1343-1350

Story Michael, and Congalton Russell. 1986. Accuracy assessment: A user's perspective. *Photogrammetic Engineering and Remote Sensing*, 52(3):397-399.

Tabachnick Barbara, and Linda Fidell. 1996. *Using Multivariate Statistics*. Third Edition. 880pp. Harper & Row Publishers, New York.

Taylor, W. Robert. 1987. Geographic Variation in Chemical Properties of New England Forest Soils: Final Report submitted to the Northeastern Forest Exp. Station Forest Service, September 1984 to September 1987.

Thoni, H. 1967. *Transformation of Variables Used in the Analysis of Experimental and Observational Data*: A Review. Tech. Rep. No. 7, Statistical Lab, Iowa State Univ., Ames, Iowa. 61 pp.

Tisdale, S.L., W. L. Nelson and Beaton. 1985. *Soil Fertility and Fertilizers*. Macmillan Publishing Company, New York.

Trangmar, B. B., R. S. Yost and G. Uehara. 1985. Application of Geostatistics to Spatial Studies of Soil Properties. p. 45-94. In N.C. Brady (ed.) *Advances in Agronomy* Vol. 38. Academic Press, Inc. Orlando.

Troedsson, T. and C. O. Tamm. 1969. Small-scale Spatial Variation in Forest Soil Properties and its implications for Sampling Procedures. *Stud. For. Succ.* 74:4-30

Uehara, G., B. B. Trangmar and R. S. Yost. 1984. Spatial variability of soil properties. p. 61-92. In D. R. Nielsen and J. Bouma (eds.) *Soil Spatial Variability*. Proceedings of A Workshop of the ISSS and SSSA, Las Vegas, USA, 30 November-1 December 1984

Upchurch, D. R., L. P. Wilding, and J. L. Hatfield. 1988. Methods to evaluate spatial variability. p. 201-229. In L. R. Hossner (ed.) *Reclamation of surface-mined lands*. Vol. II. CRC Press, Inc., Boca Raton, FL.

Upchurch, D.R., and William J. Edmonds. 1991. Statistical Procedures for Specific Objectives. p. 49-71. In M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI

USDA-Soil Conservation Service, 1993. State Soil Geographic Data Base (STATSGO) Data Users Guide. USDA-SCS Miscellaneous Publication Number 1492

USDA-Forest Service. 1980. Resources Evaluation Field Instructions for MAINE 1980-81. Renewable Resources Evaluation Research Work Unit, Northeastern Forest Experiment Station, Broomall, PA.

USDA-Forest Service. 1982. Resources Evaluation Field Instructions for VERMONT and NEW HAMPSHIRE 1982 Renewable Resources Evaluation Research Work Unit, Northeastern Forest Experiment Station, Broomall, PA.

Van Wambeke, A. and Dudal, A., 1978. Macrovariability of Soils of the Tropics. in: Diversity of soils in the Tropics. *Soil Sci. Soc of Am. special publication* 34: 13-28.

Vieira, S. R., J. L. Hatfield, D. L. Nielsen, and J. W. Biggar. 1983. Geostatistical theory and application to variability of some agronomical properties. *Hilgardia* 51:1-75.

Wambeke A. 1982. Soil Taxonomy: A Technical Language of Soil Science. *Agronomy Mimeo #83-13*. Agronomy Dept., Cornell University, Ithaca, New York.

Wambeke, A. and Dudal, A., 1978. Macrovariability of Soils of the Tropics. in: Diversity of soils in the Tropics. *Soil Sci. Soc of Am. special publication 34*: 13-28.

Wang, M. and Howarth, P. J. 1993. Modeling errors in remote sing image classification. *Remote Sens. Environ.* 45:261-271

Warrick, A. W., and D. R. Nielsen. 1980. Spatial Variability of Soil Physical Properties in the field. p. 319-344. In D. Hillel (ed.) *Application of soil physics*. Academic Press, New York.

Way, S. Douglas, 1985. Terrain Analysis: A guide to site selection using aerial photo interpretation, 2nd edition. Graduate School of Design, Harvard University, 438pp.

Webster, R. 1977. *Quantitative and Numerical methods in Soil Classification and Soil survey*. Clarendon, Oxford.

Webster, R. 1978. Mathematical treatments of soil formation. P. 161-190. Trans. Int. Congr. Soil Sci., 11th Edmontory Alberta. 19-27 July 1978. Vol. 3. Univ. Of Alberta, Edmonton.

