The ESAP User Manual and Tutorial Guide

Version 1 .O

Scott M. Lesch, James D. Rhoades, David J. Strauss,
Kennith Lin, and Mark Allan A. Co

USSL Research Report No. 138

April, 1995

U. S. Salinity Laboratory
Agricultural Research Service
United States Department of Agriculture
Riverside, California

DISCLAIMER

The information in, and/or ESAP software associated with this document has been funded and developed by the United States Department of Agriculture, Agriculture Research Service, at the U. S. Salinity Laboratory. This document (user manual) has been subjected to the ARS peer and administrative review, and has been approved for publication as an internal U. S. Salinity Laboratory research report. The ESAP software associated with this document is to be considered public domain software, and as such may be used and copied free of charge.

Although the authors of this software have endeavored to produce accurate and error free program code, this software (including instructions for its use) is provided “as is” without warranty, expressed or implied. Furthermore, neither the authors nor the United States Department of Agriculture warrant, guarantee, or make any representations regarding the use, or the results of the use of, or instructions for use of this software or manual in terms of applicability, reliability, accuracy, or correctness. The use and application of this software and manual is the sole responsibility of the user.

The mention of any trade names or commercial products is for the convenience of the user and does not imply any particular endorsement by the United States Department of Agriculture or its agents. The Geonics EM-38 Meter ® is a registered trademark of Geonics Limited.

TECHNICAL ABSTRACT

Lesch, S. M., J. D. Rhoades, D. J. Strauss, K. Lin, and M. A. A. Co. 1995. The ESAP User Manual and Tutorial Guide, Version 1 .O. Research Report No. 138, U. S. Salinity Laboratory, United States Department of Agriculture, ARS, Riverside, California.

In this report we describe and document the ESAP software package; a system of integrated, user-friendly programs designed to model and predict field scale spatial soil salinity patterns from electromagnetic induction (EM) survey readings. The ESAP software package consists of four programs which can perform the following functions; (1) analyze and validate EM signal data, (2) identify the locations of a minimal subset of EM survey sites for soil salinity sampling, (3) transform the acquired EM signal readings into predicted salinity data throughout the entire survey area using the sample soil salinity data in conjunction with spatially based statistical modeling techniques, and (4) generate report quality maps of the predicted spatial salinity pattern, by depth, for the survey area. The use and application of each program is discussed in detail. Multiple tutorial examples are included throughout this report in order to demonstrate both data input preparation and the interpretation of output file results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 .O INTRODUCTION

1.1 General EM/Salinity Modeling and Assessment Techniques

Accurate soil **salinity assessment is needed for** the **design of efficient agricultural management practices and irrigation water allocation strategies. Fortunately, the ability to diagnose and monitor field scale salinity conditions has been significantly improved through the use of electromagnetic induction (EM) survey instruments. Within the last 15 years, the adaptation of EM sensors for soil electrical conductivity measurement has greatly increased both the speed and reliability of salinity reconnaissance survey work.**

The efficient use of EM signal information requires the conversion of apparent soil conductivity (EC,) into soil salinity (EC,). A significant amount of research in recent years has been directed towards developing efficient conversion techniques (Williams and Baker, 1982; McNeill, 1986; McKenzie et. al., 1989; Rhoades and Corwin, 1990; Rhoades et. al., 1990; Rhoades, 1992; Slavich, 1990; Cook **and Walker, 1992; Diaz and Herrero, 1992; Yates et. al., 1993, Lesch et. al., 1992, 1995a,b). These conversion techniques can generally be classified into one of two methodological approaches; (1) deterministic, and (2) stochastic. In the deterministic approach, either theoretically or empirically determined models are** used to convert EC, into EC,. Deterministic models are "static"; i.e., all model parameters are considered known and no soil salinity data needs to be collected during the survey. However, these models typically require knowledge of additional soil properties (e.g., soil moisture, texture, temperature, etc.). In the stochastic approach, statistical modeling techniques such as spatial regression or cokriging are used to directly predict the soil **EC,** from EC, survey data. In this latter approach, the models are "dynamic"; i.e., the model parameters are estimated using of sample salinity data collected during the survey.

There are both advantages and disadvantages to using a stochastic as opposed to deterministic modeling approach. For example, stochastic models are usually much more accurate than deterministic models, since they are explicitly "calibrated" to the specific field being surveyed. They also typically require no knowledge of additional, secondary soil properties (although such information can sometimes be either implicitly or explicitly included into the model, if necessary). However, because stochastic models are dynamic, soil samples must be acquired during each survey expedition. Additionally, these models also tend to be both time and location dependent.

In Lesch et. al., 1995a,b, a comprehensive methodology was introduced for carrying out a field scale salinity survey using a stochastic/dynamic modeling

approach. This methodology centered around the use of spatial regression models for predicting soil salinity from EC, survey data. These models were shown to have a number of important advantages over other stochastic/dynamic modeling approaches, including (1) they facilitated the use of rapid, mobile EM surveying techniques, (2) they could be estimated using a very limited number of soil samples, (3) they could make both point and conditional probability estimates, (4) they could be used to test for changes in the geometric mean field salinity level over time, and (5) they were shown to be theoretically equivalent to cokriging models, provided the regression model residuals are spatially independent.

This manual describes and documents a series of site selection and salinity modeling software programs developed from the above mentioned methodology. It is designed to be used both as a software reference text and tutorial guide. A brief introduction to the spatial regression modeling approach is given in section 1.2. However, this manual does not represent a theoretical documentation of spatial regression modeling techniques. The theoretical details behind this approach can be found in Lesch, et. al., 1995a,b.

1.2 Spatial Regression Modeling Techniques

The spatial regression models discussed throughout this manual can all be written using the following multiple linear regression (MLR) notation:

$$u_{ij} = b_{0j} + b_{1j}w_{1i} + b_{2j}w_{2i} + \ldots + b_{kj}w_{ki} + \xi_{ij} \qquad (1.1)$$

In equation 1.1, u_{ij} represents the Ln (natural log) transformed soil salinity level within the *jth* depth of the *ith* sample site, for $i = 7, \ldots, n$ and $j = 7, \ldots, c$. Additionally, w_{1i} through w_{ki} represent either log transformed and decorrelated EM signal readings or the spatial (x,y) location coordinates associated with the *ith* survey site, and b_{0j} through b_{kj} represent the empirical model parameters (which must be estimated from the observed salinity data). The errors, ξ_{ij}, are assumed to be normally distributed with constant variance and independent from site to site, but possibly correlated between different sampling depths within the same site.

Equation 1.1 is based on a number of implicit assumptions concerning the relationship between soil salinity and apparent soil conductivity, three of which deserve special attention. The first assumption is that the relationship between salinity and conductivity is approximately linear on the Ln scale. It has been found in practice that as the soil salinity exceeds 1.0 dS/m, the relationship between salinity and electrical conductivity begins to become increasingly nonlinear. By 10 dS/m, the linear relationship between salinity and conductivity breaks down completely in most commercial EM instruments, including the Geonics EM-38 meter (McNeill, 1980). Since most surveys are performed on fields already suffering some degree of

salination, the nonlinearity in the salinity/conductivity relationship must be accounted for. Also, at higher salinity levels the micro variability of the salt content within the soil typically increases. Due to these two effects, we recommend applying a Ln transformation to both the salinity and conductivity data before estimating the regression model. First, in a strict mathematical sense, a Ln transformation helps correct for the nonlinearity induced by the quadrature component of the received magnetic field in a highly conductive environment. Second, from a statistical perspective, such a transformation also helps to stabilize the residual variance (a required assumption for the MLR model) and ensures that all predicted salinity levels will remain positive.

The second assumption in equation 1.1 is that the soil salinity within specific depth intervals can be estimated by acquiring multiple conductivity readings over the same site. We have generally found this to be true, provided that at least three separate readings can be acquired (with uniquely different signal response functions). The Geonics EM-38 meter can supply two of these readings (a horizontal and vertical dipole reading). However, the third reading will usually have to come from some other type of direct contact instrument, such as an insertion four-probe or hand held wenner array.

Insertion four-probes and small, hand held wenner arrays are both very useful for measuring the soil conductivity within the first 25 to 50 centimeters of topsoil (Rhoades, 1992). Figure 1.1 displays the relative signal response functions for the EM-38 horizontal and vertical dipole readings, and for an insertion four-probe reading acquired at a 15 centimeter depth. In theory, both EM-38 readings supply a depth weighted conductivity reading for the entire soil profile. However, this is not the case for the insertion four-probe or a suitably scaled wenner array. Both the four-probe and wenner array can be used to isolate and capture near surface conductivity values, and hence increase the depth specific signal resolution accuracy. This additional information can be critical if the soil conductivity level increases rapidly with depth, since the near surface (low conductivity) contribution to both EM-38 readings is often swamped out in strongly regular profiles (Rhoades et. al., 1991).

The third important assumption in equation 1.1 is that the regression model residuals are spatially independent. This assumption must always be verified through a thorough residual analysis, since spatially autocorrelated residuals can corrupt ordinary least squares estimation techniques and cause severe model bias. In practice, the validity of this assumption will depend on when, where, and how you conduct your EM/salinity survey. There are a number of steps you can take to both minimize the chances of observing strong spatial autocorrelation in the residuals and improve the prediction accuracy of the fitted model(s). These steps are discussed in detail in Section 4. However, the software described in this manual can only detect spatial autocorrelation; it cannot adjust the fitted regression models to reduce the prediction bias caused by it. If serious residual spatial autocorrelation is detected,

you will have to use other types of statistical prediction techniques to compensate for it. (Some alternative prediction techniques are reviewed in Lesch et. al., 1995a.)

1.3 The ESAP Software Package

The ESAP software package consists of four specialized programs: *VALIDATE, EMCCRSD, EMSMLR,* and *SALTMAP.* The *VALIDATE* and *EMCCRSD* programs have been designed to be used together to transform and decorrelate your EM survey data, and to select an appropriate subset of survey sites for soil sampling. The *EMSMLR* program has been designed to estimate and validate an appropriate spatial regression model for predicting the Ln soil salinity levels from your transformed and decorrelated EM survey data. This program can also be used to test for a change in the geometric mean field salinity level over time, provided additional sample salinity data is acquired at some point after the initial survey. Finally, the *SALTMAP* program can be used to create and print high resolution maps of the spatial salinity distribution throughout your survey area.

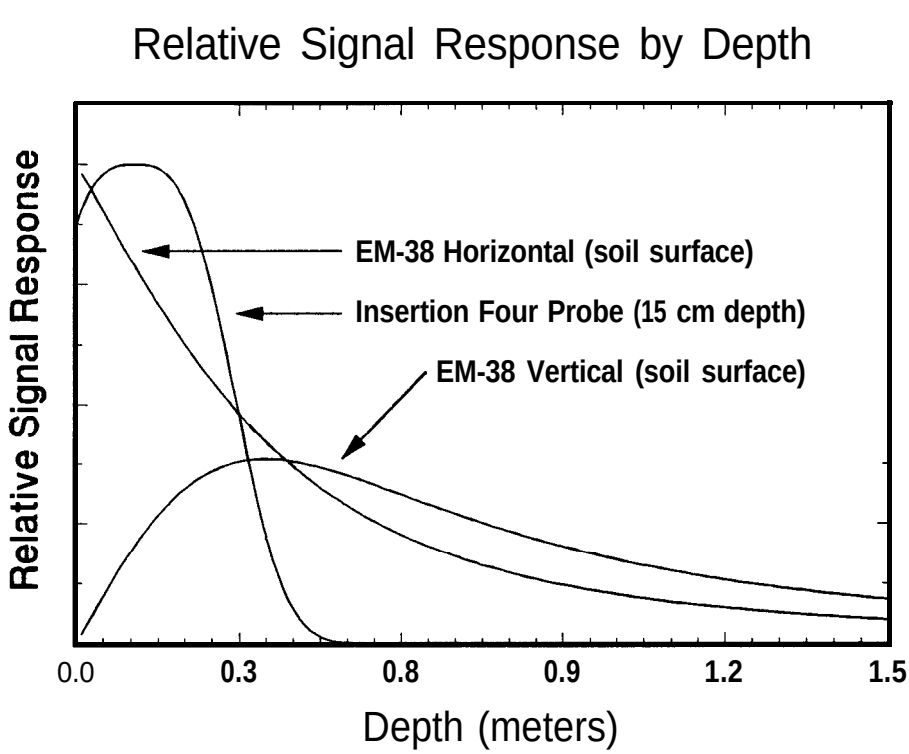
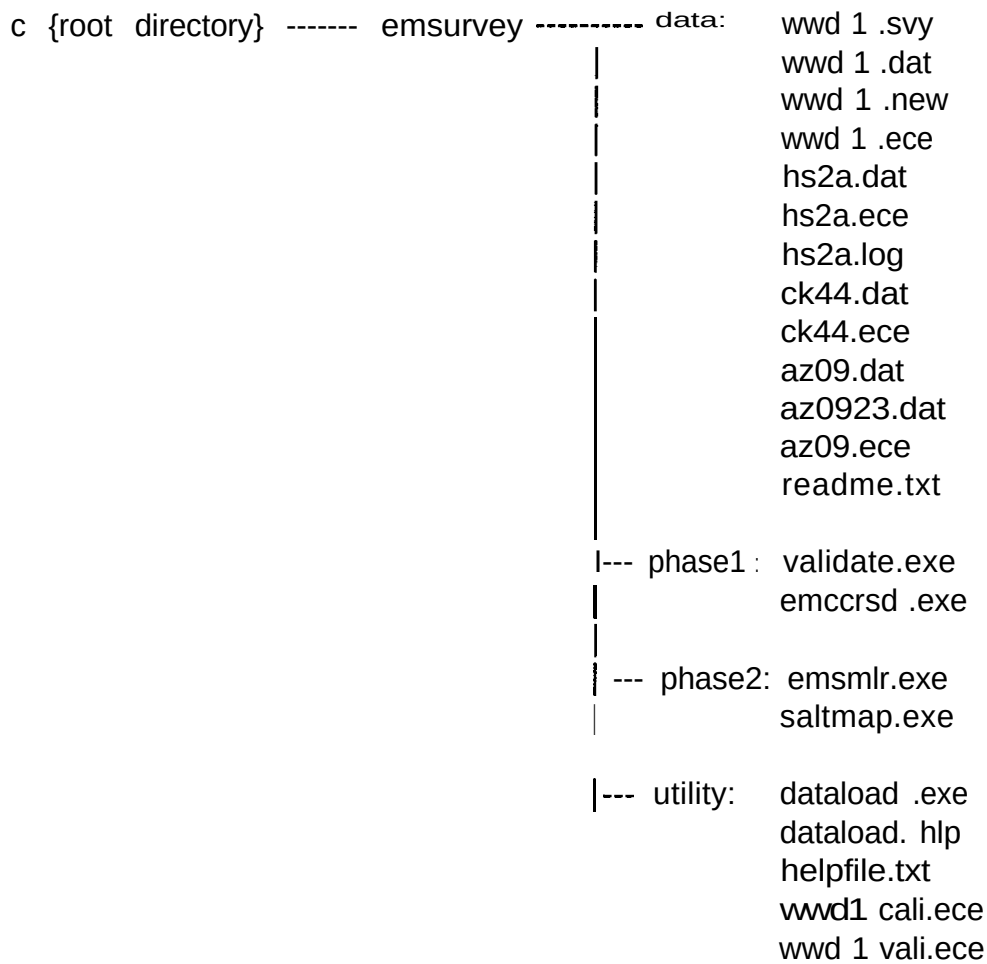


Figure 1.1 Relative signal response functions, by depth, for the Geonics EM-38 meter and an insertion four-probe inserted 15 cm into the soil.

1.4 Software Installation Directions

To install the **ESAP** software package, insert the supplied diskette into your **A** drive (or B drive), type **a:\install (or b:\install),** and follow the screen directions. If you use the default installation settings, the following subdirectories, programs, and data files will be installed:



You will not need to modify your autoexec.bat file unless you wish to run the programs from other than the c:\emsurvey\phase1 or \phase2 subdirectories, or if your computer currently has less than **579K** of free conventional memory. **(To** find out exactly how much memory your computer currently has, go to the root directory and type **mem.** If the largest executable program size is less than **579K,** consult with your computer hardware specialist about how to free up more memory.)

1.5 Manual Organization & Syntax Style

A. Organization

This manual is divided into 7 sections. Sections 2 and 3 document and describe the ***VALIDATE*** and EMCCRSD programs, respectively. An analysis of a tutorial **EM** survey data set is included at the end of each section to help you become familiar with the operation of each program. Section **4 discusses various ways of improving both the EM** surveying and soil sampling process, specifically with respect to minimizing sampling variability and maximizing the regression model prediction accuracy. This section should be read immediately after Section 3, and before conducting any field survey work. Sections 5 and 6 document and describe the ***EMSMLR*** and ***SALTMAP*** programs, respectively. An analysis of a tutorial salinity data set is again included at the end of each section. After finishing Section 6, you should be familiar enough with the ***EMSMLR*** and ***SALTMAP*** programs to advance onto Section 7. Section 7 leads you through a detailed analysis of three additional tutorial salinity data sets; in an effort to further refine your modeling and analysis skills.

This manual can be used as a self-tutorial guide. To use it as such, we suggest that within each section you first read the program documentation, then run the tutorial example(s), and then read the documentation once again. If you have not had any prior experience or training in multiple linear regression modeling, you may also wish to review these techniques before reading this manual.

The collection of tutorial data sets used in sections 2, 3, 5, and 6 are from a 1993 survey conducted in a 160 acre cotton field located in central California. This field is located within the boundaries of the Westland Water District, hence all tutorial data files associated with it have the same four character prefix: WWD1. Full descriptions of the data sets from each of the four surveys discussed in this manual are given in Section 1.6.

B. Syntax Style

The following syntax style will be used throughout this manual. All software programs will be referred to in a capitalized italic font. For example, the emccrsd.exe fortran program described within this manual will always be referred to as ***EMCCRSD.*** All program input/output files discussed in the text will also be referred to in a capitalized italic font, and will be displayed using their appropriate three character DOS extensions; e.g., ***EMCCRSD.OUT*** represents the output ASCII text file created by the ***EMCCRSD*** program. Whenever any program screen output is referred to, it will be displayed in the text using a small italic font. Additionally, any keystroke commands which must be performed by the user will be shown in small bold font, and the Enter/Return key will always be abbreviated using the following symbol: **[↵].** For example, if you see the following line of text

Please enter the field title: **My first Salinity Survey [↵]**

then you should recognize that program screen output is "Please enter the field title:" and that your keyboard input is "My first Salinity Survey", followed by an Enter/Return keystroke.

At various points throughout this manual you will encounter text preceded by one of the following two comments: NOTE or WARNING. Text associated with a NOTE comment will typically describe either special program features, helpful programming tips, or additional software information which you may find useful. Text associated with a WARNING comment will always contain critical program information which you must be aware of when using the software.

1.6 Descriptions of Tutorial Data Sets

EM signal and sample soil salinity data from four different surveys have been included with the ESAP software package. The data from three of these surveys will be used in section 7 to help you become more comfortable with the EMSMLR program. As previously mentioned, the data associated with the Westland Water District survey is used in sections 2, 3, 5, and 6 as the primary tutorial example. Some details concerning each survey are given on below.

A. WWD1 Survey

This survey was conducted in the spring of 1993 on a 160 acre cotton field located within the Westland Water District in central California. EC, data were acquired at 180 survey sites (approximate 55 meter grid) using the mobile EM, four-electrode sensing system (Rhoades, 1993). Four EC, readings were taken at each survey site; two EM-38 readings (horizontal and vertical) and two wenner array readings (1 .0 m and 2.0 m spans). All survey readings and soil samples were acquired in the furrows. Soil samples were acquired from 16 calibration and 8 validation sites, at sampling depths of 0.0-0.3, 0.3-0.6, 0.6-0.9, and 0.9-I .2 meters. Replicate soil cores were acquired at 5 of the 16 calibration sites.

B. CK44 Survey

This survey was conducted in spring of 1993 on a 36 acre fallow (disked) field located within the Choachella Valley Water District in southern California. EC, data were acquired at 139 survey sites (approximate 28 meter grid) using the mobile EM, four-electrode sensing system. Four EC, readings were taken at each survey site; two EM-38 readings (horizontal and vertical) and two wenner array readings (1 .O m and 2.0 m spans). Soil samples were acquired from 16 calibration sites at sampling depths of 0.0-0.3, 0.3-0.6, 0.6-0.9, and 0.9-I .2 meters. Replicate soil cores were

acquired at 4 of the 16 calibration sites.

C. AZ09 Survey

This survey was conducted in spring of 1993 on a 32 acre corn field located near the Gila Indian Reservation in the state of Arizona. EC, data were acquired at 114 survey sites (approximate 28 meter grid) using the mobile EM, four-electrode sensing system. Four EC,, readings were taken at each survey site; two EM-38 readings (horizontal and vertical) and two wenner array readings (1 .O m and 2.0 m spans). All survey readings and soil samples were acquired in the furrows. Soil samples were acquired from 17 calibration and 8 validation sites, at sampling depths of 0.0-0.3, 0.3-0.6, 0.6-0.9, and 0.9-l .2 meters. Replicate soil cores were acquired at 5 of the 17 calibration sites. Because of high soil textural variability, the validation data set was combined with the calibration data set, yielding a total calibration sample size of 25 sites.

D. HS2A Survey

This survey was conducted in spring of 1989 on a 40 acre cotton field located near Hanford, California. EC, data were acquired at 206 survey sites (25 meter grid) using hand held electromagnetic induction meters. Six EC, readings were taken at each survey site; two EM-38 readings (horizontal and vertical) and four four-probe readings at a 15 cm depth. At each site, the four four-probe readings were averaged into a single composite reading. Individual soil samples were acquired at each of the 206 survey sites at a sampling depth of 0.0-0.3 meters. (No replicate soil samples were acquired at any of the sites.)

The laboratory salinity analysis on the soil samples from all four surveys was performed using the methods of Rhoades et. al., 1989. The WWD1 and HS2A survey data sets have been previously discussed in Lesch et. al., 1995a,b.

2.0 VALIDATE PROGRAM DOCUMENTATION

2.1 Program Description

VALIDATE is **designed** to validate and transform your EM signal data, and produce an output file suitable for use with the EMCCRSD program. ***VALIDATE*** reads as input an ASCII text file (created by the user) containing EM survey information. This survey information can be collected at anywhere from 43 to 399 distinct survey sites within a field. ***VALIDATE*** first computes summary statistics on the signal data, and then allows the user to interactively view four different types of signal data plots. The plot types which the user may choose from are (a) bivariate scatter plots, (b) Q-Q Normal probability plots, (c) scaled distance semivariogram plots, and (d) an x/y location (survey coordinate) plot. Next, the program performs a principal components analysis on the EM signal data, transforming this data into centered and scaled principal component (PC) scores. It then returns to the interactive view mode, allowing the user to use the plots described above to examine the PC score data. At this point the user should use the bivariate scatter plots (of the PC score data) to search for outlier signal data.

If outlier PC score data are detected, ***VALIDATE*** allows the user the option of deleting the survey sites associated with these data. If the user chooses to delete one or more survey sites, the program recomputes a new set of principal component scores using the remaining EM signal data, and then once again returns to the interactive view mode. This iterative process continues until either (1) no further outlier PC score data are detected in the EM signal data set, or (2) the user requests that no further sites be deleted.

During this process, ***VALIDATE*** creates two output ASCII text files: ***RA WSTAT. TXT*** and ***PCSTA T. TXT.*** The ***RA WSTAT. TXT*** file contains summary statistics and histogram plots calculated using all the original input EM signal data. The ***PCSTAT. TXT*** file contains statistics concerning the iterative principal component transformations, as well as log notes which document the removal of any survey sites.

Once the iterative validation process is completed, ***VALIDATE*** performs two final functions. First, it allows the user to mask one or more of the remaining survey sites. Masking effectively prohibits a survey site from being selected as a calibration (sample) site, without deleting it from the data set. Second, it creates a 3rd output file, ***EMCCRSD.IN.*** This file contains the final survey location and PC score data information for use in the ***EMCCRSD*** program.

A flowchart of the program computations performed by *VALIDATE* is shown in Figure 2.1.