Webster, R. 1979. Soil Survey: Its Quality and Effectiveness. *Cornell Univ. Agron. Mimeo. No.79-12.* p. 52-63. Cornell University Ithaca, New York.

Webster, R. 1985. Quantitative Spatial Analysis of Soil in the Field. p. 1-70. In B.A. Steward (ed.) *Advances in Soil Science* 3. Springer-Verlag, New York.

Webster, R. and P. A. Burrough. 1974. Multiple Discriminant Analysis in Soil Survey. J. *Soil Science*, vol 25:120-134

Webster, R. and M. A. Oliver, 1990. Statistical methods in Soil and Land Resource Survey. Oxford University Press, New York

Webster, R. and M. A. Oliver. 1992. Sample adequately to estimate variograms of soil properties. *J. Soil Science*, 43:177-192.

Weismiller, R.A., I.D. Persinger, and O.L. Montgomery, 19977. Soil Inventory from Digital Analysis of Scanner and Topographic Data. *Soil Sci. Soc. Am. J.* 41:1166-1170

West, L. T., E. D. Bearden, and D. L. Williams. 1981. Taxonomic and Interpretive Composition of Two Map Units in Central Texas. p. 297. In Agron. Abstracts. ASA Madison, WI.

Wierenga, J. P. 1984. Spatial variability of soil-water properties in irrigated soils. p. 112-125. In D. R. Nielsen and J. Bouma (eds.) *Soil Spatial Variability.* Proceedings of A Workshop of the ISSS and SSSA, Las Vegas, USA, 30 November-1 December 1984

Wilding, L.P., R. B. Jones and G.M. Schafer. 1965. Variation of soil morphological properties within Miami, Celina and Crosby mapping units in west-central Ohio. *Soil Sci. Soc. Am. Proc.* 29: 711-717

Wilding, L. P., 1984. Spatial variability: its documentation, accommodation and implication to soil surveys. p. 166-189. In D. R. Nielsen and J. Bouma (eds.) *Soil Spatial Variability.* Proceedings of A Workshop of the ISSS and SSSA, Las Vegas, USA, 30 November-1 December 1984

Wilding, L. P. 1988. Improving our understanding of the composition of the soil-landscape. p. 13-35. In Proc. Int. Interactive Workshop Soil Resources: Their inventory, analysis and interpretation for use in the 1990's. 22-24 Mar. 1988. Edu. Dev. System, Minnesota Ext. Serv., Univ. Minnesota, St. Paul.

Wilding, L. P., and L. R. Drees. 1983. Spatial Variability and Pedology. p. 83-116. In L. P. Wilding et al. (ed) *Pedogenesis and Soil Taxonomy. I. Concepts and Interactions*. Elsevier Sci. Pub., New York.

Witty, J. E., and R. W. Arnold. 1987. Soil Taxonomy: An overview. *Outlook Agric.* 16:8-13

Yates, S. R. and A. W. Warrick. 1987. Estimating soil water content using co-kriging. Soil Sci. Soc. Am. J., 51:23-30

Young, J. F., J..M. Maatta and R. Hammer. 1991. Confidence Interval for Soil Properties Within Map Units. p. 213-229. *In* M.J. Mausbach and L.P. Wilding (ed.) *Spatial Variabilities of Soils and Landforms*. SSSA Special Pub. 28. Soil Sc. Soc. of America, Inc. Madison, WI.

Yost, S. R., G. Uehara, and R. L. Fox. 1982a. Geostatisctical analysis of soil chemical properties of large land areas. I. semi-variograms. *Soil Sci. Soc. Am. J.* 46:1028-1032.

Yost, S. R., G. Uehara, and R. L. Fox. 1982b. Geostatistical analysis of soil chemical properties of large land areas. II. kriging. *Soil Sci. Soc. Am. J.* 46:1033-1047.

Zar, H. J. 1996. *Biostatistical Analysis* (third edition). Prentice Hall, Upper Saddle River, New Jersey.

Zelazny, V.F., M.G. Hayter, C.L. Bowling, and D.A. Bewck. 1989. *Field Guide to Forest Site Classification in Brunswick: Napadogan-Tobique Site Region*. N.B. Dept. Natural Resources and Energy. Fredericton, NB. 43p.