2.2 Input/Output File Description

To function properly, *VALIDATE* must read as input an ASCII text file containing EM survey information. This file must be created by the user before initiating the program. Additionally, the data in this input file must have the following column structure:

[1] site ID [2] x [3] y [4 - 81 EM-38, wenner and/or four-probe data

The site ID column (column 1) must contain whole numbers and no two site ID numbers can be the same. Although it is not mandatory, it is a good idea to make the site ID numbers sequential; i.e., 1, 2, ... N. The x and y coordinates (columns 2 and 3) can assume any real, floating point values, however, we recommend using a scaled coordinate system. For example, if the coordinates are measured in meters, then dividing by 100 or 1000 will help prevent screen and output file format problems. All EM signal data should be placed into columns 4 through 8. There must be no less than two and no more than five columns of EM signal data. Typically, columns 4 and 5 will contain the EM-38 vertical and horizontal readings; however, any valid EM signal data may be placed into these columns.

There is no specific data format structure (the data is read in free-format mode); however, the above ordering of the columns must be maintained for *VALIDATE* to function properly. Additionally, you can have no less than 43, and no more than 399 distinct survey sites within your EM survey input file and no row can contain missing EM data. Any survey site missing one or more EM signal levels must be removed (deleted) from the input data file before initiating the *VALIDATE* program.

The files *RA WSTAT. TXT, PCSTAT. TXT,* and *EMCCRSD.IN* are automatically created by the *VALIDATE* program during execution. Upon completion of the *VALIDATE* program, RA *WSTAT. TXT* and *PCSTAT. TXT* can each be printed using the standard DOS print command. The *EMCCRSD.IN* file should not be edited in any manner, since any changes will effect the execution of the *EMCCRSD* program.

The first text file, RA *WSTAT. TXT,* contains summary statistics and histograms computed from the input EM signal data. Note that the EM signal statistics and histogram plots in *RA WSTAT. TXT* will always be based on the natural log transformed signal data values. The second text file, *PCSTAT. TXT,* lists the iterative principal component transformation statistics (eigenvalues, eigenvectors, etc.) and also documents which survey sites get deleted or masked during the validation analysis. Example output from these two text files will be shown in Section 2.4.

VALIDATE.EXE FLOWCHART

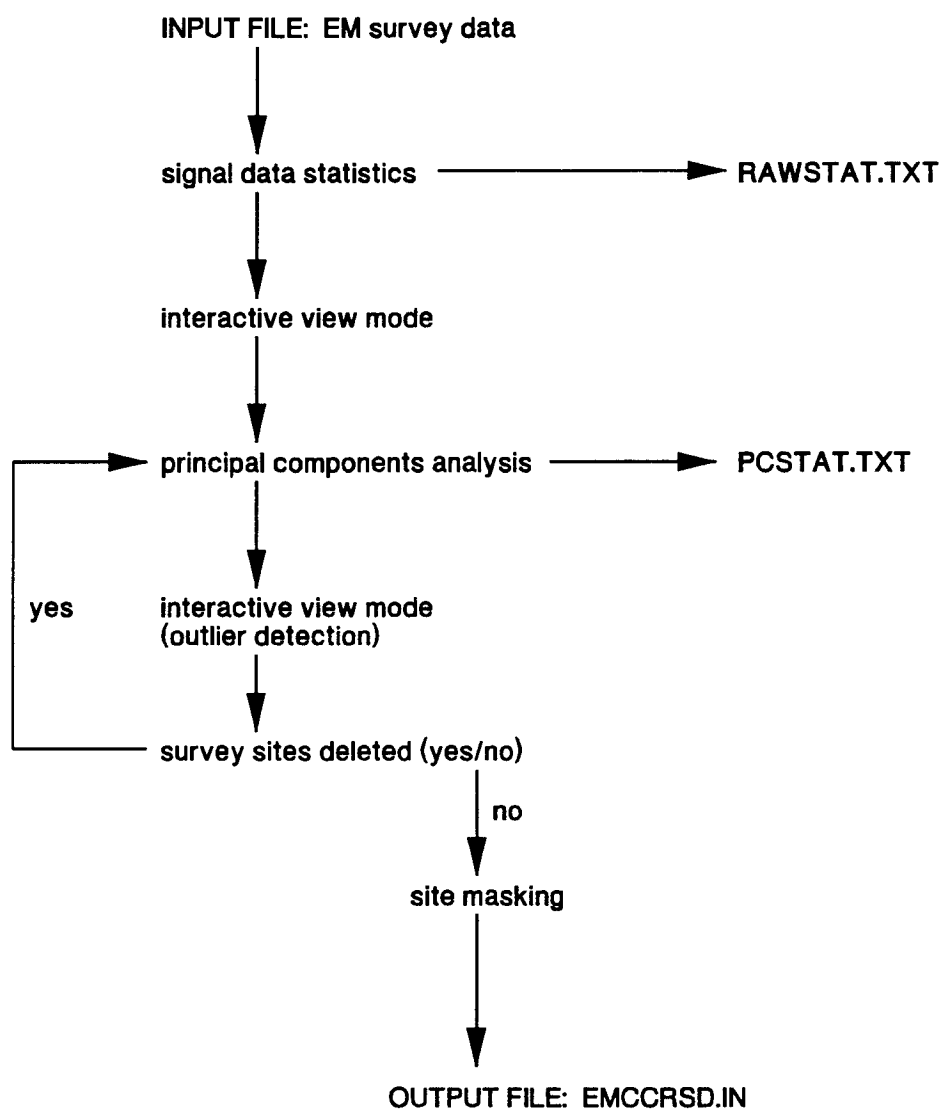


Figure 2.1 Flowchart displaying computations in *VALIDATE* program.

2.3 Program Operation

To start the *VALIDATE* program, move into the c:\emsurvey\phase1 subdirectory and type **validate** [↵] at the DOS prompt. The program will print some initial header information on the screen, along with the following prompt:

Do you need help running this program (y/n):

If you've forgotten how to use the program, type y [↵] to access the help screen, otherwise type n [↵]. The program will then ask you for the following information, (in the sequential order listed below):

Please enter pa th/fiename:

Please enter field title:

Please enter the # of survey sites (N):

Please enter the # of EM readings per site (m):

You **must enter the necessary information after each prompt, and each entry must satisfy the following restrictions: 1) the field title must be ≤ 40 characters long, 2) the number of survey sites must be 43 ≤ N ≤ 399, and 3) the number of EM signal readings must be 2 ≤ m ≤ 5. Entering an incorrect path/filename statement, or entering anything other than integers at the (N) and (m) prompts will cause the program to crash.**

Next, the program will iteratively ask you specific information about each column of EM signal data:

Please enter 1st EM signal title:
Ln transform this signal data (y/n):

Please enter mth EM signal title:
Ln transform this signal data (y/n):

At each iteration you will need to supply a EM signal title (8 characters or less) and a yes/no answer at the Ln transformation prompt.

WARNING: **The *VALIDATE, EMCCRSD, EMSMLR,* and *SALTMAP* programs are designed to work only with natural log transformed EM signal and soil salinity data. <u>If the EM signal data in vour input file has not alreadv been Ln transformed, vou must answer ves at each Ln transformation prompt.</u>**

After the final Ln transformation prompt, the following message will be printed to the screen:

Execution suspended: press Return (Enter) key to view summary statistics

Typing the **[↵]** key will allow you to view all the summary statistics associated with the Ln transformed EM signal data. The summary statistics written to the screen include the mean, variance, skewness, minimum, and maximum value associated with each column of signal data, along with the data correlation matrix. This information will be followed by another execution suspended message:

Execution suspended: press Return (Enter) key to go in to View Mode

The following statements will then appear on the screen after typing the **[↵]** key:

You may now interactively view your raw survey data.

Four plot types are available:
a) Bivaria te sca tter plots
b) Q-Q Normal probability plots
c) Scaled distance semivariogram plots
d) X/Y Coordinate Grid

Initializing Q-Q and SemiVariogram Arrays
One moment please.. .

After a few seconds, another prompt will appear:

Plot Type: a) Bivariate b) Q-Q c) SemiVariance d) x/y grid:

To initiate a specific plot, you should type the corresponding letter **("a", "b", "c",** or **"d"),** followed by the **[↵]** key. You will then be asked to specify which data columns should be plotted (by selecting the appropriate column numbers). The requested plot will then appear on the screen, followed by the prompt:

Would you like to view another plot? (y/n):

At this point, you may remain in the interactive view mode (i.e., continue viewing the various signal data plots) by answering yes, or leave the view mode by answering no.

These initial signal plots are designed to help you visually appraise your survey data. For example, the bivariate data plots will reveal the degree of correlation inherent in the Ln transformed EM signal data, while the Q-Q plots can be used to appraise the assumption of data normality. (Approximate normality should be induced by the Ln transformation, unless there is a strong trend in the signal data across the field.) Additionally, the semivariogram plots will reveal the degree of spatial autocorrelation inherent in the data, and the survey grid can be used to make sure

that the correct x/y location coordinates have been read into the program.

NOTE: In order to save on time and memory storage, the semivariogram estimation subroutine uses an approximation algorithm to determine the intrinsic lag-spacing intervals for the semivariance estimates. This algorithm assumes that the survey data has been collected on a rectangular grid, and may therefore produce unreliable results when the underlying grid pattern is highly irregular.

NOTE: The screen resolution for any given plot produced in the ***VALIDATE*** program is limited to 22 by 79 characters. Hence, there will typically be a fair amount of symbol overlap in the bivariate scatter and Q-Q normality plots. Some overlap and/or distortion may also occur in the x/y grid plot when the underlying grid pattern is either highly irregular or sufficiently dense.

The following message will appear on the screen once you have finished the 1 st interactive view session:

Interactive View session completed..

Execution suspended: press Return (Enter) key to compute principal components

The principal component scores of the Ln transformed signal data can now be computed by typing the [↵] key. If your input file contains two columns of EM signal data (i.e., EM-38 horizontal and vertical readings only), then two principal component scores will be calculated and retained for further analysis. If your input file contains three, four, or five columns of EM signal data, then the first three principal component scores will be retained for further analysis. During this computation process, another message appears:

Computing principal component scores
Please wait..

After a few seconds, this message will be followed by another execution suspended comment:

Execution suspended: press Return (Enter) key to return to View Mode

Typing the [↵] key will return you back into the interactive view mode, where you can now interactively plot the principal component data.

At this point of the program, you should sequentially create bivariate scatter plots of each pair of principal component scores. As these scatter plots are being

created, ***VALIDATE*** will check this data for outliers and warn you if any unusual PC scores are found. If an outlier is discovered, a warning message will appear on the screen (before the scatter plot appears), identifying the survey site ID number associated with these data. Note that multiple warning messages will appear if multiple outliers are discovered. You should manually record each site ID number, so that these sites can be deleted or masked out later on in the program. (You cannot delete or mask out any signal data while in the interactive view mode.)

WARNING: The user must iteratively create bivariate scatter plots of each pair of principal component scores in order to search for and identify unusual signal data. ***VALIDA*** *JE* will only search for outliers immediately before producing a bivariate scatter plot, and only within the 2 requested columns of principal component scores. It is therefore imperative that the user create bivariate scatter plots of every pair of PC scores.

NOTE: ***VALIDA*** *JE* identifies outlier PC scores by computing a χ^2 statistic which measures how far each bivariate observation deviates from 0. This statistic, referred to as a "joint PC deviation", is printed out to the screen with each warning comment. If the joint PC deviation exceeds 3.5, the corresponding bivariate observation will be flagged as an outlier. Bivariate observations with deviations exceeding 4.5 should generally be deleted from the signal data set. Bivariate observations with deviations between 3.5 and 4.5 are considered "marginal outliers", and can either be masked out or deleted.

NOTE: The other three plot types can also be requested at this point in the program. The Q-Q plots can be used to assess the approximate normality of the principal component data, and the degree of spatial autocorrelation can be inferred from the semivariogram plots. The x/y grid plot can be used to show the locations of deleted survey sites, if the principal component analysis has been re-computed on a reduce set of signal data.

Upon completion of the 2nd interactive view session, the following message will appear if any outlier principal component data has been discovered:

CAUTION! ***One or more outlier principal component scores have been detected within the PC data set.***

Do you wish to delete any sites? (y/n):

If you answer yes, the next prompt which appears is:

Please specify the survey site # to be deleted:

You should now enter the site ID number, followed by the [↵] key. The program will ask you to confirm that the site number has been entered correctly, and then delete the site. After this step, you will be allowed to delete additional sites, one at a time, until all outlier data have been removed.

WARNING: A number of error traps and confirmation loops are built into the site deletion process, to aid the user in avoiding mistakes. However, once a site is deleted, it cannot be recovered within the program. To recover an incorrectly deleted site, the user must reinitiate the ***VALIDATE*** program (i.e., start the program over again).

After exiting the site deletion process, the following message will appear:

No additional sites will be deleted.

Computing principal component scores.
Please wait.. .

followed by (a few seconds later)

Execution suspended: press Return (Enter) key to return to View Mode

Typing the [↵] will send you back into the interactive view mode, where you can once again search the new principal component scores (computed on the reduced EM signal data set) for additional outliers.

NOTE: This iterate process will repeat itself until either (1) no further outliers are discovered during the interactive view session, or ***(2)*** the user elects not to delete any further signal data from the survey data set.

After all the outlier data has been identified and removed from the survey data set, the program will print the following message to the screen:

Do you wish to mask any sites? (y/n):

If you answer yes, the next prompt which appears is:

Please specify the survey site # to be masked:

You should now enter the site ID number, followed by the [↵] key. As in the site deletion routine, the program will ask you to confirm that the site number has been entered correctly, and then mask the site. After this step, you will be allowed to mask additional sites, one at a time.

WARNING: The same warning which applies to the site deletion routine applies here. Once a site is masked, it cannot be un-masked within the program. To un-mask an incorrectly masked site, the user must reinitiate the ***VALIDATE*** program (i.e., start the program over again).

Upon exiting the site masking routine, the program terminates after printing the following comments to the screen:

Validation / transformation step completed.

Run EMCCRSD to generate the soil sampling design.

At this point, before initiating the ***EMCCRSD*** program, you should print out the RA ***WSTA T. TXT*** and ***PCSTA T. TXT*** text files. The ***PCSTA T. TXT*** file should be checked to confirm that the correct number of sites (if any) have been deleted or masked, and both printouts should be saved for further reference.

2.4 Tutorial Example

You should now try running ***VALIDATE*** using the supplied tutorial survey data set, ***WWD 1.SVY.*** The input file attributes are N = 180 and m =4. The sample data path/filename is “c:\emsurvey\data\wwd 1 .svy” (provided you used this default subdirectory when installing the ESAP software package). Enter “Westland Field 1” for the field title, and “EMv”, “EMh”, “Wn01”, and “Wn02” for the 4 EM signal titles. Be sure to also Ln transform each column of signal data.

After viewing the summary statistics, proceed into the interactive view mode and try out the four types of screen plots: (a) bivariate scatter plots, (b) Q-Q Normal probability plots, (c) scaled distance semivariogram plots, and (d) the x/y coordinate grid plot. Pay particular attention to the following attributes revealed by these plots. In the bivariate scatter plots: note the high degree of correlation between the various Ln transformed signal readings, and the 3 unusual (outlier) points in the Wn01/Wn02 scatter plot. In the Q-Q normal probability plots: note that all the Q-Q plots appear “heavy-tailed”, implying that the log transformed signal data is not Normally distributed. In the semivariogram plots: note that all semivariograms appear to be linear, suggesting that there is probably a strong directional trend in the signal data. Finally, in the x/y grid plot: note that no survey data was acquired at the last site in row 1 (upper right-hand corner of the screen), or the first site in row 3.

Next, compute the principal components transformation and then return to the interactive view mode. In the view mode, construct bivariate scatter plots of the following three pairs of principal component scores: PC1/PC2, PC1/PC3, and PC2/PC3. Note that two warning statements appear on the screen before the

PC1/PC3 and PC2/PC3 scatter plots are made. In each case, the warning statements identify survey sites 69 and 173 as outliers.

Before leaving the view mode, create Q-Q and semivariogram plots of each principal component. Note that in the semivariogram plots the spatial autocovariance structure deteriorates in the higher principal components (especially the 3rd PC score). Note also that the Q-Q plot of the 3rd PC scores clearly reveal two serious outliers (sites 69 and 173).

Now leave the view mode and answer yes at the site deletion prompt. Delete sites 69 and 173 from the data set, recompute a new set of principal component scores, re-enter the view mode, and create a x/y grid plot. Note that sites 69 and 173 are now missing from the survey grid. Next, create a new set of bivariate scatter plots. Note that one more site now gets identified as a marginal outlier (site 99) in the PC1/PC3 and PC2/PC3 plots.

Leave the view mode again, but this time answer no at the site deletion prompt. This will exit you out of the iterative validation/deletion process. Next, answer yes at the site masking prompt, mask out site 99, and then exit out of the program by answering no at the "mask another site?" prompt.

Print out the RA *WSTAT. TXT* and *PCSTAT. TXT* files and make sure they match the text file printouts shown in Tables 2.1 and 2.2. Note that histogram plots of the Ln transformed EM signal data are contained in the *RA WSTAT. TXT* file, and that these plots suggest that the Ln transformed signal data is possibly bi-modal. Note also that the ID numbers of all the deleted and masked sites are listed in the *PCSTAT. TXT* file (along with the principal component statistics).

Finally, note that a third file now exists in the c:\emsurvey\phase1 subdirectory, *EMCCRSD.IN.* This file contains all the necessary input information needed by the *EMCCRSD* program.

Table 2.1 Example of ASCII text output contained in RA *WSTAT. TXT.*

```
EM Signal Data // Summary Survey Statistics

Title:  Westland  Water  District

Total # of survey sites =        180
Total # of EM signals per site =   4

Signal.Type   Ln.Trnsfrm

    EMv           yes
    EMh           yes
    WOl           yes
    wo2           yes

   STATISTIC      EMv          EMh          W01          wo2

   mean           0.89724      0.38873      0.69318      1.54614
   variance       0.11748      0.13939      0.19607      0.19993
   skewneww      -0.11456     -0.09595      0.07002      0.09954
   minimum        0.21511     -0.32850     -0.22314      0.66320
   maximum        1.44456      0.98582      1.59127      2.45453

   CORRMATRX      EMv          EMh          W01          WO2

   EMv            1.00000      0.99242      0.85730      0.91140
   EMh            0.99242      1.00000      0.89970      0.93730
   WOl            0.85730      0.89970      1.00000      0.96268
   wo2            0.91140      0.93730      0.96268      1.00000

Ln-transformed EM Signal Histogram Distributions

   EMv            EMh            WOl            wo2
++++++++++++++ ++++++++++++++ ++++++++++++++ ++++++++++++++
HHHHHHHHHHHHHH HHHHHHHHHHHHHH HHHHHHHHHHHHHH HHHHHHHHHHHHHH
HHHHHHHHHHHHHH  HHHHHHHHHHHHH  HHHHHHHHHHHHH  HHHHHHHHHHHHH
 HHHHHHHHHHHHH  HHHHHHHHHHHHH  HHHHHHHHHHHH   HHHHHHHHHHHH
 HHHHHH HHHHH   HHHHHHHHHHHHH  HHHHHHHHHHHH   HHHHHHHHHHHH
 HHHHHH HHHHH   HHHHHHH  HHHH  HHHHHHHHHHHH   HHHHH HHHHH
 HHHHHH HHHHH   HHHHHHH  HHHH  HHHHHHHHHHHH     HHHH HHHHH
  HHHHH HHHH    HH H HH HHHH    HHHHH  HHHH     HHHH HHHHH
  H   H H HH    H   HH   HH     HHH   H HH     HHHH HHH H
  H   H   HH    H  H     HH     HHH  H  H      HH H HHH H
  HH       H    H         HH     HH     HH      HH  H  HH
  H        H    H         HH     HH    H        HH
  H        H    H          H       H   H        H
           H               H       H   H        H
           H               H                    H
           H
           H
           H
```

Table 2.2 Example of ASCII text output contained in PCSTA T. TXT.

```
EM Principal Components Data // Transformation Statistics & Log Sheet

Title:  Westland  Water  District

Total # of survey sites =        180
Total # of EM signals per site =   4

Column 1 = Eigenvalues  Column 2 = Percent of Trace
Column 3 = Cumulative Percent of Trace

             1           2           3
1      3.78090     0.94523     0.94523
2      0.18549     0.04637     0.99160
3      0.03017     0.00754     0.99914
4      0.00344     0.00086     1.00000

Principal  Axis  Matrix:  Columns  = Eigenvectors, Rows = EM Variables

             1           2           3           4
1      0.49755    -0.58057    -0.04294     0.64306
2      0.50654    -0.37998    -0.19807    -0.74820
3      0.49175     0.63948    -0.56923     0.15885
4      0.50403     0.33109     0.79681    -0.03786

Note: deleted site #:       69
Note: deleted site #:      173

Total # of survey sites =        178
Total # of EM signals per site =   4

Column 1 = Eigenvalues  Column 2 = Percent of Trace
Column  3 = Cumulative Percent of Trace

             1           2           3
1      3.79676     0.94919     0.94919
2      0.18494     0.04624     0.99543
3      0.01503     0.00376     0.99918
4      0.00327     0.00082     1.00000

Principal  Axis  Matrix:  Columns = Eigenvectors, Rows = EM Variables

             1           2           3           4
1      0.49664    -0.57937    -0.08498     0.64067
2      0.50591    -0.37377    -0.18591    -0.75484
3      0.49103     0.65724    -0.55433     0.14018
4      0.50625     0.30441     0.80681    -0.01013

Note:  masked site #:       99
```

3.0 EMCCRSD PROGRAM DOCUMENTATION

3.1 Program Description

EMCCRSD is **designed to identify between 15 to 20 sites suitable** for soil sampling, based on your principal component survey data generated by the ***VALIDATE*** **program.** ***EMCCRSD*** can also select additional monitoring sites for sampling in the future. The observed soil salinity levels at these monitoring sites can then be compared to the model predicted salinity values, and used to test for changes in the field median salinity level over time.

EMCCRSD reads as input a text file called ***EMCCRSD.IN.*** ***EMCCRSD*** **uses** this input data to select the locations of the first 14 sample sites automatically, and prompts the user to select between 1 to 6 additional sites (thereby producing a final calibration size of 15 to 20 sites). Next, the program selects 8 monitoring/validation sites, and then asks the user if the locations of these sites should be included on the sample site location map and crew log-sheet. At this point in the program, the user can either (1) discard these 8 sites entirely, (2) retain these 8 sites for future sampling (i.e., monitoring sites), or (3) retain these 8 sites for concurrent sampling (i.e., validation sites). The user also has the option of creating a multi-stage sampling plan (see section 2.5). Finally, ***EMCCRSD*** asks the user if they wish to create a spatial map of the 1st principal component score. If produced, this map can generally be interpreted as a qualitative first approximation to the bulk-average spatial soil salinity distribution.

During this process, ***EMCCRSD*** creates six output (ASCII) text files: ***SITES.TXT, SSS.MAP, CRWSHEET. TXT, ALGNOTES.OUT, EMCCRSD.OIJT*** and (if desired) ***PC 7. MAP.*** These files contain all the necessary information for implementing an optimal soil sampling design.

A flowchart of the ***EMCCRSD*** program computations is shown in Figure 3.1.

3.2 Input/Output File Description

EMCCRSD has been designed to read the ***EMCCRSD.IN*** ASCII text file created by the ***VALIDATE*** program; no other input files need to be accessed during the programs' execution. Upon initiation, this program will print the following input file information to the screen: (1) the field title associated with the input data, (2) the total number of survey sites, (3) the number of principal component scores, and (4) the number of masked survey sites. When necessary, this information can be used to verify that the correct survey data has been read into the program.

EMCCRSD.EXE FLOWCHART

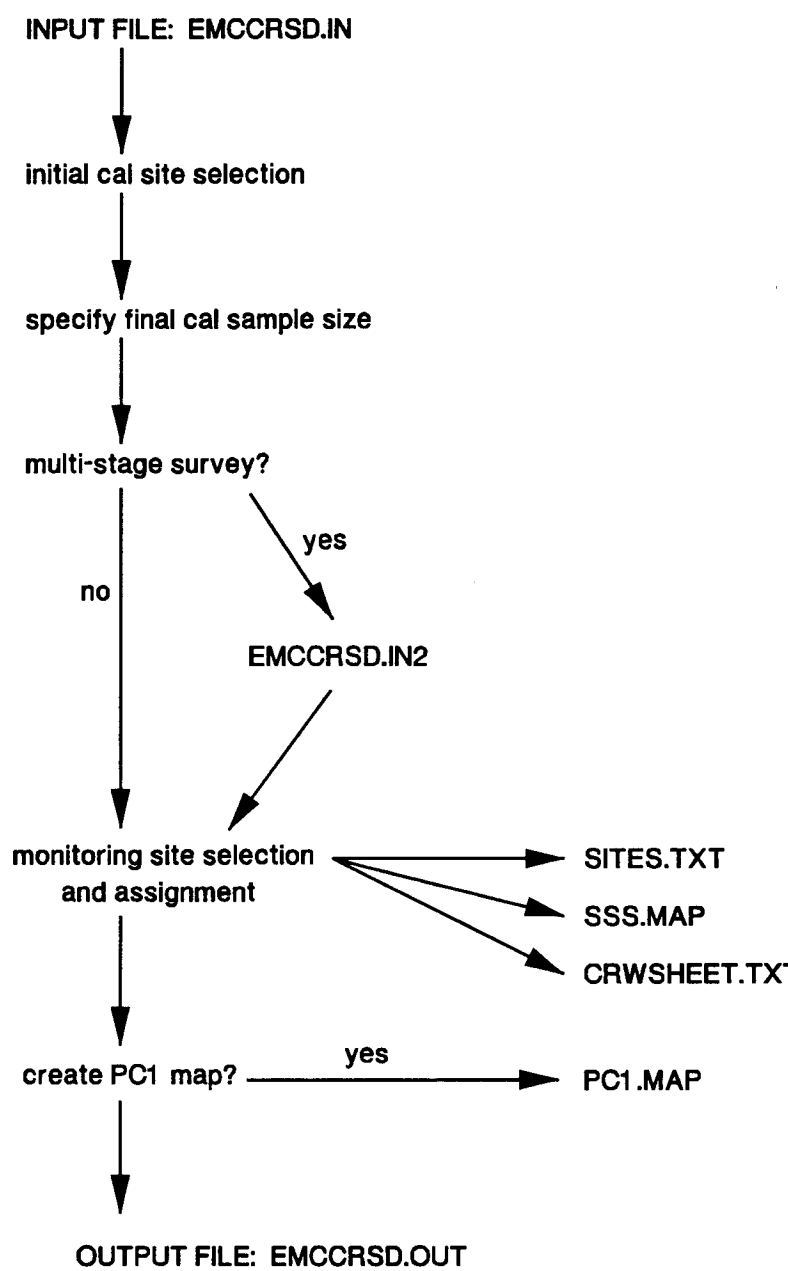


Figure 3.1 Flowchart displaying computations in EMCCRSD program.

WARNING: Do not edit and/or otherwise modify the ***EMCCRSD.IN*** input file in any manner --- modification of this file can corrupt the ***EMCCRSD*** program.

As mentioned in Section 2.1, ***EMCCRSD*** produces six output ASCII text files; all but the ***ALGNOTES.OUT*** and ***EMCCRSD. OUT*** text files should be printed out after exiting the program. (You can obtain a printout of the four remaining files using the standard DOS print command.) The ***SITES. TXT*** file contains the theoretical target design levels and the observed principal component scores for each of the calibration and monitoring/validation sites, along with the sample site ID numbers and x/y location coordinates. The ***SSS.MAP*** file contains a map of the survey grid and identifies the locations of all the calibration and monitoring/validation sites, as well as any masked sites. The ***CRWSHEET. TXT*** file lists the calibration sites in sequential order (by site ID), and can be used as a log-sheet for field notes during soil sampling. (The 8 monitoring/validation sites will also be listed sequentially in this file, if requested by the user.) The ***PC1.MAP*** file contains a spatial map of the first principal component score. This file should only be created if a systematic survey grid was employed to collect the EM survey data. The ***ALGNOTES.OUT*** file generates a summary of the internal computations pertaining to the various iterative site selection procedures -- hence, it does not typically need to be printed.

The ***EMCCRSD.OUT*** text file must always be renamed and/or copied into another subdirectory once the ***EMCCRSD*** program has terminated. This file contains the final principal component scores and coordinates of all the survey sites; you will need this information in order to create a valid input file for the ***EMSMLR*** program (see section 5). We recommend moving a renamed version of the ***EMCCRSD. OUT*** text file into the c:\emsurvey\data subdirectory.

WARNING: Each time the ***EMCCRSD*** program is executed, the ***EMCCRSD.OUT*** text file is overwritten. The name of this file must be changed, and/or this file must be moved into another subdirectory in order to save the transformed and decorrelated signal information.

3.3 Program Operation

To start the ***EMCCRSD*** program, make sure you are still in the c:\emsurvey\phase1 subdirectory and type emccrsd [↵] at the DOS prompt. The program will print some initial header information to the screen, along with the following prompt:

Do you need help running this program (y/n):

If you've forgotten how to use the program, type y [↵] to access the help screen,

otherwise type n [↵]. The program will then print out the EMCCRSD.IN input file information described in Section 2.2, followed by:

Execution suspended: press Return (Enter) key to continue

Type the [↵] key to initiate the first stage of the iterative site selection process; these iterations can take anywhere from 3 seconds (on a 586/60 PC) to about 60 seconds (on a 386/1 6 PC). During this stage, the program will automatically select the first 14 sample site locations, and then print the following statement:

Please enter the # of additional spatial support sites to be included in the sampling design (1 ≤ SSS ≤ 6):

You can now specify the final sample size to be between 15 to 20 by entering the appropriate number of additional support sites (1 through 6). For example, if you wish to select a total of 16 sample sites, type 2 [↵] at this prompt.

NOTE: **You must choose at least 1, and no more than 6 support sites.**

After you have specified the number of additional support sites, the program will initiate the second stage of the site selection process. EMCCRSD will first determine the locations of the support sites, and then select 8 additional monitoring/validation sites. The following question will then appear on the screen:

Will this be a multi-stage sampling design (y/n):

Answer no at this prompt (multi-stage sampling techniques will be described in section 2.5). Next, the following messages will appear:

Note: you have the option of not listing the monitoring sites on the sample site map or crew log-sheet.

Do you wish to include the monitoring sites? (y/n):

If you do not wish to sample at these 8 sites, type n [↵]. If you answer no, then these sites will not be listed on either the sample site map *(SSS.MAP)* or crew log-sheet *(CRWSHEET. TXT).* If you answer yes, then the locations of these 8 sites will be shown on the sample site map. Additionally, the following question will appear:

When will samples be collected at the monitoring sites?
a) during calibration sampling (i. e., validation)
b) in the future (i.e., monitoring):

If you type a **[↵],** these sites will be listed on the crew log-sheet as validation points. If you type **b [↵],** these sites will be specified on the log-sheet as monitoring points.

NOTE: The EMCCRSD program will assign a "cup code" to every survey site which is selected as a sample site; these cup codes are printed on the crew log-sheet next to the sample site ID numbers. All calibration sites automatically receive cup codes; however, monitoring sites only receive cup codes if you define these sites to be validation sites. If desired, cup codes can be used as laboratory identification numbers.

After the sample site map and crew log-sheet are created, EMCCRSD will print to the screen one final question:

Do you wish to produce a spatial map of the 1st principal component score?
Note: answer yes only if you used a rectangular, systematic survey grid.. .

Create this map (y/n):

Type **y [↵]** if you wish to produce this map. If you answer yes, then a four level raster map of the 1st principal component score will be produced and written to the PC1.MAP text file. Note that this raster map will always appear square, regardless of the actual field dimensions. Hence, if the ratio of the horizontal to vertical field boundaries is significantly different from 1, then this map will tend to distort the true spatial pattern of the principal component score.

NOTE: The spatial estimation technique used to create this map may produce unreliable results if a highly non-systematic sampling grid is used to collect the survey data.

Upon completion of the spatial map estimation routine, the program will write a message to the screen reminding you to print out and/or rename the various output text files, and then terminate.

3.4 Tutorial Example

Type emccrsd **[↵]** to initiate the EMCCRSD program (as described in the first paragraph of Section 3.3). Answer no at the help prompt, confirm that the input file information is correct, and then type the **[↵]** key to begin the calibration site selection process. Enter 2 **[↵]** at the support site prompt and answer no at the multi-stage survey prompt. Next, answer yes at the monitoring site inclusion prompt, and type a **[↵]** to identify these 8 sites as validation points. Finally, type **y [↵]** at the map prompt to generate a spatial map of the 1st principal component score.

When the program terminates, print out the SITES. TXT, SSS.MAP, CRWSHEET.TXT, and PC1.MAP text files; their output should match the text file output shown in tables 3.1 through 3.4.

After the last text file has been printed, type the following command at the DOS prompt:

copy emccrsd.out c:\emsurvey\data\wwd1 .out

This will copy the *EMCCRSD.OUT* text file into a text file called *WWD1.OUT* (and place the *WWD 1.OUT* file into the c:\emsurvey\data subdirectory). In Section 5.4 you will learn how to merge this information with the sample soil salinity information in order to create a valid input file for the *EMSMLR* program.

You may now wish to run the *EMCCRSD* program again and change your answers at the monitoring site inclusion prompt, to better understand how these answers effect the SSS. *MAP* and *CRWSHEET. TXT* output files.

3.5 Multi-stage Sampling Designs

Occasionally, there may be situations where you wish to acquire more than one set of monitoring sites. For example, you may wish to collect both calibration and validation sites within your EM surveyed field, and also select two sets of monitoring sites to sample at in the future. These types of sampling plans, referred to in this manual as "multi-stage sampling designs", can be generated by repeatedly running the *EMCCRSD* program on the same survey data set. Multi-stage sampling designs can be very useful if you need to monitor a field over a long period of time, and/or need to collect soil samples at multiple points throughout the life of an experiment (in order to measure the rate of change in the geometric mean field salinity level with respect to a change in the experimental conditions).

When you answer yes at the multi-stage survey prompt, an output text file called *EMCCRSD.IN2* is created. This file can then be used as a new input file for the *EMCCRSD* program. Executing the program again produces a new, second set of calibration sites (which can be used as either additional validation sites or as future monitoring sites) and, if requested, a new *EMCCRSD.IN2* text file. This process may be repeated more than once, and hence used to generate multiple sets of monitoring sites.

In the example which follows, the *EMCCRSD* program is run three times in order to generate a sampling design with 20 calibration sites, 8 validation sites, and 2 sets of monitoring sites, with 15 sites in each monitoring set.

Table 3.1 Example of ASCII text output contained in *SITES. TXT.*

Spatial CCRSD Calibration & Monitoring Site Information

Title: Westland Field 1

Total # of survey sites = 178
Total # of PC scores per depth = 3
Total # of calibration sites = 16

Central Composite Spatial RSD Calibration Sites

Site ID	X	Y	PC scores			RSD levels		
30	5.8750	2.9250	1.46	0.63	1.12	1.13	1.13	1.13
104	2.8750	0.7250	-1.04	-0.89	0.88	-1.13	-1.13	1.13
25	6.3750	0.7250	-0.89	0.97	-1.03	-1.13	1.13	-1.13
114	2.8750	6.2250	1.04	-1.00	-0.84	1.13	-1.13	-1.13
160	0.8750	2.9250	-0.99	1.44	1.09	-1.13	1.13	1.13
71	4.3750	4.0250	0.95	-0.99	1.00	1.13	-1.13	1.13
80	3.8750	1.8250	-1.16	-1.70	-1.23	-1.13	-1.13	-1.13
89	3.8750	6.7750	1.00	1.24	-1.40	1.13	1.13	-1.13
170	0.3750	6.7750	1.44	-0.12	-0.33	1.96	0.00	0.00
180	0.3750	1.2750	-1.49	0.33	0.12	-1.96	0.00	0.00
14	6.3750	6.7750	0.20	2.35	-0.37	0.00	1.96	0.00
124	2.3750	3.4750	-0.16	-2.08	-0.63	0.00	-1.96	0.00
61	4.8750	5.6750	-0.31	-0.44	1.70	0.00	0.00	1.96
164	0.8750	5.1250	0.49	0.10	-2.00	0.00	0.00	-1.96
18	6.3750	4.5750	-0.59	0.57	-1.24	spatial support		
127	2.3750	1.8250	-1.07	-0.83	0.30	spatial support		

Monitoring/Validation Sites

Site ID	X	Y	PC scores			RSD levels		
29	5.8750	2.3750	1.46	1.01	0.85	1.13	1.13	1.13
100	3.3750	2.3750	-0.70	-1.38	0.52	-1.13	-1.13	1.13
159	0.8750	2.3750	-1.16	1.42	-1.11	-1.13	1.13	-1.13
117	2.3750	7.3250	0.85	-1.23	-1.07	1.13	-1.13	-1.13
9	6.8750	5.1250	-0.53	1.16	1.24	-1.13	1.13	1.13
97	3.3750	4.0250	0.89	-0.73	0.64	1.13	-1.13	1.13
79	3.8750	1.2750	-1.42	-0.99	-0.83	-1.13	-1.13	-1.13
36	5.8750	6.2250	0.88	1.07	-0.88	1.13	1.13	-1.13

Table 3.2 Example of ASCII text output contained in *SSS.MAP*.

```
Soil Sample Site X/Y Location Grid

Title:  Westland Field 1

. = survey site          $ = calibration site
+ = masked survey site   o = monitoring/validation site

   .   .   .   .   o   .   .   .   .   .   .   .   .

   $   .   .   .   .   .   .   $   .   .   .   .   $   .

   .   .   .   .   .   $   .   .   .   .   .   o   .   .

   .   .   .   .   .   .   .   .   .   $   .   .   .   .

       $   .   .   .   .   .   .       .   .   .   .   o

   .   .   .   .   .   .   .   .   .   .   .   .   $   .

   .   .   .   .   .   .   o   .   $   .   .   .   .   .

   .   .   .   .   $   .   .   .   .   .   .   .   .   .

   .   $   .   .   .   .   +   .   .   .   .   $   .   .

   .   o   .   .   .   .   o   .   .   .   .   o   .   .

   .   .   .   .   $   .   .   $   .   .   .   .   .   .

   $   .   .   .   .   .   .   o   .   .   .   .   .   .

   .   .   .   .   .   $   .   .   .   .   .       $   .
```

Table 3.3 Example of ASCII text output contained in CRWSHEET. *TXT.*

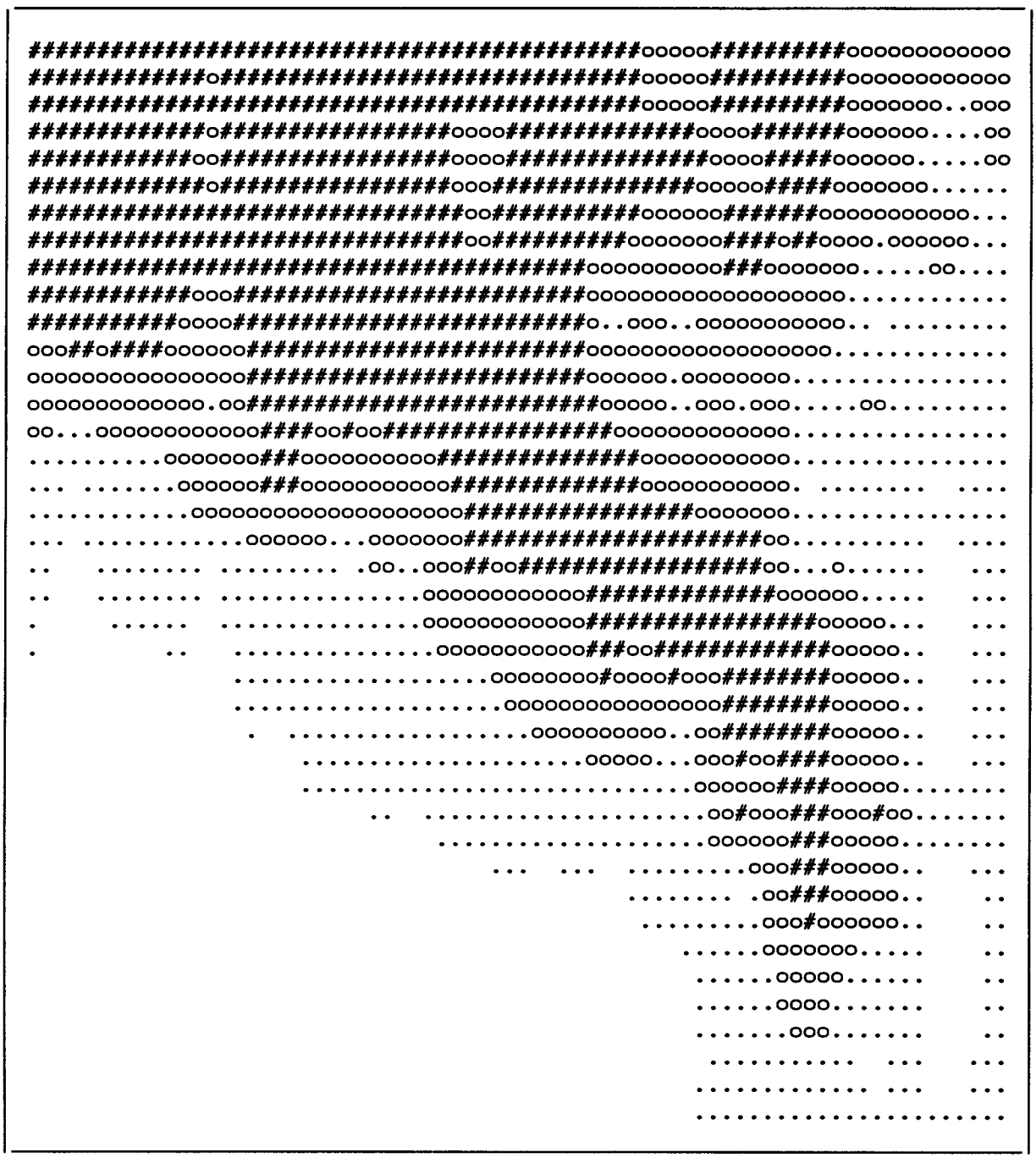
```
Sampling Crew Log-Sheet:  Ordered soil sample sites
Title:                    Westland Field 1
Sampling Date:            -       -       -

  9   cup #:   1  --> Validation Site Notes: _____
 14   cup #:   2     Calibration Site Notes: _____
 18   cup #:   3     Calibration Site Notes: _____
 25   cup #:   4     Calibration Site Notes: _____
 29   cup #:   5  --> Validation Site Notes: _____
 30   Cup #:   6     Calibration Site Notes: _____
 36   cup #:   7  --> Validation Site Notes: _____
 61   Cup #:   8     Calibration Site Notes: _____
 71   cup #:   9     Calibration Site Notes: _____
 79   cup #:  10  --> Validation Site Notes: _____
 80   cup #:  11     Calibration Site Notes: _____
 89   cup #:  12     Calibration Site Notes: _____
 97   cup #:  13  --> Validation Site Notes: _____
100   cup #:  14  --> Validation Site Notes: _____
104   cup #:  15     Calibration Site Notes: _____
114   cup #:  16     Calibration Site Notes: _____
117   cup #:  17  --> Validation Site Notes: _____
124   Cup #:  18     Calibration Site Notes: _____
127   cup #:  19     Calibration Site Notes: _____
159   cup #:  20  --> Validation Site Notes: _____
160   cup #:  21     Calibration Site Notes: _____
164   cup #:  22     Calibration Site Notes: _____
170   Cup #:  23     Calibration Site Notes: _____
180   cup #:  24     Calibration Site Notes: _____
```

Table 3.4 Example of ASCII text file map contained in *PC1.MAP*.

```
Relative 1st PC Spatial Distribution (assuming Normality)
Title:  Westland Field 1

[    ]   0% < Quantile < 25%            pc1 < -0.67
[...]  25% < Quantile < 50%    -0.67 < pc1 <  0.00
[ooo]  50% < Quantile < 75%     0.00 < pc1 <  0.67
[###]  75% < Quantile < 100%           pc1 >  0.67
```



Step 1.

Initiate the ***EMCCRSD*** program. Enter 6 **[↵]** at the support site prompt and answer yes at the multi-stage prompt. The following question will appear:

Do you wish to mask out the eight 1st stage monitoring sites? (y/n):

Answer yes. Next, answer yes at the monitoring site inclusion prompt, type **a [↵] to** identify these 8 sites as validation points, and answer no at the map prompt to terminate the program. At the DOS prompt, type the following file copy commands:

copy sites.txt sites.1 [↵]
copy sss.map sss.1 [↵]
copy crwsheet.txt crwsheet. 1 [↵]
copy emccrsd .in emccrsd. 1 [↵]
copy emccrsd.in2 emccrsd.in [↵]

Step ***2.***

Initiate the ***EMCCRSD*** program a second time (note that the program now detects 29 masked survey sites). Enter 1 **[↵]** at the support site prompt and answer yes at the multi-stage prompt. The following question will again be printed to the screen:

Do you wish to mask out the eight 1st stage monitoring sites? (y/n):

Answer no. Also answer no at the monitoring site inclusion prompt and spatial map prompt. At the DOS prompt, type the following second set of file copy commands:

copy sites.txt sites.2 [↵]
copy sss.map sss.2 [↵]
copy crwsheet.txt crwsheet.2 [↵]
copy emccrsd.in emccrsd.2 [↵]
copy emccrsd.in2 emccrsd.in [↵]

Step ***3.***

Initiate the ***EMCCRSD*** program a third and final time. (The program will now report 44 masked survey sites.) Enter 1 **[↵]** at the support site prompt and answer no at all remaining program prompts. At the DOS prompt, use the following file print commands to view the site information and final sampling maps generated by this three-stage sampling design:

print sites.1 [↵] {to see the information pertaining to the 20 1st stage
print sss.1 [↵] calibration and 8 1st stage validation sites}
print crwsheet. 1 [↵]

print sites.2 [↵] {to see the information pertaining to the 15 2nd stage
print sss.2 [↵] monitoring sites}
print crwsheet.2 [↵]

print sites.txt [↵] {to see the information pertaining to the 15 3rd stage
print sss.txt [↵] monitoring sites}
print crwsheet.txt [↵]

If you follow the instructions given in the example above, you should find that the site numbers listed in *CRWSHEET. 7, CRWSHEET.2,* and *CRWSHEET. TXT* match the numbers shown below:

CR WSHEET. 7 :
calibration sites:
{14,18,22,25,30,53,61,71,80,89,104,111,114,124,127,141,160,164,170,180}
validation sites: {9,29,36,79,97,100,117,159}

CRWSHEET.2:
calibration sites [to be used as 2nd stage monitoring sites]:
{3,15,20,44,52,83,88,95,107,128,140,148,152,165,181}

CRWSHEET. TXT:
calibration sites [to be used as 3rd stage monitoring sites]:
{2,11,19,48,59,65,74,106,112,116,122,144,153,172,178}

4.0 EM SURVEYING AND SOIL SAMPLING CONSIDERATIONS

4.1 Spatial Variability

The ***VALIDATE*** and ***EMCCRSD*** programs are designed to optimize the choice of sampling locations with respect to describing the spatial variability present within your survey area. These soil salinity data are then used to estimate a regression equation which, in turn, can be used to predict the soil salinity levels at all of the remaining non-sampled locations. The sampling design used in the ***EMCCRSD*** program actually exploits the spatial soil salinity variability present in the field. This is done by deterministically selecting a sample data set which will typically be more appropriate (for estimating the regression equation) than one would expect to observe under simple random sampling strategies.

However, there are a number of additional factors that can cause both soil salinity and EM signal variation in irrigated farmland, and which must be accounted for to implement an efficient soil sampling plan. Furthermore, unlike spatial variation (which can, at least in part, be considered stochastic), many of these additional sources of variability are typically induced by management practices. In this section we will summarize some of the additional deterministic and stochastic mechanisms which create salinity and EM signal variations in irrigated farmland, and discuss appropriate surveying and sampling methods to minimize their effects.

4.2 EC, Variation Induced by Sampling Depth

Probably the most common source of EC, variation comes from fluctuation in salinity levels with respect to sampling depth. Soil salinity levels can change quite rapidly with depth; it is not unusual to see the relative salinity level increase an order of magnitude within the first 1 .O meter of soil. Maintaining an accurate and consistent sampling depth throughout the survey area is therefore critical, and sampling practices which minimize depth variations must be rigorously followed.

It should be pointed out that many EM instruments acquire a depth weighted signal reading throughout the soil profile. The Geonics EM-38 meter is one example of such an instrument; both the horizontal and vertical EM-38 readings represent depth weighted average conductivity throughout the first 1 .O to 2.0 meters of soil. Without prior knowledge of the salinity profile shape, it can be difficult to infer the exact depth of maximum salinity concentration. For this reason we recommend that soil cores be acquired to a depth of at least 1 .O meter at each sample site. If resources permit, each core can be sliced into subsamples, thereby facilitating the estimation of prediction functions (regression models) for multiple sample depths.

One sampling strategy we commonly use is to acquire soil samples at each sample site in 30 centimeter increments, typically down to a depth of either 1.2 or 1.5 meters. When sampling by hand (i.e., using a hand auger), each soil sample can be removed individually. If a drilling rig is available, then the entire core is usually bored at one time and then split into subsamples after being brought to the surface. However, regardless of the actual sampling techniques, the surveying crew should always attempt to absolutely minimize any depth variations between sample sites.

As discussed in Section 1.2, we also strongly recommend that a second type of electromagnetic induction survey instrument be simultaneously used along with the EM-38 during all survey work. Insertion four probes and small, hand held wenner arrays are both very useful for measuring the soil conductivity within the first 25 to 50 centimeters of topsoil. This added EM signal information can greatly increase the accuracy of the fitted prediction functions, particularly those associated with the near-surface sampling depths.

4.3 EC, Variation Induced by the Bed-Furrow Environment

The micro, bed-furrow environment can be another source of considerable salinity variation, particularly towards the end of the cropping season (or throughout the season in a fixed bed system). A percentage of the irrigation water deposited into the furrows will absorb upwards into the bed, due to preferential capillary flow. This water movement in turn will carry the majority of near surface soluble salts up into the bed. In flood irrigated fields the relative difference between the near surface furrow and bed salinity levels can become quite pronounced over time.

Figure 4.1 displays the geometrical distribution of soil salinity throughout the near surface bed-furrow environment within a fixed bed, flood irrigated cotton field in the Coachella Valley, California (sampled in 1992). Note that at the high mean salinity level (23.2 dS/m throughout the bed-furrow environment), the ratio of bed to furrow near surface salinity levels was 4:1 At the low level (5.8 dS/m) this ratio actually increased to 8:I Figure 4.1 indicates that the overall mean salinity level would have been very poorly estimated by samples acquired either only in the furrows or in the bed. (In this particular survey, it took 14 soil samples at each sample site to adequately describe the 2-dimensional pattern of salinity present within the bed-furrow environment.)

The main point here is not to suggest that such extensive sampling schemes must always be employed within bed-furrow systems (unless, of course, estimating the micro, 2-dimensional salinity distribution is the primary goal of the survey). Rather, it is to point out the critical need for consistency with respect to borehole locations; all soil cores should be sampled from the same place within the bed-furrow

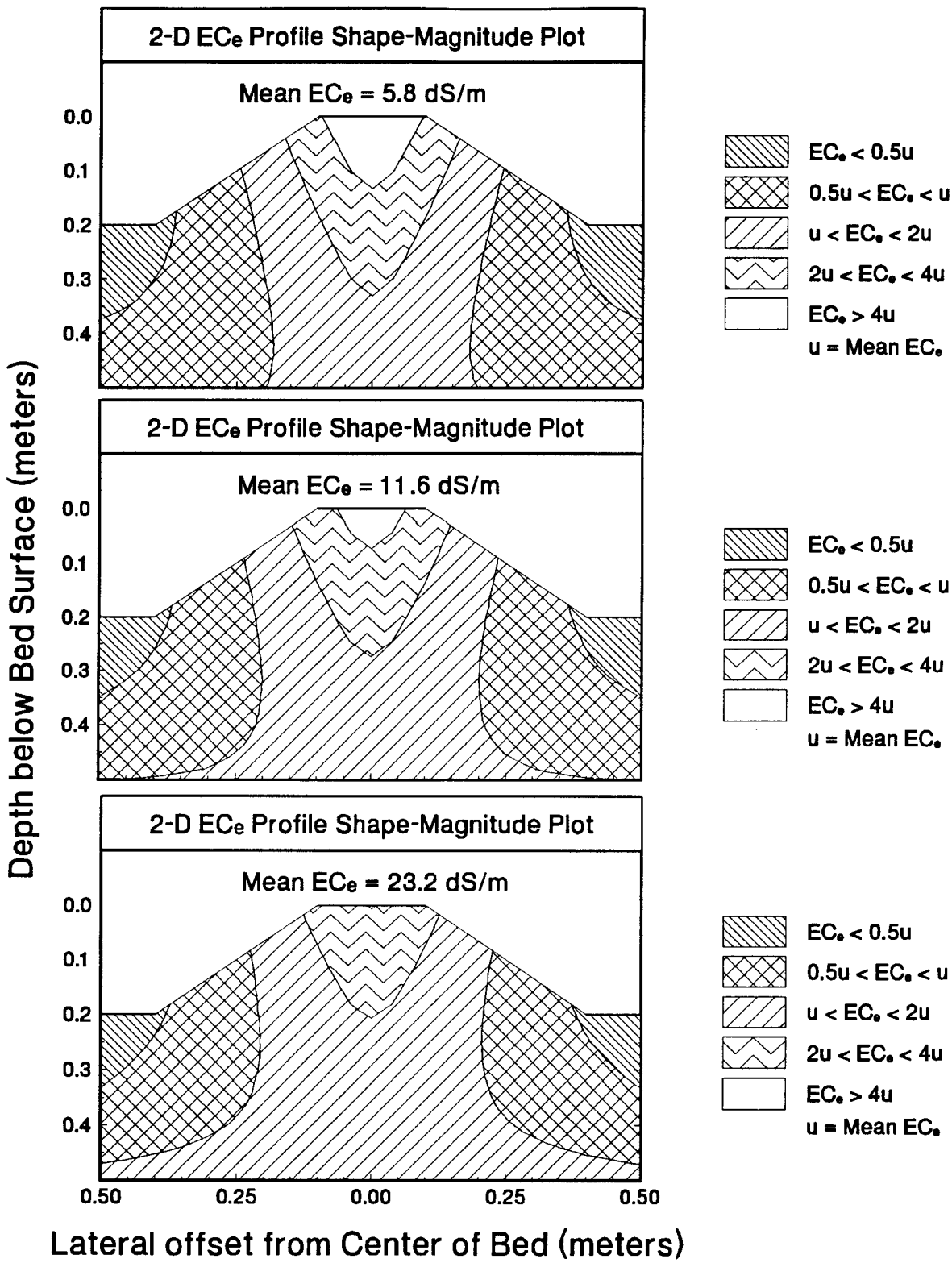


Figure 4.1 2D salinity distribution within the bed-furrow environment of a fixed bed system.

environment. Furthermore, the EM survey data and sample soil cores should be acquired from exactly the same area within this environment. If the EM signal data are acquired over the furrows, then the soil samples should also be acquired from the furrows, etc.

Some authors have suggested using composite sampling strategies (also known as "bulking") for averaging out bed-furrow variations. In composite sampling, soil samples would be acquired from both bed and furrow locations and then mixed together in an effort to obtain a more "representative" sample. We generally do not recommend such strategies for two reasons. First, it doubles the field work without providing any knowledge of the 2-dimensional, bed-furrow salinity distribution and second, it often introduces more variability into the sample data (through poor mixing processes) than it removes through averaging.

4.4 EC, Variation Induced by Traffic Patterns

Another source of potential EC, variation arises from soil compaction caused by repetitive traffic operations. In many fields, tractor operators consistently drive down the same set of furrows when performing various tillage and cultivation operations throughout the growing season. This leads to a systematic pattern of excessive compaction in a subset of furrows throughout the field, which in turn can cause a cyclic variation in the irrigation water infiltration rate (Wu et. al., 1995).

Figure 4.2 displays EM-38 horizontal readings acquired along 30 adjacent furrows in a buried drip irrigated cotton field subject to repetitive traffic influences (Westlands Water District, 1991). In this case the traffic pattern induced a clearly cyclic pattern in the EM, readings; the highest conductivity readings consistently occurred in the compacted furrows.

The data shown in Figure 4.2 is rather atypical; in general, we have not found compaction induced cyclic patterns in soil conductivity to be this pronounced in most fields. None the less, it is a good idea to systematically avoid the compacted furrows when conducting the EM survey if obvious traffic patterns are present in a field. Excessive soil compaction will nearly always have at least some effect on the soil salinity levels. Random surveying (and sampling) of both compacted and non-compacted furrows in the same field can introduce additional variability into the regression model salinity predictions, and should be avoided whenever possible.

4.5 EC, Variation Induced by Irrigation Management Practices

Irrigation management has a pronounced effect on determining the apparent salinity distribution within a field, both spatially and with respect to depth throughout

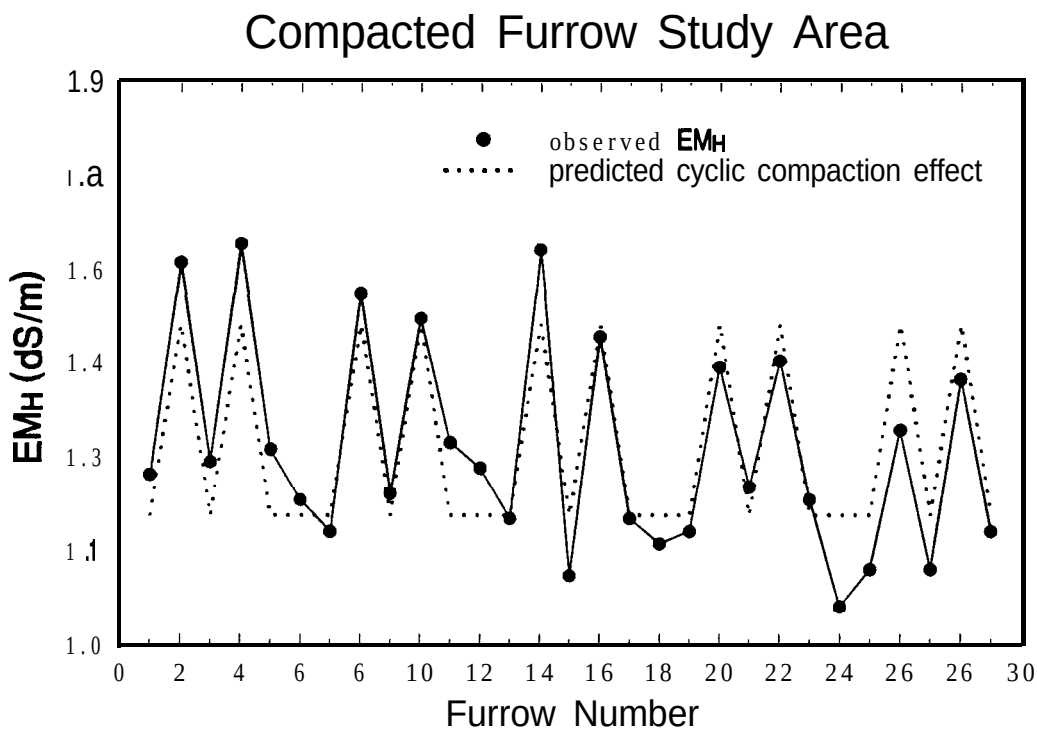


Figure 4.2 Tillage equipment compaction effect on the observed soil conductivity in a drip irrigated cotton field.

the profile. The amount and frequency of irrigation will directly influence the movement of soluble salts through the profile and across the field. The percent water content of the soil during the electromagnetic survey will also be at least partially determined by the elapsed time from the last irrigation event.

Since a change in irrigation management can seriously effect the 3-dimensional salinity distribution within a field, it is important to avoid collecting signal readings (and/or soil samples) from an area under more than one irrigation management strategy during the electromagnetic survey. In practice, this means that any survey area must be restricted to farmland under identical water management practices. For example, suppose you wish to survey a 160 acre cotton field which has been sub-divided into four 40 acre sections, where each section has been subject to different irrigation techniques. Under such a scenario, you cannot conduct a single survey across all 4 sections together; each 40 acre section must be individually surveyed and soil sampled. This is necessary because each survey must be conducted entirely within a single, homogeneous irrigation management area.

The critical importance of “blocking” during an electromagnetic survey should never be overlooked. The failure to restrict electromagnetic readings to an area under

a single water management strategy can result in serious regression model bias and inflated prediction errors, and corrupt the entire surveying process.

4.6 EC, Variation Induced by Deviations **in Surface Elevation**

Most farmland is typically laser-leveled to improve irrigation efficiency. Various leveling designs are used, depending on the method of irrigation and the agricultural crop under production. The three most common designs include (1) dead leveling, (2) single-slope leveling, and (3) dual slope leveling. All three of these designs create a theoretical plane which can be written mathematically as:

$$\text{Surface Elevation} = a, + \alpha_1 x + \alpha_2 y, \qquad (4.1)$$

where x and y represent the physical (x,y) coordinates, and the a,, a,, and α_2 coefficients determine the primary and secondary slopes of the field. Note that in the statistical literature, equation 4.1 is known as a "first order trend surface equation" (Box & Draper, 1987).

The *EMSMLR* program can incorporate both first and second order trend surface equations into the fitted salinity prediction model. Hence, any linear and/or quadratic surface elevation effects can be explicitly compensated for. These trend surface equations can also be used to compensate for gradual changes in both soil texture and water content across the survey area, if any such changes effect the soil salinity distribution and/or electromagnetic signal response.

Occasionally, you may have to survey non-graded farmland which exhibits significant local variation in surface elevation. In such a scenario it will usually be necessary to conduct a surface elevation survey along with the electromagnetic survey. This elevation data can then be directly incorporated into the regression model, if this data contributes to an improvement in the prediction accuracy.

The *EMSMLR* program does not presently allow for the explicit inclusion of surface elevation data in the modeling process. However, there are two ways you can implicitly include elevation data into the prediction equation. First, you can incorporate the surface elevation data directly into the sampling design by including these data in the input survey file for use in the *VALIDATE* program. In other words, one of your columns of electromagnetic signal readings would actually be surface elevation data. This approach can be advantageous if there are severe local deviations in the surface elevation across the field, and if you wish to ensure that your sampling locations adequately reflect these deviations. Alternatively, you can collect the ancillary elevation information during the survey process, and simply replace the third principal component column in the *EMSMLR* input file with the surface elevation readings during the modeling stage of the analysis. In this latter

approach, the elevation data can be easily discarded **if it does not significantly improve the prediction accuracy of the regression model.**

4.7 Electromagnetic Signal Variation Induced by Changes in Soil Texture

Of all the potential secondary factors which contribute to increased EM **signal variability, fluctuations in** soil texture are the most troublesome and difficult to adjust for. Textural variations across the field can rarely ever be measured in enough detail to facilitate either efficient blocking strategies or the explicit inclusion of such information directly into the regression model. Additionally, significant textural variation often effects both the spatial EM signal data and the 3-dimensional soil salinity pattern in dissimilar ways.

The nature and degree of textural variation will ultimately determine whether a regression modeling approach can successfully estimate the spatial salinity pattern in your survey area. It has been our experience that the regression modeling approach works well in the following two types of textural variation scenarios: (1) nearly all fields with minimal soil texture variability, and (2) most fields with moderate soil texture variability, provided either the texture changes in a smooth and gradual manner across the field, or the texture and salinity variations are strongly correlated. On the other hand, we have seen the regression modeling approach fail in fields with severe, chaotic texture variations, relatively low soil salinity levels, and poor correspondence between the soil texture and salinity readings.

Fortunately, most all irrigated farmland typically encountered during electromagnetic survey work falls into one of the first two categories described above. Furthermore, if a chaotic, low saline field is encountered, the series of statistical tests and prediction diagnostics contained within the *EMSMLR* program will reveal the degree of bias in the fitted regression model. A quantitative decision can then be made as to the merits of collecting additional samples and/or pursuing other modeling techniques.

An example data set displaying high texture and low salinity variability (AZ09.DAT) is included with the ESAP software. These data can be used to demonstrate the various *EMSMLR* residual assessment (model validity) tests, and will be discussed in detail in section 7.4.

4.8 Electromagnetic Signal Variation Induced by Other Soil Properties

Other physical soil properties also theoretically affect the EM signal readings to various degrees. While these properties do not control and/or influence the soil salinity directly, they can increase the apparent salinity variation by systematically

confounding the signal response data. Examples of these properties include the soils organic matter level, magnetic susceptibility, temperature, and water content level.

In most practical applications, significant variation in these properties needs to be present before any meaningful corruption of the signal reading occurs. We have not encountered a field with enough variation in either the organic matter level or magnetic susceptibility to noticeably affect any type of EM reading. Likewise, a one degree centigrade change in temperature throughout the entire soil profile typically causes no more than a two percent change in the EM-38 signal readings. Since soil temperature fluctuations below the 30 centimeter depth level in the soil profile occur rather slowly, the entire survey process can usually be completed before a significant change in the bulk-average soil profile temperature occurs.

Soil water content variations do effect the apparent conductivity readings. However, in areas under uniform irrigation management practices, the degree of spatial water content variability is typically minimal (provided significant soil texture variation is not present). Furthermore, a gradual fluxuation in the soil water content level across the survey area can usually be compensated for through the use of trend surface parameters in the regression model. Note also that variation in the water content level through the soil profile (i.e., by depth) will not effect the regression model prediction accuracy, provided the water content levels vary with depth in a reasonably consistent manner throughout the field.

It is important to remember that if the water content of the soil drops too low (i.e., the soil becomes too dry) then the electromagnetic signal readings can become seriously dampened. Indeed, if the survey area is absolutely void of any moisture throughout the entire soil profile, then it will be impossible to measure any conductance whatsoever. In most practical applications, we have found that good EM signal data can be obtained providing the survey area (field) is at or above 30 percent field capacity. As already mentioned, the surveying of especially dry areas should be avoided, since the conductivity of dry soil is not reflected in the signal information.

4.9 Assessing Short Scale (Nugget) Salinity Variation

It is not uncommon to observe a certain degree of variation in soil salinity levels over very short distances. For example, near surface furrow samples acquired 50 cm apart can vary anywhere from 10 to 50 percent due to micro-scale soil composition characteristics and/or fluctuations in preferential water flow.

Electromagnetic induction instrument readings tend to average out this micro-variation. The degree to which this averaging occurs depends directly upon the instruments “foot-print”; e.g., the volume of soil incorporated into the signal

response. For example, an EM-38 signal will be influenced by any electrically conductive material within about 1 to 2 meters of the instrument (both laterally and vertically). Hence, it is typically assumed to have about a 1.5 meter foot-print.

Because the volume of soil measured by the EM-38 is so much larger than the volume obtained by conventual soil sampling techniques, an estimate of the degree of short scale salinity variation needs to be acquired for model validation purposes. Such an estimate can be acquired by obtaining "replicate" sample cores (two sample cores spaced about 50 centimeters apart) at some of the calibration sites during the soil sampling process. The replicate cores can then be used to estimate the short scale salinity variation (referred to as the "nugget variation" in geostatistical models and as the "pure error estimate" in spatial regression models). The measured salinity data from these cores also allows the *EMSMLR* program to construct a very useful residual autocorrelation test, known as a "lack-of-fit" test, for assessing the spatial residual independence assumption (Lesch et. al., 1995a).

In a standard 16 site calibration sampling design, we commonly acquire replicate sample cores at 4 to 6 of the 16 calibration sites. These four to six sites can either be chosen at random (from amongst the 16 sites) or selected throughout the survey area (which is usually preferable). Additionally, the core separation spacing (usually 50 centimeters) should be the same at all sites and both the primary and replicate cores should always come from the same location with respect to the bed-furrow environment (i.e., both from the bed, or both from the furrow).

Techniques for estimating the short scale salinity variation and performing residual lack-of-fit tests are described in Lesch et. al., 1995a. These estimates and testing procedures are automatically calculated by the *EMSMLR* program when replication salinity data are available (Section 5.3).

4.10 Surveying & Sampling Considerations: An Overview

We have described a number of deterministic and stochastic mechanisms which can create both soil salinity and EM signal variation in irrigated farmland, and discussed appropriate sampling methods to minimize their effects. A summary of these methods is given on the following pages, along with some additional tips and/or suggestions for implementing effective surveying and sampling practices.

A. EM Surveying Techniques

1. Decide on the survey grid size before beginning the survey process. Use only centric, systematic grids (square, rectangular, or triangular) and keep the total number of survey sites at a manageable level so that the survey process can

be completed in a reasonable time-frame. For example, in a 40 acre field, a 12 by 12 grid of survey sites can be acquired in one day, and is usually more than adequate for producing good field median EC, estimates and a reliable salinity map.

2. Acquire all EM-38, wenner, and/or four-probe readings in the same locations with respect to the bed-furrow environment at every survey site (for example, only in the furrows). If four-probe readings are being acquired, try to take either two readings (spaced 50 centimeters apart in the furrow) or three readings (spaced 30 centimeters apart in the furrow) at each site, and use the average of these two or three readings as the input four-probe data within the *VALIDATE* program. (Averaging your four-probe data will produce a four-probe foot-print which is more consistent with the size of the EM-38 footprint, and reduce the four-probe signal variability induced by micro-salinity fluctuations.) Collect only near surface four-probe data (for example, a O-30 cm depth reading) to minimize the survey workload at each site.

3. Leave some sort of identifiable marker at each survey site (such as a numbered bag) so that the exact location of each site can be ascertained during the soil sampling phase of the survey.

4. Make sure that your survey area is contained entirely within an area under a single irrigation management practice. (This typically implies that you limit your survey area to an individual field.)

5. Try to conduct the survey on soil at or above 30 percent field capacity, and make sure that the soil water content is relatively consistent across the entire survey area.

6. Avoid collecting survey data in obviously compacted furrows.

7. If possible, try to limit your survey area to landscape with reasonably similar soil texture characteristics.

B. Soil Sampling Techniques

1. Acquire the soil cores (at all the selected calibration sites) directly over the location of the EM signal readings, and keep the sampling depth consistent from site to site. When acquiring samples at multiple depths, always attempt to absolutely minimize any sample depth variations between sites.

2. Avoid including the dry, crumbly, immediate topsoil in the near surface soil sample. This topsoil “fluff” is not reflected in any of the EM signal readings (because it has zero moisture content), nor is it typically reflective of the

average salinity level within the first 25 to 50 centimeters of the soil profile.

3. Collect replicate samples at four to six of the calibration sites. Sample the primary and replicate soil cores 50 centimeters apart, such that each core is approximately 25 centimeters from the center of the location where the EM signal data was acquired. Be sure to acquire all primary and replicate calibration samples from the same location with respect to the bed-furrow environment at every site.

4. Decide on a reasonable identification system for the soil samples, and number all the soil sample containers before beginning the sampling process. (You're much less likely to make container labeling errors if all the labeling is completed before hand.) Use the ***SITES. TXT*** and ***SSS.MAP*** printouts produced by the ***EMCCRSD*** program to identify the calibration site locations. Additionally, use the ***CRWSHEET. TXT*** printout while soil sampling to keep track of the sampling progress and to record any unusual observations about the individual soil samples; such as depth to water table, abrupt changes in soil texture, etc.

5.0 EMSMLR PROGRAM DOCUMENTATION'

5.1 Program Description

EMSMLR is a comprehensive, menu driven spatial regression modeling software program designed to predict the soil salinity levels across a survey area, based on the observed EM signal and calibration soil sample data.

EMSMLR reads as input an ASCII text file containing the calibration sample soil salinity levels, along with the survey output data from the *EMCCRSD.OUT* text file (created by the *EMCCRSD* program). It then allows the user to estimate and validate an optimal spatial regression model for predicting the soil salinity levels at all non-sampled survey locations, creates a salinity diagnostic report for the survey area, and produces an output file suitable for use in the *SALTMAP* program. The *EMSMLR* program can also be used to test for changes in the field geometric mean salinity level over time, if salinity data are acquired at one or more preselected monitoring sites in the future.

EMSMLR creates a number of different output files which document the various model statistics, parameter estimates, and residual tests performed during the analysis. Nearly all of these files can be either printed out while still running the program or written to the hard disk for future reference. A detailed explanation of the file naming system, along with a discussion of the contents of each file is given in Section 5.4.

5.2 An **Overview of the EMSMLR Menu System**

All components of the *EMSMLR* program are controlled through the use of a menu system. In order to successfully use the program, you must be familiar with both the operation of the menu system and the specific function that each menu subroutine performs. This section presents an overview of the menu system; the specific menu subroutine functions will be described in Section 5.3.

A hieratical layout of the *EMSMLR* menu system is shown in table 5.1. After initiating the program, the SMLR MODELING menu screen will appear (referred to in Table 5.1 as the main menu screen). This menu allows you to perform one of six distinct operations: (1) read an input file into the program, (2) identify one or more regression models most suitable for predicting the spatial soil salinity distribution throughout your survey area, (3) estimate and validate a specific regression model, (4) create a salinity diagnostic report for your survey area, (5) test for a change in the predicted field geometric mean salinity level over time, and (6) exit to DOS.

Table 5.1 A hieratical layout of the EMSMLR Menu System.

MAiN MENU SCREEN

SMLR MODELING MENU

1. **Read Input File**
2. **Model Identification**
3. **Model Estimation / Validation**
4. **Salinity Diagnostic Report**
5. **Site Monitoring: Net Flux Testing**
6. **Exit to DOS**

SUB LEVEL 1 MENU SCREEN

ESTIMATION / VALIDATION MENU

1. **Display Model Estimation Results**
2. **Display Residual Diagnostics**
3. **Display Prediction Diagnostics**
4. **Print Results**
5. **Return to Main Menu**

SUB LEVEL 2 MENU SCREENS

MODEL ESTIMATION SUBMENU

1. **Model Summary Statistics**
2. **ANOVA Tables**
3. **Parameter Estimates**
4. **Return to ESTIMATION / VALIDATION Menu**

RESIDUAL DIAGNOSTICS SUBMENU

1. **Univariate Statistics**
2. **Influence Diagnostics**
3. **Correlation Matrix**
4. **Moran Spatial Autocorrelation Tests**
5. **Residual Plots (Against Predictions)**
6. **Residual Plots (Against Locations)**
7. **Residual Plots (Normality Assessment)**
a. **Return to ESTIMATION / VALIDATION Menu**

PREDICTION DIAGNOSTICS SUBMENU

1. **Summary Predictive Criteria**
2. **Univ Stats: Prediction Estimates**
3. **Univ Stats: Prediction Variance Estimates**
4. **Prediction Plots: (Log Data)**
5. **Prediction Plots: (Back-Transformed Data)**
6. **Return to ESTIMATION / VALIDATION Menu**

FILE PRINT SUBMENU

1. **Print Model Estimation Results**
2. **Print Residual Diagnostics**
3. **Print Prediction Diagnostics**
4. **Write all Results to Files**
5. **Return to ESTIMATION / VALIDATION Menu**

After initiating the program, you must first perform option 1; i.e., read in an input file. Next, you can invoke option 2 (the model identification subroutine), or jump directly to option 3 (the model estimation and validation subroutines). Once a suitable regression model has been estimated and validated, you can invoke options 4 and/or 5. Note that you can quit the *EMSMLR* program at any time from this menu by invoking option 6.

The various regression model estimation and validation operations must be initiated from the ESTIMATION / VALIDATION menu screen (referred to as the sub level 1 menu screen in table 5.1). This screen will appear whenever you choose option 3 in the SMLR MODELING menu. This menu screen allows you access all model estimation and validation operations, which include (1) displaying the model statistics and parameter estimates, (2) displaying the residual plots and test statistics, (3) displaying the salinity prediction estimates and prediction plots, and (4) creating and printing the output text files. Note that option 5 on the ESTIMATION / VALIDATION menu screen can be used to get back to the SMLR MODELING menu screen,

When you choose one of the first four options from the ESTIMATION / VALIDATION menu, a new submenu screen will appear. Examples of each of these submenus are shown at the bottom of Table 5.1 (referred to as the sub level 2 menu screens). Choosing option 1 from the ESTIMATION / VALIDATION menu will cause the MODEL ESTIMATION submenu screen to appear. Likewise, choosing options 2, 3, or 4 will cause the RESIDUAL DIAGNOSTICS submenu, PREDICTION DIAGNOSTICS submenu, or FILE PRINT submenu screens to appear, respectively. Note that the specific estimation, validation, and/or file print subroutines can only be invoked from these submenu screens.

Option 3 in the SMLR MODELING menu is the only option associated with additional sub level menus; choosing any other option from this menu will invoke operations which are completely self contained (i.e., operations which do not branch off to other menu systems).

5.3 Program Operation: [Specific Menu Subroutine Functions]

To start the *EMSMLR* program, move to the c:\emsurvey\phase2 subdirectory and type emsmlr [↵] at the DOS prompt. The program will print some initial header information, and ask you to strike any key to continue. At this point, typing the [↵] key will send you into the SMLR MODELING menu, which is the program's main menu system. From here you can invoke the various program options discussed on the following pages.

1. **Read Input File:**

Choosing this option allows you to enter your input data file. The first statement printed to the screen will be:

Please enter the path/file name:

Note that your input data file can reside anywhere on your hard drive, although we recommend that you store your data in the c:\emsurvey\data subdirectory. After you type in the correct path and filename, the next statement printed to the screen will be:

Please enter the survey code (4 character max):

At this point, you must specify a four character survey code, such as "wwd1", which will be used as the first four characters in all output text file names (see Section 5.4). After you verify the input file and survey code information, the following information will appear on the screen:

File name:	C:\XXXX\XXX\XXX.XXX
Title:	xxxxxxxxxxxxxxxxx
Survey Code:	xxxx
Abbreviation of survey type:	XX/XX/XX
Number of survey sites:	xxx
Number of calibration sites:	xx
Number of sites with two samples per depth:	x
Number of sample depths:	x
Number of principal component scores:	x

You should use this information to verify that the correct file has been read into the EMSMLR program (this information is explained in more detail in Section 5.4).

Next, the following question will appear:

Verify input data (y/n):

You should answer no at this prompt, unless you want to view each line of input data.

After reading in your data, the program will print to the screen the mean, variance, and correlation matrix associated with the input calibration data. The following prompt will then appear:

View bivariate log salinity plots? (y/n):

Answer yes if you wish to view these plots. Next, another plot prompt will appear on the screen:

View the plots of 1st principal component against the log salinity data? (y/n):

Again, answer yes if you wish to view these additional plots. After these two plot prompts, the final statement printed to the screen will be:

Press *any key to return to main menu.*

2. **Model Identification:**

This option can be used to identify one or more spatial regression models which fit your sample data well.

In the *EMSMLR* program, you are allowed to choose from among 50 different parameter combinations for the final regression model (if you are using only 2 principal component scores, you can choose from 30 parameter combinations). There are two types of model parameters used in the regression equations; signal (S) parameters and trend surface (T) parameters. Signal parameters are associated with the principal component scores, and trend surface parameters are associated with the x,y location coordinates of your **survey data.**

EMSMLR uses the following abbreviation system for identifying parameter combinations in a regression model:

Signal Parameters	Abbreviation	Trend Surface Parameters	Abbreviation
PC1	S1-	none	TO
pc1, pc2	s2-	x	Tx1
pcl, pc2, pc12	S2i-	y	Ty1
pc1, pc2, PC3	s3-	**x, Y**	Tx1y1
pcl, pc2, pc12, pc3	S3i-	x, x^2	Tx2
		y, y^2	Ty2
		X, **Y,** x^2	Tx2y1
		X, **y,** y^2	Tx1y2
		X, **Y,** x^2, y^2	Tx2y2
		X, **y, xy,** x^2, y^2	Tquad

Hence, if the regression model contains parameters for the 1st, 2nd, and 3rd principal component scores, and parameters for linear drift in both the x and y directions, the model is abbreviated as S3-Tx1y1. Likewise, if the model includes parameters for the 1st and 2nd principal components scores, an interaction parameter between the 1st and 2nd scores, and a full second order trend surface equation, then the model is abbreviated as S2i-Tquad.

When you invoke option 2, a series of regression models which reflect all possible parameter combinations are fitted to the calibration data. After each model is estimated, both the PRESS residual statistic and average prediction variance estimate (APVE) are computed. When the program is finished estimating all of the models, these PRESS and APVE statistics are ranked and then the model abbreviations associated with the 10 best statistics in each category are displayed to the screen.

Both the PRESS and APVE statistics are summarized and displayed in two manners; first as an absolute score and then as a relative (scaled) score. There will be no difference in the ranking of the absolute and relative scores when you are modeling salinity data from only one sample depth. However, if you have acquired multiple sample depths, then these scores can differ. Furthermore, the ranking of the PRESS and APVE scores will typically differ, even if the salinity data comes from only one sample depth. Therefore, these scores should usually be used as a guideline to help you narrow down the choices of parameter combinations in your final regression model, rather than a rigorous criteria for choosing a single best model.

The mathematical details behind the estimation of each statistic are discussed in Lesch et. al., 1995. Intuitive definitions of each score are given below.

PRESS statistic: a jack-knifed, prediction sum of squares estimate; i.e., the squared difference between the observed and prediction Ln salinity level at each calibration site, computed by the regression model without the model being fitted to the Ln salinity data at that site.

APVE statistic: the theoretical, average prediction variance of the Ln salinity predictions associated with all the remaining, non-sampled survey sites, based on the fitted regression model and assuming that the model contains the correct number of parameters.

Both statistics are computed for each sample depth, and then summed across the sample depths to create the absolute and relative scores. The absolute scores are simply the sum of the individual statistics computed for each sample depth, divided by the total number of sample depths. In theory, the best regression model should produce the smallest absolute PRESS and APVE scores . The same holds true for the

relative scores (the best model should again produce the smallest scores); however, note that the relative scores are computed differently. After all 50 PRESS and APVE statistics have been estimated for all the sample depths, the minimum statistics from each sample depth are identified. All 50 PRESS and APVE statistics are then divided by these minimum statistics, and then summed across the sample depths. Hence, if k sample depths are acquired at each calibration site, then the minimum relative PRESS and APVE scores must be $\geq k$.

After the 10 best models in each score category are displayed to the screen, the following prompt will appear:

Print the full results? (y/n):

If you answer yes, the program will print four output text files, which will list the individual absolute and relative PRESS and APVE statistics for each model at each depth. If you answer no, then another prompt will appear:

Print the top 10 results? (y/n):

If you answer yes at this prompt, then you will receive a single printout which lists the top 10 models in each category, based on the absolute and relative PRESS and APVE summary scores. Regardless of whether you answer yes or no to either prompt, the program will create and save four ASCII text files which contain all of the individual statistics for each model (see Section 5.4).

Finally, the last prompt which will appear on the screen will be:

Press any key to return to main menu.

3. Model Estimation / Validation

Option 3 can be used to estimate and validate any one of the 50 different regression models analyzed by the model identification subroutine. Immediately after choosing option 3, the information menu shown on the following page will be printed to the screen:

Model Selection Menu

	PC Scores				TS Components	
7.	PC7				0.	none
2.	PC7	PC2			7.	X
3.	PC7	PC2	PC72		2.	Y
4.	PC7	PC2	PC3		3.	XY
5.	PC7	PC2	PC72	PC3	4.	x X^2
					5.	Y Y^2
					6.	x Y X^2
					7.	x Y Y^2
					8.	x Y X^2 Y^2
					9.	x Y XY X^2 Y^2

Please enter a 2-digit model code (70 - 59):

You must select the parameter combination (for the regression model) by specifying the appropriate two digit code. For example, if you wish to use the S3-Tx1y1 regression model, the two digit code would be 43. Likewise, 20 would be the correct two digit code for the S2-TO regression model; i.e., the model containing only the 1st and 2nd principal component scores and no trend surface parameters.

NOTE: If you collected only EM-38 signal data, you will be using only 2 principal component scores and you will have only 30 models to choose from. Therefore, lines 4 and 5 under the PC Scores column will not appear on the screen, and the valid range for the two digit model codes will be from 10 to 39.

After confirming your regression model selection, the ***EMSMLR*** program will print the following message to the screen:

Es tima ting model, please wait.. .

Once the program computations are completed, the ESTIMATION / VALIDATION menu will appear. At this point, you can access any one of the following four submenus; the MODEL ESTIMATION submenu, the RESIDUAL DIAGNOSTICS submenu, the PREDICTION DIAGNOSTICS submenu, and the FILE PRINT submenu. The specific subroutines associated with each of these submenus are shown in table 5.1 and described below.

From within the MODEL ESTIMATION submenu, choose option 1 to view the model summery statistics, option 2 to view the model analysis of variance (ANOVA) tables, and option 3 to view the model parameter estimates. The model summary

statistics include the R^2 values, adjusted R^2 values, mean square error (MSE) estimates, root MSE estimates, and back-transformed model coefficient of variation scores (estimated as CV = $100[(e^{MSE}-1)^{0.5}]$). The ANOVA tables display the degrees of freedom, sum of squares, and mean square error estimates, along with the F test scores and probability levels. If replicate sample cores have been collected, the lack-of-fit (LOF) test scores and probability levels will also be displayed here. The parameter information includes the individual parameter estimates, standard errors, t-test scores and probability levels. Note that when sample data from multiple depths are being modeled, all of the above information will be displayed for each depth.

There are seven separate options available from within the RESIDUAL DIAGNOSTICS submenu. These seven options allow you to test and/or display various features of the residual distribution(s), in order to access the validity of the regression modeling assumptions. To view the univariate statistics associated with the studentized residuals, choose option 1. These summary statistics include the mean, variance, skewness, quantile rankings, and residual stem-leaf (histogram) plots for each sampling depth. Option 2 can be used to list the individual studentized residual values (for each sampling depth), along with the hat leverage values associated with each sample calibration site. The raw (non-studentized) residual depth correlation matrix can be displayed using option 3, and the Moran spatial autocorrelation tests scores and approximate probability levels can be displayed using option 4. Options 5, 6, and 7 can be used to display various types of residual plots. Choose option 5 to plot the studentized residuals against the predicted Ln salinity levels, option 6 to plot the studentized residuals against the x and y sampling coordinates, and option 7 to create residual QQ plots (for assessing the residual normality assumption).

There are five options available from within the PREDICTION DIAGNOSTICS submenu. Options 1, 4, and 5 allow you to access how well the estimated regression model fits the observed sample data. Option 1 can be used to display the regression model PRESS and APVE statistics (for each sampling depth), while options 4 and 5 will display plots of the observed versus model predicted salinity levels in Ln and back-transformed units, respectively. Options 2 and 3 display some pertinent summary statistics concerning the regression model predictions associated with the remaining non-sampled survey sites. Option 2 can be used to display the mean, variance, skewness, and quantile estimates of the predicted salinity levels (in both Ln and back-transformed units) at all the non-sampled survey sites. Likewise, option 3 can be used to display the mean, variance, skewness, and quantile estimates associated with the theoretical prediction variance estimates.

The FILE PRINT submenu can be used to print and/or save most all of the information accessed through the MODEL ESTIMATION, RESIDUAL DIAGNOSTICS, and PREDICTION DIAGNOSTICS submenus. You should use option 1 to print the model estimation information, option 2 to print the residual diagnostics, and option

3 to print the prediction diagnostics. Additionally, choose option 4 if you wish to save all of this information to the hard disk.

NOTE: **The *EMSMLR* program can not print out or save to the hard disk any of the screen plots displayed from the RESIDUAL DIAGNOSTICS or PREDICTION DIAGNOSTICS submenus.**

The various estimation statistics and residual/prediction diagnostics discussed above can be used for two purposes. First, to determine if the general assumptions associated with the regression modeling approach are satisfied. This should be done by (1) checking for spatially correlated residuals using the LOF and/or Moran residual test scores -- significant scores imply that the residuals are spatially autocorrelated, (2) checking the constant variance assumption by plotting the residuals against both the sampling locations and the predicted log salinity levels -- the residual/location and residual/predicted plots should be devoid of any pattern, and (3) using the univariate residual statistics and residual QQ plots to check for outliers and to assess the residual normality assumption -- the residuals should fall along an approximately straight line in the QQ plots. Most of the remaining statistics and diagnostics can be used for the second purpose, which is differentiating between various regression model parameter combinations. The regression model summary statistics, parameter estimates, residual influence diagnostics, and the assorted prediction statistics and diagnostics can all be used to help select the final combination of model parameters. Ideally, the selected model should have at least some of the following properties at each sample depth; (1) a high R^2 and low MSE estimate, (2) significant parameter estimates, (3) no apparent residual outliers, and (4) reasonably low hat leverage scores. Furthermore, the predicted salinity levels should display a strong, linear relationship with the observed salinity data, on both the natural log and back-transformed scale.

NOTE: **Most all of the residual and/or prediction diagnostics used in the *EMSMLR* program are discussed in detail in Atkinson, 1985; Myers, 1986; and Weisberg, 1985.**

4. Salinity Diagnostic Report

Once a suitable regression model has been estimated and validated, a salinity diagnostic report can be generated. The generated report will include the following information; (1) field median point estimates, (2) range interval estimates, (3) mapping classification accuracy scores, and (4) spatial variation index scores. The point estimates consist of depth specific mean Ln salinity and median (geometric mean) salinity estimates for the entire survey area, along with their respective variances and confidence intervals. The range interval estimates display the percentage of the survey area, by depth, which falls into one of five distinct salinity classes. The

mapping classification scores estimate the theoretical accuracy of the predicted spatial salinity map, assuming the mapping contours correspond to the range interval cutoff levels. Finally, the spatial variation indices divide the total observed salinity variation within the survey area into two components; short scale (micro) variation, and field scale (macro) variation.

The first statement which appears after invoking option 4 is:

The current model is: Sx- Txx
Proceed with this model? (y/n):

After answering yes, the following statement will appear

Please choose the desired confidence level: (1) 90%, (2) 95%, (3) 99%:

followed by

The current range cutoff levels are 2.00, 4.00, 8.00, 16.00
Change the current range setting? (y/n):

At the confidence interval prompt, you can select either at 90%, 95%, or 99% confidence interval by entering the numbers 1, 2, or 3, respectively. At the range interval prompt, you can select a new set of cutoff levels by answering yes and entering the new levels at the appropriate prompts.

NOTE: The new range interval cutoff levels must be entered in increasing, sequential order. Additionally, the default range intervals are assumed to be in dS/m units.

The last question printed to the screen is

This program assumes that your instrument IEM) data was collected on a systematic grid across the entire survey zone. Is this correct? (y/n):

If you answer no, a warning statement will appear on the screen, notifying you that all the field diagnostic summary statistics are biased. After answering this last question, the program will perform the necessary report calculations and then display then to the screen. After displaying all four parts of the salinity diagnostic report, you will also have the option of printing out the results. At this point, the entire diagnostic report will also be saved to the hard disk.

5. Site Monitoring: Net Flux Testing

This option allows you to test for a change in the survey areas average Ln salinity level over time. To use this option, you must collect additional salinity samples at one or more monitoring/validation sites, and have this additional data summarized and stored in a separate ASCII text file (see section 5.4).

The first statement which appears after invoking option 5 is:

The current model is: Sx- Txx
Proceed with this model? (y/n):

After answering yes, the following statement will appear

Please enter the path/file name:

followed by

Please choose the desired confidence level: (1) 9O%, (2) 95%, (3) 99 %:

After entering the correct input path/file information and selecting the appropriate confidence level, the program will display the net flux calculations on the screen. This information will include the observed (sampled) and model predicted mean Ln salinity levels, by depth, along with the standard deviations of the predicted levels, t test scores and probability levels. The average net-flux at each depth will also be estimated, along with the appropriately specified confidence intervals (for these net flux estimates).

NOTE: The net flux in the salinity level is estimated as $100[e^{(O-P)}-1]$, where 0 represents the observed average Ln salinity level across all the monitoring sites, and P represents the average Ln predicted level.

After displaying the net flux test results, you will have the option of printing out these results. At this point, these test results will also be saved to the hard disk.

6. Exit to DOS

As previously mentioned, this option should be used to end the ***EMSMLR*** program. Note that it can be invoked at any time from the SMLR MODELING menu.

5.4 Input/Output File Description

There are two input ASCII text files accepted by the EMSMLR program; the calibration survey/salinity data file (which is required) and the data file containing the monitoring/validation soil salinity information (which is optional). Both files have specific text and column structure requirements, which will be described shortly.

There are two ways that you can create either of these input ASCII text files. The first way to create these files is by using the supplied utility program ***DATALOAD*** (this program is located in the c:\emsurvey\utility subdirectory). To use the ***DATALOAD*** program, you must have saved the ***EMCCRSD.OUT*** output text file (the output file produced by the ***EMCCRSD*** program, see Section 3.2). You also need to have your laboratory determined calibration and/or monitoring/validation salinity data either saved as ASCII text files, or in front of you when you initiate the ***DATALOAD*** program (so that this information can be typed in via the keyboard). To use the ***DATALOAD*** program, move into the c:\emsurvey\utility subdirectory, type dataload [↵] at the DOS prompt, and follow the instructions contained in the on-line help facility.

If you prefer, you can create either of the ***EMSMLR*** input files using your own favorite database management program. In order to do this, you must have access to the ***EMCCRSD.OUT*** text file, your laboratory analyzed soil salinity data, and a database program (or text editor) capable of producing a format specific text file by merging together multiple ASCII input files. You must also be familiar with the specific text and column structure requirements of each ***EMSMLR*** input file. Even if you never intend to create any input files in this manner, you should become familiar with the text and column structure requirements. You will need to know this information, should you ever need to modify an input file after it has been created by the ***DATALOAD*** program.

The first two lines of the input calibration survey/salinity data file must have the following structure in order to be successfully read by the ***EMSMLR*** program:

Line 1: survey area title (40 character or less)

Line 2: n1 n2 n3 n4 n5 (all integer values)
where
n1 = the total # of EM survey sites throughout the survey area,
n2 = the total # of soil salinity calibration sites throughout the survey area,
n3 = the total # of calibration sites with two salinity samples per depth (i.e., the # of sites where replicate salinity cores were acquired)

n4 = the # of salinity sample depths at each site,
n5 = the # of principal component scores in the input file.

Note that the acceptable bounds on nl through n5 are as follows:

43 ≤ nl ≤ 399
15 ≤ n2 ≤ 32
0 ≤ n3 ≤ n2, with the restriction that (n2+n3) ≤ 40
1 ≤ n4 ≤ 6, and
n5 = 2 or 3.

The abbreviated survey code attached to your input file will be generated from line 2 using the calibration, replication, and sample depth information (i.e., n2, n3, and n4). The survey code will always be expressed as [n2]-[n3]r/[n4]d. For example, if your input data consisted of 18 calibration sites with replication cores from 4 sites and 5 sampling depths acquired at each site, then the survey code would be written as 18-4r/5d.

The remaining lines in the input file must list the survey information in the following order; (1) the site ID, (2) the replication code, (3) the sample Ln salinity levels across all n4 sample depths, (4) the n5 principal component scores, and (5) the spatial x,y coordinates. The site ID column must contain distinct, integer value site identification codes. The replication code column must contain one of three integer values; 0, 1, or 2. It should be set to 0 if no salinity data were acquired at the site, 1 if salinity data were acquired at the site and these data come from a primary core, and 2 if the acquired salinity data come from a replicate core sample. The next n4 columns must contain either the natural log transformed levels of the sample salinity data (listed in increasing depth order) or periods to indicate that no data were acquired at the site. After this, the next n5 columns must contain the principal component scores, listed in increasing order. Finally, the last two columns must contain the spatial x and y coordinate data, respectively. Note that the *EMSMLR* program can accept EC, information at each site from one primary and one replicate core only. Note also that all calibration sites (i.e., sites associated with sample salinity data) must be listed at the beginning of the file, and that the salinity information associated with a replication core must be listed immediately after the primary salinity core information in the input file (on a site by site basis).

An abbreviated version of the tutorial input data set is shown in table 5.2. The full data set resides in the c:\emsurvey\data subdirectory, in a file entitled *WWD 1.DAT.*

A calibration survey/salinity input data file can have any valid DOS name and extension, and reside anywhere on you hard disk. However, we recommend that you

Table 5.2 Partial listing of the ASCII input text file *WWD1.DAT*.

```
  Westland Field 1
178   16    5    3    3
 14 1 2.13983 2.28707 2.37388  0.19527  2.35261 -0.36745 6.375 6.775
 18 1 0.44725 1.35196 2.04368 -0.58996  0.57199 -1.23556 6.375 4.575
 18 2 1.45115 1.85848 1.94762 -0.58996  0.57199 -1.23556 6.375 4.575
 25 1 0.54812 0.48551 1.08620 -0.89427  0.96710 -1.03467 6.375 0.725
 30 1 2.28054 2.53576 2.66445  1.46112  0.62648  1.12343 5.875 2.925
 61 1 1.78440 2.32601 2.48582 -0.30900 -0.44208  1.70459 4.875 5.675
 71 1 1.97269 2.43502 2.65169  0.94645 -0.98932  0.99645 4.375 4.025
 71 2 2.11541 2.58415 2.77128  0.94645 -0.98932  0.99645 4.375 4.025
 80 1 0.12751 1.77360 2.34861 -1.16429 -1.69552 -1.22889 3.875 1.825
 80 2 0.08618 1.98142 2.26768 -1.16429 -1.69552 -1.22889 3.875 1.825
 89 1 2.19467 2.55202 2.60992  1.00235  1.23840 -1.40197 3.875 6.775
104 1 1.45815 2.30298 2.44790 -1.03820 -0.88657  0.87646 2.875 0.725
114 1 1.80187 2.54866 2.69679  1.03919 -0.99689 -0.83981 2.875 6.225
124 1 1.96235 2.66166 2.67511 -0.15531 -2.07773 -0.63067 2.375 3.475
127 1 0.92980 1.96009 2.32082 -1.06985 -0.82795  0.30467 2.375 1.825
160 1 0.44340 0.86415 1.39946 -0.99402  1.43640  1.08764 0.875 2.925
164 1 1.60322 2.38223 2.45342  0.48703  0.10197 -1.99581 0.875 5.125
170 1 2.25518 2.69773 2.69273  1.43900 -0.11955 -0.33477 0.375 6.775
170 2 2.21964 2.77614 2.78062  1.43900 -0.11955 -0.33477 0.375 6.775
180 1 1.43198 1.14994 1.88964 -1.49107  0.32601  0.11692 0.375 1.275
180 2 0.25542 0.41078 1.12590 -1.49107  0.32601  0.11692 0.375 1.275
  1 0  .       .       .      -0.88096  0.95718 -2.45959 6.875 0.725
  2 0  .       .       .      -1.21476  0.74876 -0.84224 6.875 1.275
  3 0  .       .       .      -1.33616  1.37840 -1.36585 6.875 1.825
  4 0  .       .       .      -0.59666 -0.06650 -1.89432 6.875 2.375
  5 0  .       .       .      -1.25330  0.77502 -0.12878 6.875 2.925
  6 0  .       .       .      -1.18066  0.91266 -0.89863 6.875 3.475
  7 0  .       .       .      -1.07546  0.48231  0.36584 6.875 4.025
  8 0  .       .       .      -0.77222  0.42652 -0.78172 6.875 4.575
  9 0  .       .       .      -0.52699  1.16352  1.24036 6.875 5.125
 10 0  .       .       .      -0.53795  2.68214  0.76743 6.875 5.675
                                    .
                                    .
                                    .
171 0  .       .       .       1.11078 -0.35516 -1.63649 0.375 6.225
172 0  .       .       .       0.79410 -1.07959 -1.29098 0.375 5.675
174 0  .       .       .      -0.84571 -0.45999 -0.69455 0.375 4.575
175 0  .       .       .      -1.02428 -0.19180  2.17142 0.375 4.025
176 0  .       .       .      -1.33019  0.41186 -0.08571 0.375 3.475
177 0  .       .       .      -1.34593  0.50541  0.36075 0.375 2.925
178 0  .       .       .      -1.20500 -0.44132 -0.99610 0.375 2.375
179 0  .       .       .      -1.31753 -0.15898 -0.33702 0.375 1.825
181 0  .       .       .      -1.42338 -0.22541 -0.02352 0.375 0.725
```

store this file in the c:\emsurvey\data subdirectory, and that you give it a four character DOS name, followed by a .DAT extension. Additionally, we recommend that you use this four character file name as your four character survey code in the *EMSMLR* program. Note that you must always create a calibration input file before running the *EMSMLR* program.

WARNING: The *EMSMLR* program will produce spurious results and/or crash if the input calibration file structure is not properly specified.

NOTE: The *EMSMLR* program assumes that the calibration salinity data is measured in Ln(dS/m) units. If your sample data has been measured in some other units, you should convert your readings to dS/m units before using the DA *TALOAD* program (DA *TALOAD* will automatically apply a Ln transform to the user specified soil salinity data).

As already mentioned, the input file containing the average Ln salinity levels associated with one or more monitoring sites can also be created using either the *DATALOAD* utility program or a text editor. If you create this file yourself, it should have the following file structure:

Line 1: $m_1, m_2 \ldots m_j$ j = n4
where
m_1 = the average Ln salinity level in the first sampling depth for all v monitoring samples,
m_2 = the average Ln salinity level in the second sampling depth for all v monitoring samples,

m_j = the average Ln salinity level in the jth sampling depth for all v monitoring samples.

Line 2: n6 (integer value)
where
n6 = the total number of monitoring samples.

Lines 3 through n6+2: the individual monitoring site id numbers (one per line).

Note that the average Ln salinity levels at all depths must be based on the same number of monitoring sites. If the average salinity level can not be computed for a specific depth (because no samples at this depth were obtained at any of the monitoring sites), then a period should be used to indicate this missing data on line 1. Also, note that the *EMSMLR* program will not accept any monitoring site id

numbers which coincide with calibration site id numbers. In other words, you cannot specify an individual site to be both a calibration and monitoring site in the same field.

As with the calibration input file, we suggest that you store this file in the c:\emsurvey\data subdirectory. We also suggest that you assign the file name to be the same four character name used for the calibration file, and that you specify the DOS extension to be .NEW. The tutorial data file containing the validation salinity data is shown in table 5.3. Note that this file resides in the c:\emsurvey\data subdirectory, as is entitled WWD 7. *NEW.*

The *EMSMLR* **program has to capability to generate up to 20 output text files which list the various model estimation, validation, and prediction results. A structured file naming system is used by the** *EMSMLR* **program in order to help the user keep track of the various output files. All output files receive eight character names, where the first four characters are determined by the user specified four character survey code and the last four characters are determined by the program. Additionally, only two types of three character DOS extension codes can be assigned to these files. The first possible DOS extension code is .MIS, which represents an abbreviation for "model identification score". This extension code is placed after the four files created in the model identification subroutine. The second possible DOS extension code is .M&&, where the && are wildcard symbols which stand for the appropriate two digit model code used in the model estimation and validation subroutine. For example, the two digit model code for the S3i-Tyl regression model is 52; hence, all estimation and validation output text files associated with this model will receive a three character DOS extension of .M52.**

Table 5.3 Listing of ASCII input text file WWD 7. *NEW.*

```
1.29964  2.20656  2.49076
8
29
36
79
97
100
117
139
159
```

Table 5.4 lists the names, extensions, and contents of the 20 output text files which can be generated by the program. Table 5.4 also indicates when and how these files are created, and whether or not they are automatically written to the hard disk.

NOTE: **Because you may wish to fit many different models before making your final model selection, none of the output text files created in the model estimation and validation subroutines will be written to the hard disk unless you specifically request the EMSMLR program to do so. If you specifically request to save output files from more than one regression model, you will need to use the .M&& extensions to determine which output text files associate with each model.**

NOTE: **All output text files are written to the c:\emsurvey\phase2 subdirectory.**

NOTE: **During execution, the *EMSMLR* program will create a file called W. *MTX.* This file contains the proximity matrix for the Moran spatial residual test, and does not need to be printed out.**

5.5 Tutorial Example

You should now try running *EMSMLR* using the supplied tutorial survey/sample data set, *WWDI.DAT.* Initiate the *EMSMLR* program, move to the SMLR MODELING menu and select option 1. Type "c:\emsurvey\data\wwd1 .dat" at the path/file prompt and "wwd1" at the survey code prompt. You should then see the following information scroll across the screen:

File name:	***c:\emsurvey\data\wwdl.dat***
Title:	***Westland Field 7***
Survey Code:	***wwd1***
Abbreviation of survey type:	***7 6-5r/3d***
Number of survey sites:	***778***
Number of calibration sites:	***76***
Number of sites with two samples per depth:	***5***
Number of sample depths:	***3***
Number of principal component scores:	***3***

After confirming that the correct input file has been read into the program, answer no at the verify input data prompt and observe the summary input data statistics. Note that the sample mean Ln salinity levels increase and the sample variances decrease with depth, and that the salinity data from depths 2 and 3 appear

Table 5.4 EMSMLR output text file characteristics (four character survey code is displayed as xxxx, 2 digit model code is displayed as &&).

8 Character File Code	DOS Extension	Contents	Created in Main Menu Option #	Automatically written to Hard Disk
xxxxAPRS	.MIS	absolute PRESS statistics	2	yes
xxxxRPRS	.MIS	relative PRESS statistics	2	yes
xxxxAPVE	.MIS	absolute APUE statistics	2	yes
XXXXRPVE	.MIS	relative APUE statistics	2	yes
xxxxECOP[1]	.M&&	observed and predicted log salinity levels at all calibration sites	3	yes
xxxxMSMS	.M&&	summary model statistics	3	no
xxxxNOVA	.M&&	ANOUA tables	3	no
xxxxPARA	.M&&	model parameter estimates	3	no
xxxxUNIV	.M&&	studentized residual summary statistics and stem-leaf plots	3	no
xxxxRESI	.M&&	studentized residuals and hat leverage scores	3	no
xxxxCORR	.M&&	residual correlation matrix	3	no
xxxxMSAT	.M&&	Moran residual spatial auto-correlation test statistics	3	no
xxxxSPDC	.M&&	summary prediction statistics	3	no
XXXXUSPD	.M&&	prediction statistics: non-sampled sites	3	no
XXXXUSPV	.M&&	prediction variance statistics: non-sampled sites	3	no
xxxxSDRP	.M&&	salinity diagnostic report	4	yes
xxxxFFIT[1]	.M&&	predicted log salinity levels for all survey sites	4	yes
xxxxTPRB[1,2]	.M&&	prediction variance factors for all survey sites	4	yes
xxxxHEAD[1]	.M&&	header information file	4	yes
XXXXFLUX	.M&&	net flux report / monitoring test results	5	yes

Note[1]: the following output text files can not be printed from within the EMSMLR program: xxxxECOP.M&&, xxxxFFIT.M&&, xxxxTPRB.M&&, and xxxxHEAD.M&&.

Note[2]: the prediction variance factors in the xxxxTPRB.M&& file can be converted into theoretical prediction variance estimates by multiplying each factor with the appropriate regression model MSE estimate.

highly correlated (r > 0.97). The Ln salinity depth correlation plots can be viewed by answering yes at the bivariate plot prompt. Also answer yes at the 1st PC plot prompt, and note that the relationship between the 1st principal component and the Ln salinity data appears to be nonlinear at each depth.

Upon returning to the main menu, select option 2. The program will inform you that there are 50 models to analyze; type the [↵] key to begin the analysis. (The entire analysis may take a few minutes to complete, if you are running the *EMSMLR* program on a 386/16 platform.) When the analysis is finished, a summary listing of the ten best models in each score category will appear on the screen; note that the S3i-Ty1 model is ranked 1st in all four score categories. Before returning to the main menu system, request a printout of the top 10 results.

Now choose option 3, and enter 52 at the two digit model code prompt. Note that the program asks you to confirm your model code choice by printing to the screen the abbreviated model parameter combination; S3i-Ty1. After answering yes, *EMSMLR* will estimate this model and send you to the ESTIMATION / VALIDATION menu. Selecting option 1 from this menu will send you to the MODEL ESTIMATION submenu, where you can now view the various model estimation results. Sequentially select options 1, 2, and 3 to display the model summary statistics, ANOVA tables, and parameter estimates. Note that the model R^2 values increase and the MSE estimates decrease with depth. In the ANOVA tables, note that all the model F-test statistics are highly significant (at or below the 0.001 level) and that the LOF test statistics are all non-significant. Finally, note that the majority of the depth 1 model parameter estimates do not appear to be significantly different from zero (as determined by the t-test significance levels). However, nearly all the parameter estimates within the depth 2 and depth 3 models appear to be statistically significant.

After returning to the ESTIMATION / VALIDATION menu, select option 2 to move to the RESIDUAL DIAGNOSTICS submenu and use the various options from within this submenu to view the residual diagnostics. Note that the univariate statistics, influence diagnostics, and residual normality plots confirm that the model residuals at each depth are approximately normally distributed and contain no outliers. Additionally, the residual/location and residual/prediction plots are devoid of any meaningful pattern, and the Moran residual test scores are all non-significant. Now leave the RESIDUAL DIAGNOSTICS submenu and move into the PREDICTION DIAGNOSTICS submenu. Use options 4 and 5 to view the observed versus predicted Ln and back transformed salinity levels. Note that there appears to be some degree of nonlinearity present in the observed/predicted relationship for the first depth. However, the observed/predicted relationships in the second and third depths appear linear and highly correlated.

At this point, you should be prepared to make a decision concerning the adequacy of the S3i-Ty1 models. Aside from the nonlinearity apparent in the low end

predicted Ln salinity data within the first sample depth, the fitted regression models appear to describe the sample data quite well. Note also that all the general regression modeling assumptions are satisfied. Neither the LOF or Moran tests were statistically significant, suggesting that no meaningful spatial autocorrelation could be detected in the residuals. Additionally, all three residual data sets appeared normally distributed with constant variance.

Based on these results, we choose to use the S3i-Ty1 model for predicting the spatial salinity distributions within the survey area at the 0.0-0.3, 0.3-0.6, and 0.6-0.9 meter sample depths (Lesch et. al., 1995). Hence, at least for purposes of this tutorial example, assume that the S3i-Ty1 model is adequate. Move to the FILE PRINT submenu and use options 1, 2, and 3 to print out the estimation and validation results. After this, select option 4 and save all of these results to the hard disk.

NOTE: You may have noticed that the intercept and y parameter estimates appear different from the estimates shown in Lesch et. al., 1995a,b. This is because the physical x,y survey site locations were expressed on a different scale during the original analysis.

NOTE: With regards to the final assessment of model adequacy, you will rarely ever be fortunate enough to find a model which performs exceptionally well across all sampling depths. For example, the S3i-Ty1 model describes the salinity data from the second and third sample depths quite well, but appears to contain too many parameters within the first sample depth (only the PC1 parameter appears statistically significant). Additionally, the MSE estimate in the first depth is 0.22, which is much higher than 0.10 (the approximate cutoff point for a “good” regression equation). However, the pure error estimate in the first depth is 0.24, which suggests that this near surface salinity variability is real (as opposed to being caused by a poorly fitted model). Additionally, any reduction in the number of model parameters will cause a significant increase in prediction bias in the second and third depths (try fitting the S1-TO model to this data set).

After exiting back up to the SMLR MODELING menu, select option 4 to produce the salinity diagnostic report. Make sure the active model is S3i-Ty1, select the 95% confidence level, and use the default range cutoff levels (2, 4, 8, and 16). After answering yes at the systematic grid prompt, EMSMLR will perform the diagnostic calculations and display them to the screen. Note that the predicted field median salinity levels increase with depth, as does the percent area of the field falling into the higher range interval estimates. The mapping accuracy scores suggest that the salinity maps at the deeper depths will be more accurate, and the spatial variation indices suggest that the first depth is the most locally variable. (The short scale variation accounts for 34% of the total variability in depth 1, as opposed to only

about 14% in depths 2 and 3.) After viewing these results, answer yes at the print results prompt; your printed output should look similar to the text output shown in table 5.5.

Now return to the SMLR MODELING menu and select option 5. The program will again ask you to confirm that the active model is correct; make sure that the model is still S3i-Tyl . Type c:\emsurvey\data\wwd1 .new at the path/file prompt, choose the 95% confidence level, and observe the net flux calculations displayed to the screen. Note that the average Ln salinity levels observed across the eight validation sites are not statistically different from the model predicted average levels at each depth. Once again, answer yes at the print results prompt; your printed output should match the text output shown in table 5.6.

Upon returning to the SMLR MODELING menu, you can now use option 6 to end the EMSMLR program. At this point, you may wish to view the output text files residing in the c:\emsurvey\phase2 subdirectory. You should find that all the files listed in table 5.4 now exist in this subdirectory (with either a .MIS or .M52 extension). Note that the *WD1FFIT.M52* file will be used as the input file for the *SALTMAP* program discussed in Section 6.

Table 5.5 Salinity diagnostic report produced by the *EMSMLR* program (using a S3i-Ty1 model and the *WWD1.DAT* input file).

Survey Type: [16-5r/3d] Survey Code: [wwd1]
Westland Field 1

Active Model: S3iTy1

I. POINT ESTIMATES

Depth	Mean(LnECe)	Var(LnECe)	95% Confidence Interval
1	1.509448	0.010475	[1.291343, 1.727552]
2	2.039121	0.003934	[1.905456, 2.172786]
3	2.298280	0.001996	[2.203073, 2.393487]

Depth	Median ECe	95% Confidence Interval
1	4.52	[3.64, 5.63]
2	7.68	[6.72, 8.78]
3	9.96	[9.05, 10.95]

II. RANGE INTERVAL ESTIMATES

Depth	Range 1	Range 2	Range 3	Range 4	Range 5
1	18.67%	24.45%	29.91%	20.52%	6.44%
2	7.06%	14.56%	24.10%	37.14%	17.15%
3	0.31%	7.75%	23.16%	49.23%	19.57%

Ranges: (1) [0.0,2.0] (2) [2.0,4.0] (3) [4.0,8.0]
(4) [8.0,16.0] (5) [> 16.0]

III. MAPPING CLASSIFICATION ACCURACY

Depth	CA
1	47.00%
2	62.96%
3	72.35%

IV. SPATIAL VARIATION INDICES

Depth	% Micro	% Macro	Micro CV	Macro CV
1	34.21%	65.79%	49.57%	72.51%
2	13.49%	86.51%	29.33%	83.51%
3	13.67%	86.33%	20.68%	55.01%

Table 5.6 Net flux calculations produced by the *EMSMLR* program (using a S3i-Ty1 model and the WWD 7. *NEW* input file).

```
Survey Type: [16-5r/3d]     Survey Code: (wwdl]
Westland Field 1

Active Model: S3iTyl
Source File <c:\emsurvey\data\wwdl.new>
Number of Sites: 8
Depth Level:  3[123]

Monitoring  Sites:     29   36   79   97  100  117  139  159


I.  MEAN  MONITORING  LEVELS:  Test  Statistics

Depth    Observed    Predicted    Std.Dev.    Observed  t    Prob  >  |t|

 1       1.29964     1.55830      0.20166     -1.28265       0.21908
 2       2.20656     2.14172      0.12359      0.52462       0.60752
 3       2.49076     2.36971      0.08803      1.37506       0.18930


II. PERCENT FLUX IN FIELD AVERAGE ECe

Depth      Net  Flux      95%  Confidence  Interval           Effect

 1        -22.792%        [ -49.762%,     18.659%]       non-significant
 2          6.699%        [ -18.007%,     38.847%]       non-significant
 3         12.868%        [  -6.438%,     36.157%]       non-significant
```

6.0 SALTMAP PROGRAM DOCUMENTATION

6.1 Program Description

SALTMAP is designed to display the salinity prediction map(s) for your survey area, and produce high resolution output graphics which can be printed on a PCL3 compatible printer. ***SALTMAP*** can read as input either a ***xxxxFFIT. M&&*** output file created by the ***EMSMLR*** program or an ASCII test file created by the user.

SALTMAP uses a modified version of an inverse distance squared spatial estimation routine to produce high quality raster maps of the predicted salinity distributions across your survey area. You can display and print either individual, depth specific salinity maps, or composite maps (showing up to four depths simultaneously). You can also interactively change both the number and magnitude of salinity contour levels and the map boundaries from within the program.

6.2 Input/Output File Description

SAL TMAP has been designed to automatically accept all ***xxxxFFIT. M&&*** files generated by the ***EMSMLR*** program. Do not modify these files in any manner if you wish to import them into the ***SALTMAP*** program.

SALTMAP will also accept an ASCII text file containing observed and/or predicted Ln salinity data, provided the file has the following column structure:

[I]: site ID [2]: x [3]: y [4 - 91: $LnEC_1$ $LnEC_2$ . . . $LnEC_6$

The program will accept anywhere from one to six columns of Ln salinity data, hence columns 5 through 9 are optional.

WARNING: When reading in a user specified ASCII text file, the site ID column should contain integer values only, and all columns of salinity data must be expressed in natural log units.

6.3 Program Operation

To initiate the ***SALTMAP*** program, move into the c:\emsurvey\phase2 subdirectory and type **saltmap [↵].** When the program appears, the screen will be separated into two sections; a drawing section and a program command section. All program options must be executed from the command section, using either the mouse

or keystroke (hot-key) commands.

After invoking *SALTMAP, you* will see two main menu buttons within the program command section of the screen; a File button and an Options button. You can display either of these menus by using the mouse to click on the appropriate button or by typing the underlined letter (hot-key) in the menu button title. You have two options to choose from within the File menu; (1) you can read in an input file using the Input File button, or (2) you can print a displayed map using the Print button. You can choose from among five options from within the Options menu; (1) you can display a single, depth specific salinity map at maximum resolution using the Single button, (2) you can display a composition of multiple salinity maps at minimum resolution using the Composite button, (3) you can change the map title and/or sample depth labels using the Titles button, *(4) you* can change both the number and magnitude of the salinity contour levels using the Legend button, and (5) you can change the default map boundaries using the Boundary button. Note that you can quit the *SALTMAP* program using either the Quit button or the Esc key (provided that neither the File or Options menus are currently activated).

Each command section option is shown (and discussed) below:

1.	**FILE BUTTON**	**2.**	**OPTIONS BUT-TON**
1 .a	**Input File**	**2.a**	**Single**
1.b	**Print**	**2.b**	**Composite**
		2.c	**Tiltles**
		2.d	**Legend**
		2.e	**Boundary**

1.a Input File:

You should use this option to read in your input file. You must specify the path and filename, whether or not the input file was generated by the *EMSMLR* program, and the number of sample depths. Your input file can contain up to six sample depths. However, if your input file contains more than four depths you will have to choose four specific depths to work with. *(SALTMAP* can not process more than four depths of salinity data at any one time.)

WARNING: **You must know beforehand how many depths there are in the input file. If an erroneous number is entered into the *SALTMAP* program, then the interpolation process will become corrupted and the displayed maps will make no sense.**

NOTE: **If you are reading in a file residing within the c:\emsurvey\phase2 subdirectory, you do not have to specify the subdirectory path.**

1 .b Print:

You can use this option to print the displayed map, provided you have a PCL3 compatible printer (such as a HP-Laserjet II, Ill, IV, or HP-Deskjet printer).

2.a Single:

This option can be used to increase the interpolated resolution of a salinity map at a particular depth. After choosing this option, you can select which depth to enlarge by using the mouse or by typing 1, 2, 3, or 4 on the keyboard. A map of the salinity distribution at the requested depth will then be displayed to the screen at maximum resolution. After the map is displayed, you can switch to another depth by selecting this option again and typing 1, 2, 3, or 4. If you read in an input file with only one depth, SALTMAP will automatically display the salinity map at maximum resolution. Note that once this option is invoked, the screen will automatically be redrawn.

2. b Composite:

This option can be used to simultaneously display up to four salinity maps (representing different depth levels) at the same time. Each salinity map within the display area will be shown at minimum resolution. If you read in an input file with multiple depths, ***SALTMAP*** will simultaneously display all the initial maps at minimum resolution. Note that once this option is invoked, the screen will automatically be redrawn.

2.c Titles:

You can use this option to change the default map title and/or the labels associated with each sample depth. This option will also be automatically invoked immediately after reading in a new input file. After typing in your new title and/or labels, you can either redraw the screen immediately or postpone the redrawing process in order to perform other options.

NOTE: The default title will be set to the name of your input file, and the default legends will be set to 0.0-0.3 m, 0.3-0.6 m, 0.6-0.9 m, and ***0.9-I*** .2 m, respectively.

2.d Legend:

You can use this option to change the number and/or magnitude of the contour cutoff levels. The default number of contours is 5, and the default cutoff levels are 2, 4, 8, and 16 dS/m, respectively. You can vary the number of contours anywhere

from 2 to 5, and that you must specify one less cutoff level than the number of contours you request. For example, if you want to display your map using the four contour levels [0,5], 15,101, [10,20], and [> 201, request 4 levels and specify the bounds to be 5, 10, and 20.

After changing the number and/or magnitude of the contour levels, you can either redraw the screen immediately or postpone the redrawing process in order to perform other options.

2.e Boundary:

You should use this option to change the minimum and maximum x and y boundary coordinates of the map. Note that the default values will always be set equal to the observed minimum and maximum x,y survey coordinates. After changing these coordinates, you can either redraw the screen immediately or postpone the redrawing process in order to perform other options.

NOTE: **SALTMAP will print a warning message if the new map boundaries are more than 25% larger than the default boundaries. However, you will still be allowed to create the map, regardless of the new boundary size.**

NOTE: **SALTMAP will automatically "true-size" each map up to a 4:1 ratio. If the ratio of the longest to shortest boundary exceeds 4: 1, then the map boundary will default to a 4:1 ratio. Note also that all maps produced by SALTMAP will have rectangular boundaries, regardless of the underlying shape of the survey area.**

6.4 Tutorial Example

In the following tutorial example, we have assumed that you are using a mouse. If you do not have access to a mouse, all of the following point and click operations should be replaced by the appropriate keystroke commands.

Initiate the SALTMAP program, click on the File button, and then click on the Input File option button. Type wwd1ffit.m52 at the filename prompt, y at the EMSMLR prompt, 3 [↵] at the sample depth prompt, and y at the verification prompt. Next, type Westland Field 1: MLR Model S3i-Ty1 [↵] at the title prompt, strike the [↵] key three times to accept the default sample depth labels, and type y at the verification prompt. At this point, the SALTMAP program will begin estimating the spatial salinity maps at the 0.0-0.3 m, 0.3-0.6 m, and 0.6-0.9 m depths and writing them to the display area of the screen.

When this process is completed, click on the Options button, and then click on the Legend option button. Type 4 **[↵]** at the contour level prompt, and 4 **[↵],** 8 **[↵],** and 16 **[↵]** at the cutoff level prompts. Next, type y at the verification prompt and **n** at the redraw screen prompt (the display area of the screen will now go blank). Click on the Options button again, and then click on the Boundary option button. Type 0.0 **[↵],** 7.0 **[↵],** 0.0 **[↵],** and 7.5 **[↵]** at the minimum & maximum x and y prompts, respectively. (In this example we are using a scaled coordinate system of 1 map unit = 100 meters for the WWD1 survey data. Hence, you just defined the east/west and north/south boundaries to be 700 and 750 meters, respectively.) Finally, type y at the verification prompt and y at the redraw screen prompt to once again estimate and display all the salinity maps simultaneously.

When the redraw process has finished, click on the File button. If you have a PCL3 compatible printer, you can now create a print out of the screen display area by clicking on the Print option button. Your printed output should look like Figure 6.1.

Next, click on the Options button again, and then click on the Single option button. Type 1 to draw a map of the salinity distribution at the 0.0-0.3 m sample depth at maximum resolution. After the entire map is displayed to the screen, click on the File button and then on the Print option button to produce a print out. Now repeat this process for the 0.3-0.6 m and 0.6-0.9 m sample depths. When you are finished, your printed output for the 0.0-0.3 m, 0.3-0.6 m, and 0.6-0.9 m depths should look like Figures 6.2, 6.3, and 6.4, respectively.

A clear understanding of the apparent spatial salinity distribution within the Westland field can be gained from Figure 6.1. Note that there appears to be a buildup of salinity along the northern end of the field, along with an apparent incursion of salinity moving from the northwest to southeast areas of the field. Note also that this incursion appears to become more pronounced with depth, suggesting that some sort of systematic change in one or more soil attributes are occurring within the profile as one moves from the southern to northern ends of the survey area. Within this field, it is quite possible that the salinity pattern was strongly influenced by soil textural changes; the saturation percent tended to increase not only with depth, but also spatially (in a pattern similar to the predicted salinity maps).

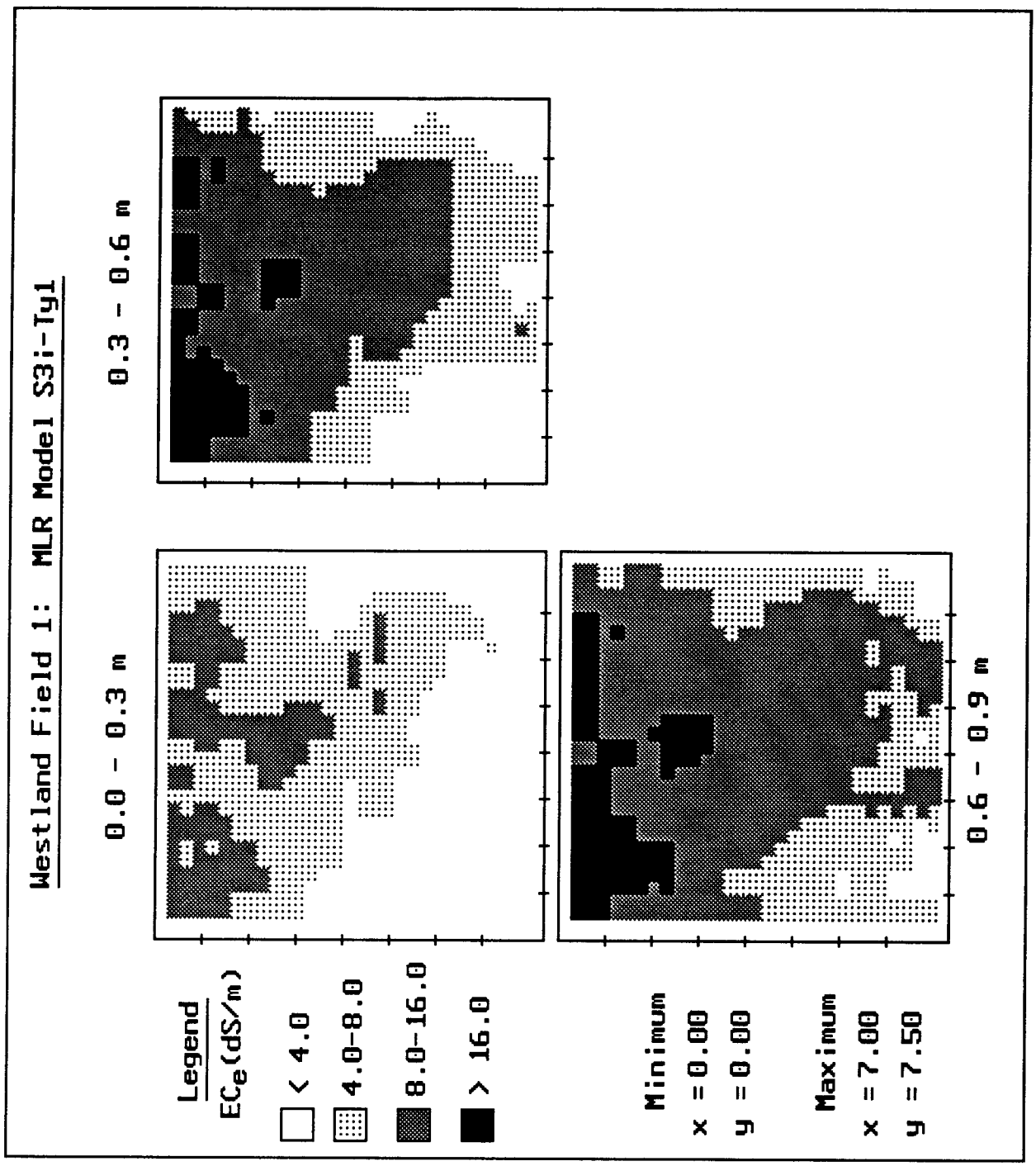


Figure 6.1 Composite printout of the Westland field salinity maps for the 0.0-0.3 m, 0.3-0.6 m, and 0.6-0.9 m depths; input file is *WWD1FFIT.M52*.

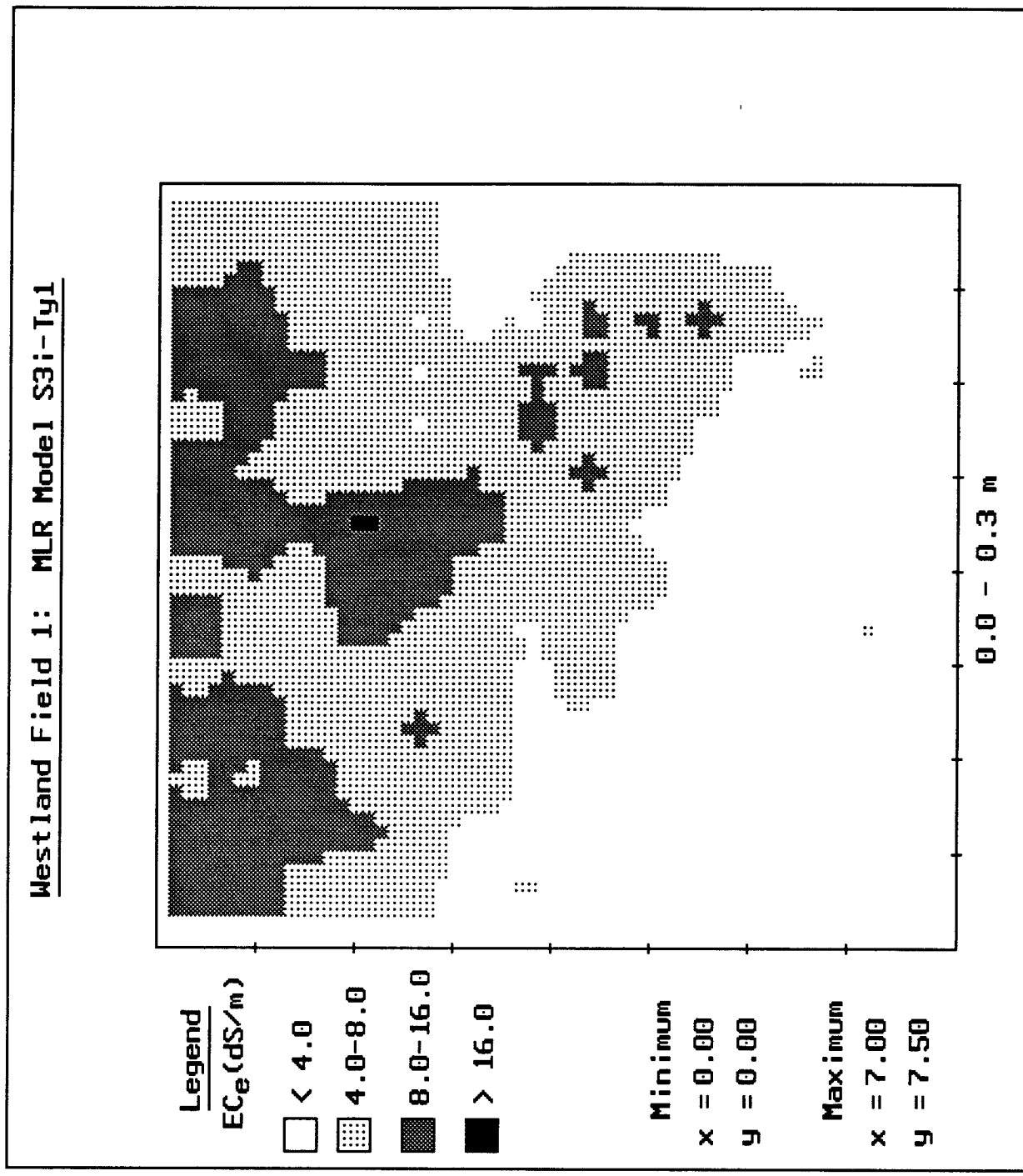


Figure 6.2 Individual printout of the Westland field salinity map for the 0.0-0.3 m depth; input file is *WWD1FFIT.M52*.

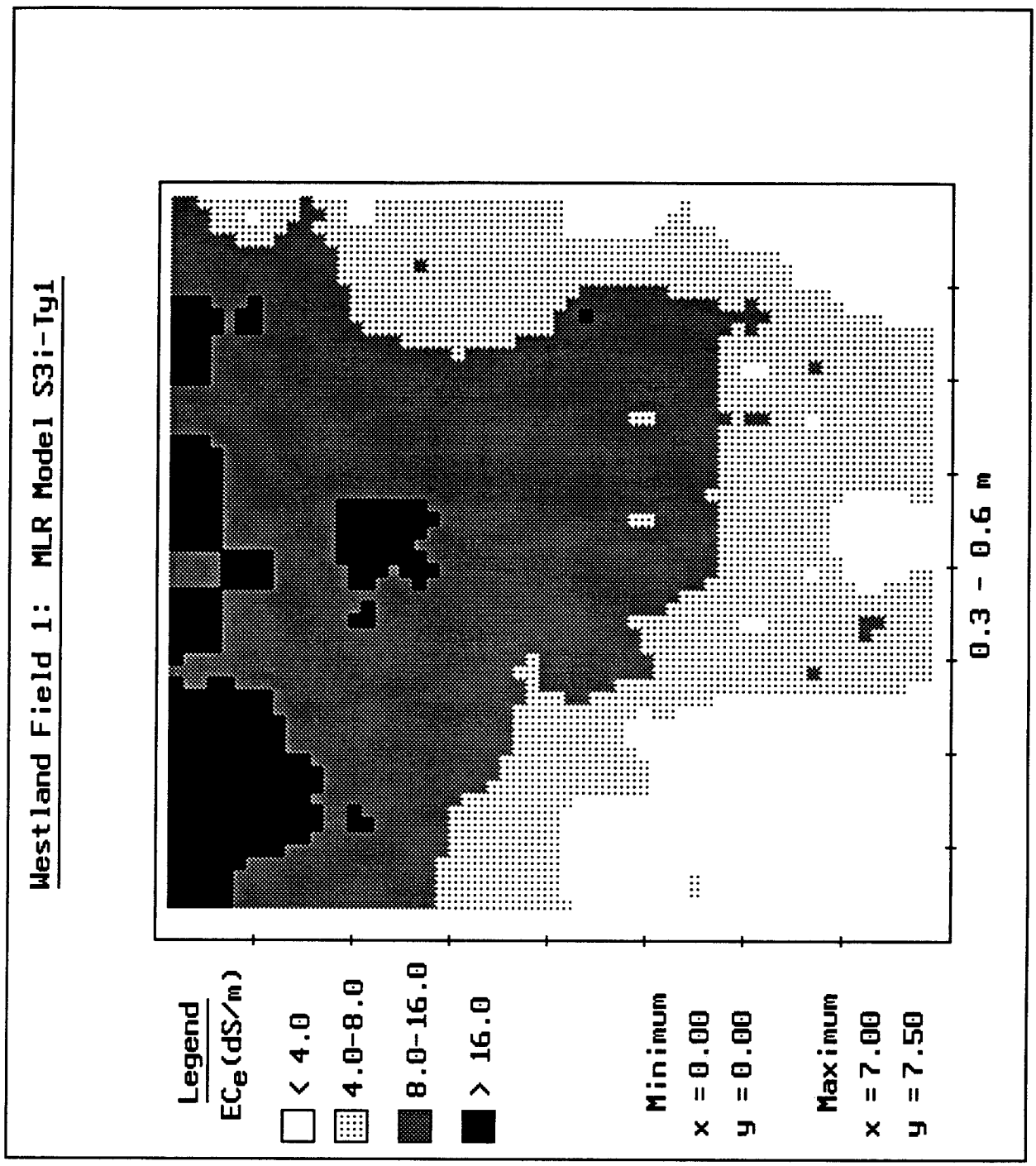


Figure 6.3 Individual printout of the Westland field salinity map for the 0.3-0.6 m depth; input file is *WWD1FFIT.M52*.

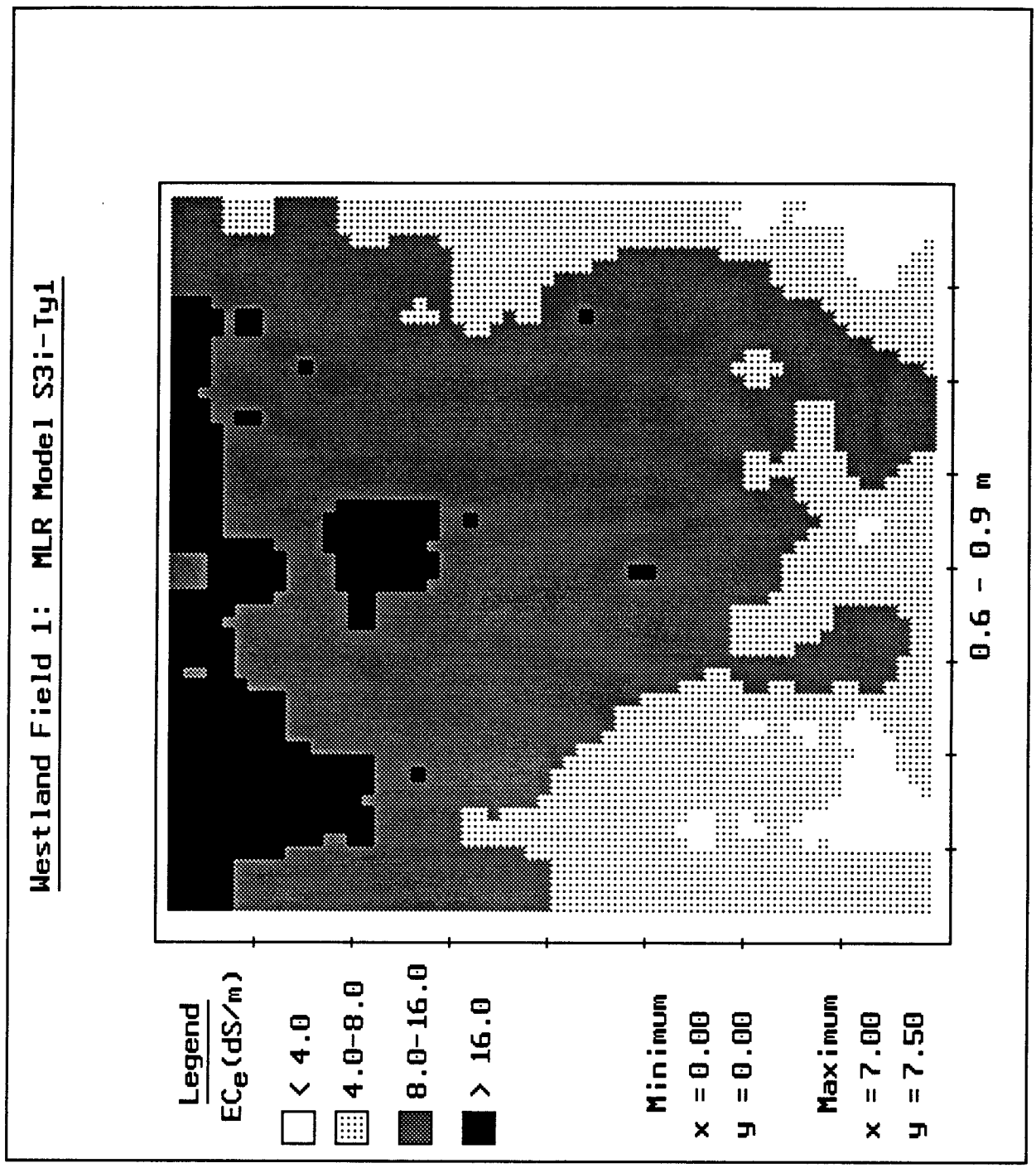


Figure 6.4 Individual printout of the Westland field salinity map for the 0.6-0.9 m depth; input file is *WWD1FFIT.M52*.

7.0 ADVANCED TUTORIAL EXAMPLES

7.1 Appropriate Model Selection/Validation Methodology

In Section 1.1 we pointed out some of the advantages that spatial regression models had over other types of stochastic/dynamic modeling approaches. One of the most important advantages not mentioned in Section 1.1 is the wealth of prediction and residual diagnostics available to the analyst when estimating a regression model. In any regression analysis, model validation is critical. The ultimate prediction accuracy will depend on the appropriateness of the fitted model, which in turn depends directly on the underlying validity of the modeling assumptions. There are numerous tests and diagnostics designed to assess both the residual assumptions and the degree of prediction accuracy and/or bias within a regression model; proper use of these diagnostic tools can significantly increase the likelihood of estimating a "good" (i.e., accurate and unbiased) model.

In Sections 5.3 and 5.5, a series of diagnostic tools designed for model validation and selection purposes were introduced and briefly discussed. This section is designed to give you additional experience in the efficient use of these diagnostic tools, by leading you through the analysis of three additional survey data sets using the *EMSMLR* program.

Before turning to the first tutorial example, a review of the different model diagnostic tools available within the *EMSMLR* program is in order. All *EMSMLR* model diagnostics can be classified into one of two categories; (1) residual diagnostics, and (2) assessment/prediction diagnostics. Residual diagnostics include all of the various residual tests and plotting techniques designed to detect residual assumption violations. On the other hand, the assessment/prediction diagnostics include all model goodness-of-fit statistics, parameter tests, and the various prediction assessment methods designed to appraise either the accuracy or reliability of the final predictions.

In order to choose between two or more competing regression equations, the following sequential validation/selection process should be employed. The residual diagnostics should be examined first, and any models displaying obvious residual assumption violations should be discarded. If more than one model passes all of the critical residual tests, then the various assessment/prediction diagnostics should be used to make the final model selection. To facilitate this validation/selection process, you should compile a list of the more important diagnostic test results for each model. An example of a "model comparison worksheet" which includes most of the useful residual and assessment/prediction diagnostics is shown in table 7.1.

Table 7.1 Example format for a model comparison worksheet.

Depth level: _____ **meters**

Survey Code: _____

	MODEL:					
1.	**LOF F prob level:**					
2.	**MORAN score:**					
3.	**Linearity of Q-Q Plot:**					
4.	**Res Skew:**					
5.	**Balance:**					
6.	**Abs > 2.5:**					
7.	**R/P Plot:**					
8.	**R/X Plot:**					
9.	**R/Y Plot:**					
10.	**Para #:**					
11.	**R^2:**					
12.	**Adj R^2:**					
13.	**MSE:**					
14.	**Model F prob level:**					
15.	**Para |t| prob levels:**	**< .01: .01-.05: .05-. 10: > .10:**	**< .01: .01-.05: .05- 10: > .10:**	**< .01: .01-.05: .05- 10: > .10:**	**< .01: .01-.05: .05- 10: > .10:**	**< .01: .01-.05: .05-. 10: > .10:**
16.	**PRESS/n:**					
17.	**APVE:**					
18.	**Predicted min - max:**					
19.	**Prd Var 50%:**					
	75%:					
	100%:					
20.	**Linearity of Ln Plot:**					

There are references to 20 specific diagnostic items in the worksheet shown in table 7.1. The first nine items represent residual tests and/or diagnostics, the next six represent model goodness-of-fit statistics, and the last five represent prediction diagnostics. Each of these items are explained in detail below.

1. LOF F prob level:

This space is for the probability level associated with the lack-of-fit F test. If replicate sample data was acquired, the lack-of-fit F test results will be displayed in the ANOVA tables (option 2 within the MODEL ESTIMATION submenu). A significant F test implies model bias and short range residual autocorrelation, and represents a critical residual assumption violation.

2. Moran Score:

This space is for the normalized Moran score (option 4 within the RESIDUAL DIAGNOSTICS submenu). Note that the EMSMLR program prints a one-sided probability level with this score; however, a two-sided test may sometimes be more appropriate. (Significant negative spatial autocorrelation will occur if the residuals display a cyclic pattern across the field; such a residual pattern is usually indicative of a pronounced, repetitive oscillation in the soil texture.) For a two sided test, scores above 1.645, 1.960, and 2.575 indicate positive residual autocorrelation at the 0.10, 0.05, and 0.01 significance levels; likewise, scores which fall below -1.645, -1.960, and -2.575 indicate negative autocorrelation at the same significance levels. Significant Moran scores (above 1.96 or below -1.96) represent a critical residual assumption violation.

3. Linearity of Q-Q Plot:

The assessment of the residual normality assumption can be done using a number of diagnostic tools; however, the most intuitive tool is probably the normality quantile-quantile plot (option 7 within the RESIDUAL DIAGNOSTICS submenu). Note that a linear pattern in the Q-Q plot indicates normally distributed residuals, while a nonlinear and/or fractured pattern is indicative of non-normality. A qualitative grade of the residual linearity within the Q-Q plot (i.e., good, ok, fair, poor) should be written in this space. When a fair or poor grade is recorded, the reason should also be listed (e.g., curvature, outlier, etc.).

4. Residual Skewness:

This space is for recording the residual skewness estimate (option 1 within the RESIDUAL DIAGNOSTICS submenu). Theoretically, the skewness should be equal to 0. In small sample sizes ($n \approx 20$), a skewness estimate between -0.5 to 0.5 can be assumed to be reasonable. Estimates below -1 .O or above 1 .O usually indicate

negative or positive skewness, respectively.

5. Residual Balance:

This space is for recording the estimated residual balance, which is another way of gauging the symmetry of the residual distribution. The residual balance is defined as follows:

balance = r100 / |r0| if r100 > |r0|
= |r0| / r100 if r100 < |r0|

where r0 and r100 represent the 0 and 100 residual percentile values (option 1 within the RESIDUAL DIAGNOSTICS submenu). Skewness and/or outliers may be present if the balance exceeds 1.5.

6. Abs > 2.5:

If the absolute value of any residuals (negative or positive) exceed 2.5, answer yes in this space and write down the value, otherwise answer no. (Individual studentized residual values can be shown using option 2 within the RESIDUAL DIAGNOSTICS submenu.) A residual with an absolute value exceeding 2.5 is considered a marginal outlier; a residual exceeding 3.5 should be considered a critical outlier.

7. R/P Plot:

A plot of the residuals against the model predicted Ln salinity levels can be shown using option 5 within the RESIDUAL DIAGNOSTICS submenu. This space is for a qualitative description of the R/P scatter plot (i.e., random, nonrandom). Note that the R/P plot should appear random; any obvious pattern suggests the presence of prediction model bias.

8. & .9 R/X and R/Y Plots:

Plots of the residuals against the x and y sample location coordinates can be shown using option 6 within the RESIDUAL DIAGNOSTICS submenu. Spaces 8 and 9 are for qualitative descriptions of these R/X and R/Y scatter plots, which should also appear random. Any obvious linear or quadratic relationships between the residuals and the spatial location coordinates suggests that the model is missing one or more important trend surface parameters.

10. Para #:

The number of model parameters (not including the intercept) should be written

down in this space. All else being equal, a model containing only a few parameters is preferable to a model containing many parameters.

11. R^2:

This space is for recording the coefficient of determination (R^2), which represents the proportion of variation in the response data that is explained by the model.

12. Adj R^2:

This space is for recording the adjusted coefficient of determination, which represents an R^2 value adjusted for the total number of model parameters. A large difference between the R^2 and the adjusted R^2 values usually implies that the model is "over-fit" (i.e., it contains too many meaningless parameters).

13. MSE

This space is for the recording the model mean square error estimate. A small MSE estimate usually implies good prediction accuracy.

NOTE: Items 10, 11, 12, and 13 can be displayed using option 1 within the MODEL ESTIMATION submenu.

14. Model F prob level:

This space if for the probability level associated with the overall model F test (option 2 within the MODEL ESTIMATION submenu). A significant F test implies that at least one parameter within the model (besides the intercept) is different from 0.

15. Para |t| prob levels:

This space is for recording the number of model parameters which are significant at the following probability levels: < 0.01, 0.01 to 0.05, 0.05 to 0.10, and > 0.10 (option 3 within the MODEL ESTIMATION submenu). In a good model, the majority of model parameters should be significant at or below the 0.1 level.

16. PRESS/n:

This space is for recording the jack-knifed estimate of the MSE, which is simply the PRESS statistic divided by the calibration sample size (option 1 within the PREDICTION DIAGNOSTICS submenu). A good model should have a small PRESS/n estimate, and the difference between the MSE and PRESS/n estimates should also be small.

17. APVE:

This space is for recording the average prediction variance estimate; a theoretical estimate of the average prediction variance associated with a non-sampled survey site (option 1 within the PREDICTION DIAGNOSTICS submenu). A good model should produce a small APVE.

18. Predicted min - max:

The predicted 0 and 100 salinity percentile levels should be written in this space (option 2 within the PREDICTION DIAGNOSTICS submenu). A biased model will tend to produce minimum or maximum salinity predictions which are significantly smaller or larger than the predicted levels in an unbiased model.

19. Prd Var:

These spaces are for recording the 50%, 75%, and 100% percentile prediction variance estimates (option 3 within the PREDICTION DIAGNOSTICS submenu). Large variance estimates indicate unreliable predictions.

20. Linearity of Ln Plot:

Another assessment of prediction reliability can be done using the Ln prediction plot (option 4 within the PREDICTION DIAGNOSTICS submenu). A nonlinear pattern in this plot indicates possible prediction bias in the model. A qualitative grade of the prediction linearity within the Ln prediction plot (i.e., good, ok, fair, poor) should be written in this space.

When soil samples are acquired at *k* sampling depths, there will be *k* individual regression models to validate (and hence it will be necessary to fill out *k* worksheets). There will often be times when you find that the parameter combination within a model does not perform well across all the sample depths. In such a scenario, it is a good idea to choose a parameter combination which performs satisfactory across most sampling depths, as opposed to a parameter combination which performs very well in some depths and very poorly in others.

7.2 Analysis of the H2SA Survey Data

The survey/salinity data associated with the 1989 H2SA project resides in the c:\emsurvey\data subdirectory, and is contained in a file called H2SA.DAT. Initiate the EMSMLR program and use option 1 to read in this data set. You should find that the data contains 206 survey sites, 20 calibration sites sampled from one depth only

(0.0-0.3 m), and that no replication cores were acquired at any calibration sites. Note also that the sample Ln salinity data appear well correlated with the 1st principal component scores.

Now use the Model identification option to compute the APVE and PRESS scores for all 50 models. Request a printout of the top 10 results after they have been displayed to the screen. The absolute PRESS rankings for the five best models should be S3-Tx2y1 (0.10310), S3-Ty1 (0.104391, S3-Tx2y2 (0.10585), S3-Ty2 (0.106091, and S2-Tx2y1 (0.11531). Likewise, the absolute APVE rankings for the five best models should be S2-Tx2y1 (0.106471, S3-Tx2yl (0.10722), S1-Tx2yl (0.11144), S3-Ty1 (0.11204), and S3-Ty2 (0.11584). Note that there are four models with high (top five) rankings in each category: S3-Ty1, S3-Ty2, S2-Tx2y1, and S3-Tx2y1.

At this point you should use option 3 in the SMLR MODELING menu to estimate and validate each one of the four models mentioned above. For comparative purposes only, we also recommend that you estimate and validate the Sl-TO model. As you validate each model, keep track of the model diagnostic results using a worksheet like the one shown in table 7.1.

When you are finished, your results should look like the worksheet data displayed in table 7.2. A comparison of the residual diagnostic results in table 7.2 suggests that the S3-Tx2y1 and S2-Tx2y1 models are the least subject to any residual assumption violations. Note that the Moran score in the Sl-TO model is significant below a 0.05 level. Furthermore, there is obvious linear and/or quadratic drift apparent in both the R/X and R/Y plots; hence this model should be discarded. There is a fair amount of skewness apparent in both the S3-Ty2 and S3-Ty1 residuals (both Q-Q plots appear curvelinear, and both residual distributions have poor skewness and balance estimates). While this residual skewness does not absolutely disqualify either model, it does tend to discredit them. On the other hand, the residual diagnostics from both the S3-Tx2y1 and S2-Tx2y1 models appear reasonable. The only two noticeable differences between the S3-Tx2y1 and S2-Tx2y1 models are (1) the Moran scores suggest that the S3-Tx2y1 residuals may be slightly less negatively correlated, and (2) there is a small fracture in the S3-Tx2y1 residual Q-Q plot, suggesting a mild violation in the residual normality assumption.

The residual diagnostic results suggest that we discard the Sl-TO model entirely and give preferential consideration to the S3-Tx2y1 or S2-Tx2y1 models, provided the remaining statistics due not clearly favor either the S3-Ty2 or S3-Ty1 models. None of the information in the goodness-of-fit statistics or the prediction diagnostics suggest that the S3-Ty1 or S3-Ty2 models are superior, hence we will limit our final choice to either the S3-Tx2y1 or S2-Tx2y1 model. Note that the S3-Tx2y1 model has slightly higher R^2 and adjusted R^2 statistics, and lower MSE and PRESS/n estimates. On the other hand, the S2-Tx2y1 model has a higher percentage

Table 7.2 Detailed comparison of five different prediction models for HS2A survey data.

Depth level: 0.0-0.3 meters
Survey Code: 20-0r/1d (N=206)

MODEL:	S3-Tx2y1	S2-Tx2y1	S3-Ty2	S3-Ty1	S1-T0
LOF F prob level:	*no reps*	*no reps*	*no reps*	*no reps*	*no reps*
MORAN score:	-0.915	-1.398	-0.086	0.266	1.997
Linearity of Q-Q Plot:	*fair (fracture)*	*good*	*fair (curvature)*	*poor (curvature)*	*good*
Res Skew:	-0.520	0.202	-1.099	-1.100	-0.367
Balance:	1.067	1.113	1.942	2.270	1.152
Abs > 2.5:	no	no	no	no	no
R/P Plot:	*random*	*random*	*random*	*random*	*random*
R/X Plot:	*random*	*random*	*random*	*random*	*drift (quadratic)*
R/Y Plot:	*random*	*random*	*random*	*random*	*drift (linear)*
Para #:	6	5	5	4	1
R^2:	0.933	0.920	0.916	0.907	0.761
Adj R^2:	0.903	0.892	0.886	0.883	0.747
MSE:	0.0708	0.0787	0.0829	0.0852	0.1836
Model F prob level:	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Para |t| prob levels:	< .01: 1 .01-.05: 3 .05-.10: 0 > .10: 2	< .01: 4 .01-.05: 0 .05-.10: 0 > .10: 1	< .01: 2 .01-.05: 0 .05-.10: 0 > .10: 3	< .01: 2 .01-.05: 0 .05-.10: 1 > .10: 1	< .01: 1 .01-.05: 0 .05-.10: 0 > .10: 0
PRESS/n:	0.1031	0.1152	0.1061	0.1044	0.1927
APVE:	0.1072	0.1065	0.1158	0.1120	0.2064
Predicted min - max:	0.17, 49.8	0.14, 50.9	0.20, 42.1	0.25, 49.2	0.19, 37.1
Prd Var 50%:	0.0971	0.1020	0.1079	0.1064	0.1981
75%:	0.1140	0.1147	0.1257	0.1166	0.2099
100%:	0.2779	0.1696	0.2664	0.2172	0.3144
Linearity of Ln Plot:	*good*	*good*	*good*	*good*	*fair*

of individually significant parameter estimates and less high end prediction variability. The APVE are nearly identical, and there is good linearity in both Ln prediction plots. Taken together, these results do not suggest a clearly preferable model.

In practice, either of these two models could have been used to construct the final Ln salinity predictions. Personally, we prefer the S3-Tx2y1 model, primarily because it retains all three principal component scores in the regression equation. However, both models yield nearly identical field average Ln salinity and range interval estimates, and produce very similar prediction maps. If you have not already done so, you should verify this by producing a salinity diagnostic report using each model now (your results should match the estimates shown in table 7.3). Note that the "true values" displayed in table 7.3 are based on all N = 206 observed salinity samples.

Figures 7.1 and 7.2 display the predicted field salinity maps created by the *SALTMAP* program using the S3-Tx2y1 and S2-Tx2y1 models, respectively. (You can recreate these maps yourself by importing the *HS2AFFIT.M46* and *HS2AFFIT.M26* output text files into the *SALTMAP* program.) Like the prediction estimates shown in table 7.3, these two maps are nearly equivalent. The only noticeable difference in the prediction maps occurs in the southeast corner of the field (lower right hand corner of the printed map), and this difference is minor.

Table 7.3 Predicted HS2A field average Ln salinity and range interval estimates from the S3-Tx2y1 and S2-Tx2y1 models. True HS2A values also shown (N =206).

	MLR Model S3-Tx2y 1	MLR Model S2-Tx2y1	True Values
Ln salinity estimate:	0.957	0.967	1.017
95% CI:	10.832, 1.0821	10.837, 1.097]	
Range Interval Estimates			
0.0 - 2.0:	43.0%	42.7%	42.5%
2.0 - 4.0:	24.8%	23.6%	23.1%
4.0 - 8.0:	17.9%	19.0%	16.1%
8.0 - 16. 0:	9.2%	10.0%	13.4%
> 16.0:	5.1%	4.7%	4.9%

Figure 7.3 displays the observed salinity map, based on all N = 206 sample sites. This map was generated by the *SALTMAP* program using the *HS2A.LOG* input data file (located in the c:\emsurvey\data subdirectory). Note that both predicted maps appear quite similar to the true map. Along with table 7.3, these results confirm that a spatial regression model can produce highly reliable predictions, when the model is properly estimated.

Figure 7.4 displays the predicted field salinity map created by the *SALTMAP* program using the Sl-TO model. We have included this figure for comparative purposes only, to demonstrate what can happen when an improper model is used to make the final predictions. Recall that this model not only failed a number of residual assumptions, but it's goodness-of-fit and prediction diagnostic results also appeared inferior in comparison to the remaining models. The bias in this model is clearly evident in the prediction map. The salinity levels tend to be under predicted in the northern end of the field and over predicted in the southeast quadrant.

7.3 Analysis of the CK44 Survey Data

The survey/salinity data associated with the 1993 CK44 project resides in the c:\emsurvey\data subdirectory, and is contained in a file called CK44.DAT. Initiate the *EMSMLR* program and use option 1 to read in this data set. You should find that the data contains 139 survey sites, 16 calibration sites, and 4 replication sites. Additionally, note that samples were acquired at four depths at each site (0.0-0.3, 0.3-0.6, 0.6-0.9, and 0.9-l .2 m). Use the bivariate Ln salinity plots to display the sample salinity correlation plots (for various sampling depths); note that the only meaningful correlation seems to occur between adjacent sampling depths. Note also that the Ln salinity data appears well correlated with the 1st principal component score in the first two sample depths only.

Now use the Model Identification option to compute the APVE and PRESS scores for all 50 models. Note that the top two scores in each category are associated with either the S3-TO or S3i-TO models. These scores suggests that the S3-TO and S3i-TO models should be estimated and validated first. As before, use a worksheet like the one shown in table 7.1 to keep track of the model diagnostic results. (Note that you will need four worksheets; i.e., one for each sampling depth).

When you are finished, your results should look similar to the first two columns of worksheet data displayed in tables 7.3a - 7.3d. For both models, you should have noticed a number of potential problems revealed by the various residual diagnostics. These problems include (I) fractures, curvature, and/or outliers in the residual Q-Q plots, (2) residual drift in 0.0-0.3 m, 0.3-0.6 m, and 0.9-l .2 m R/X and/or R/Y plots, and (3) a significant Moran test score (0.9-l .2 m depth, S3i-TO model).

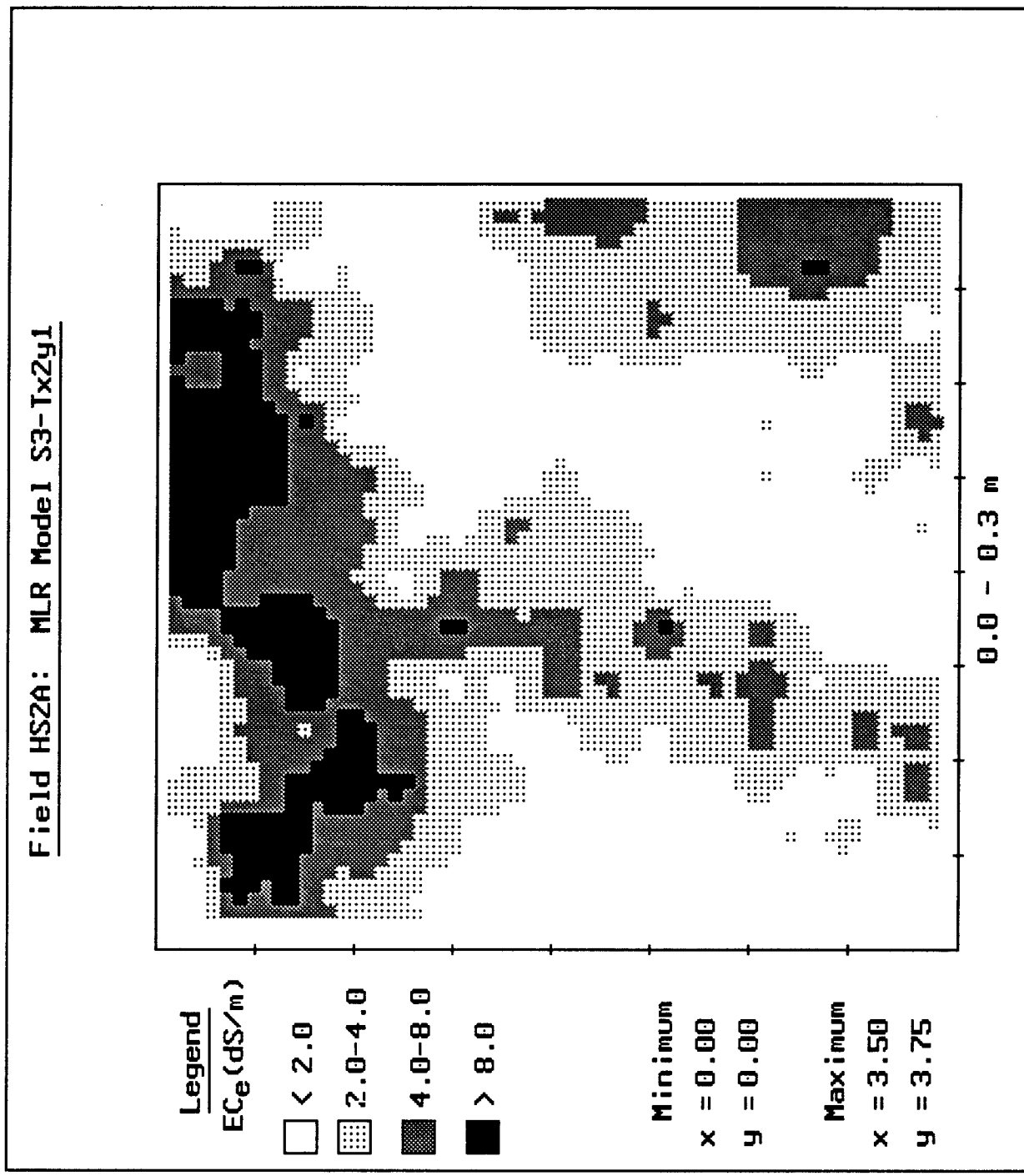


Figure 7.1 Printout of the predicted HS2A field salinity map for 0.0-0.3 m depth; input file is *HS2AFFIT.M46*.

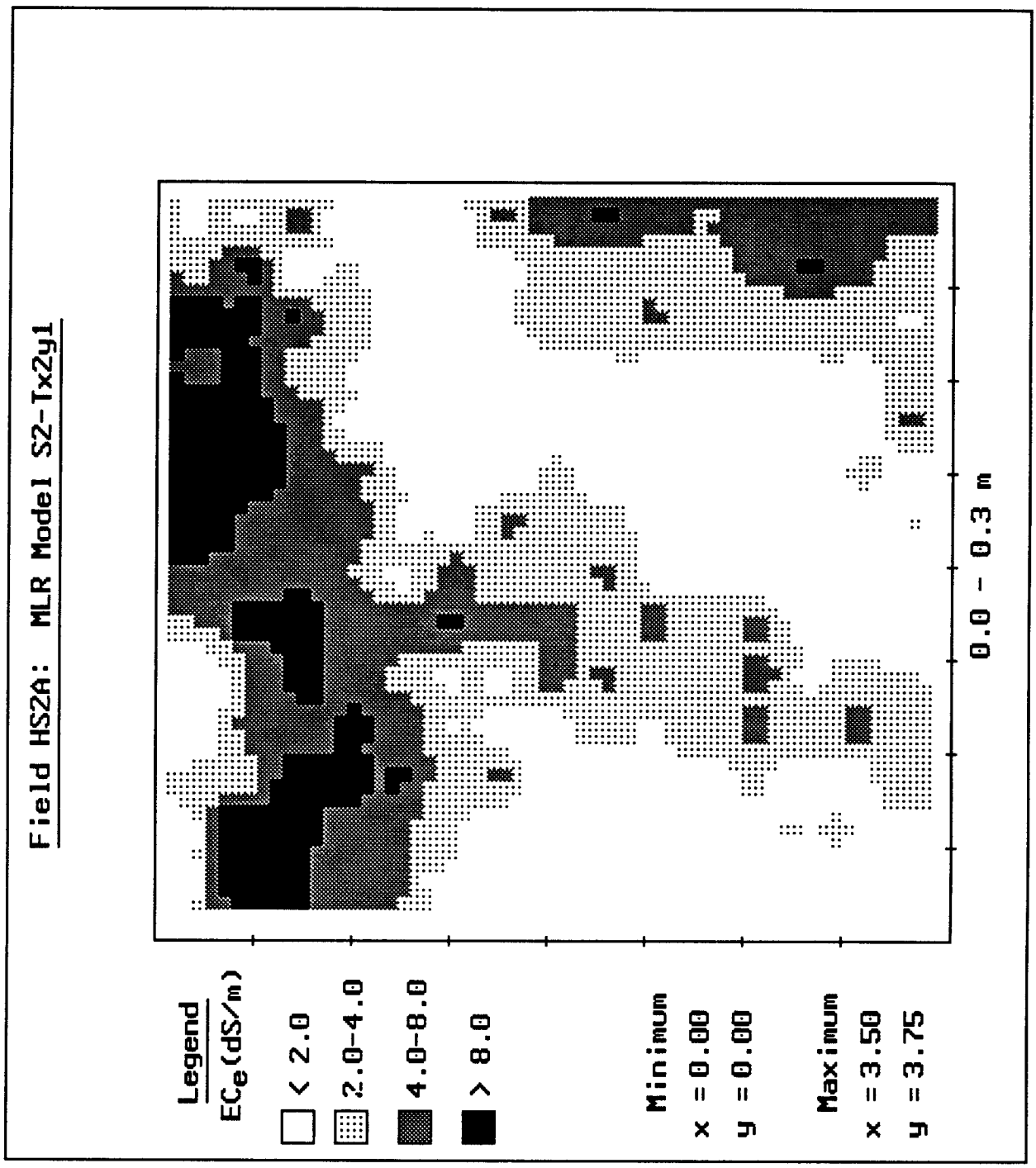


Figure 7.2 Printout of the predicted HS2A field salinity map for 0.0-0.3 m depth; input file is *HS2AFFIT.M26*.

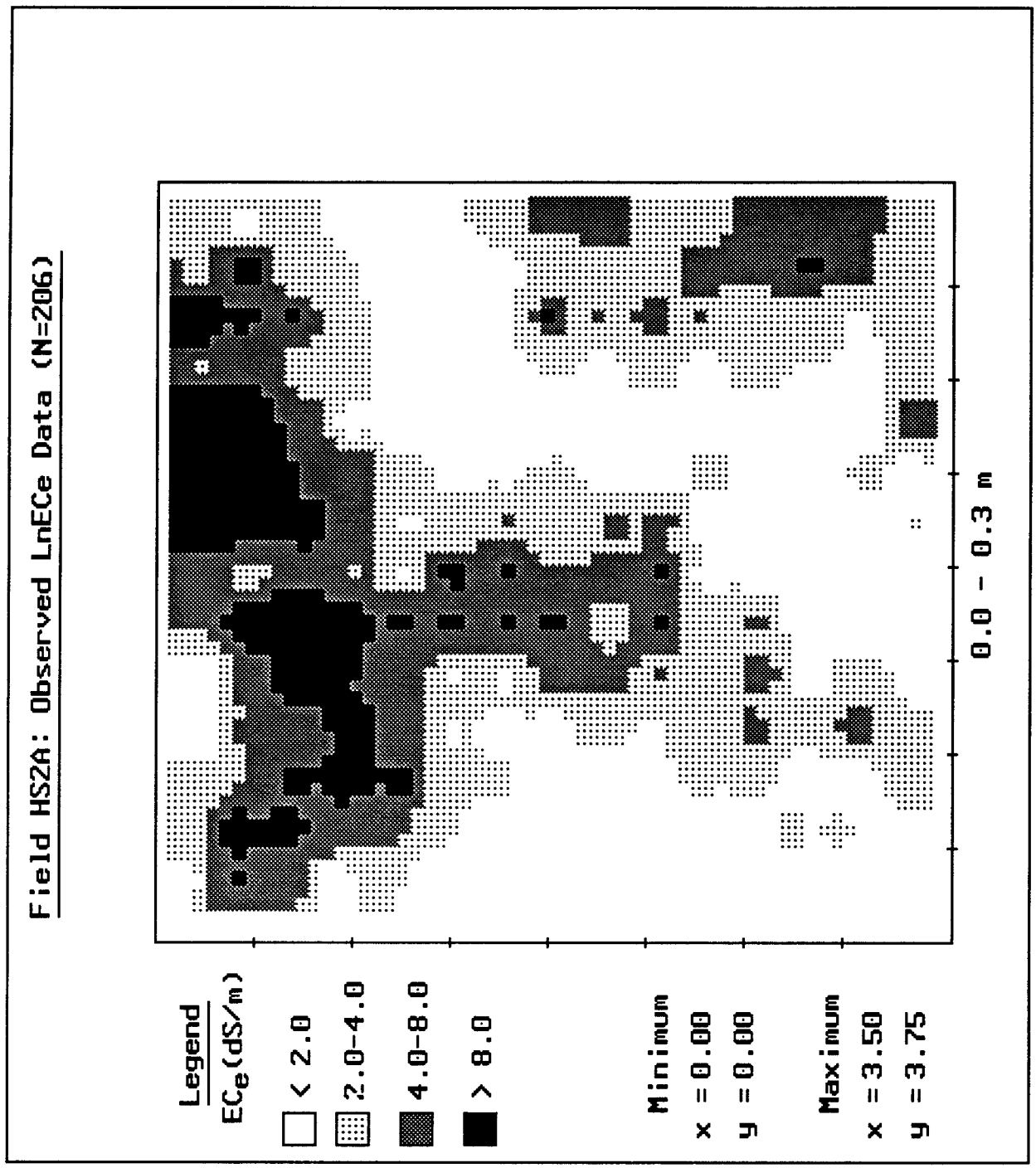


Figure 7.3 Printout of the observed HS2A field salinity map for 0.0-0.3 m depth; based on N = 206 sample sites (input file is *HS2A.LOG*).

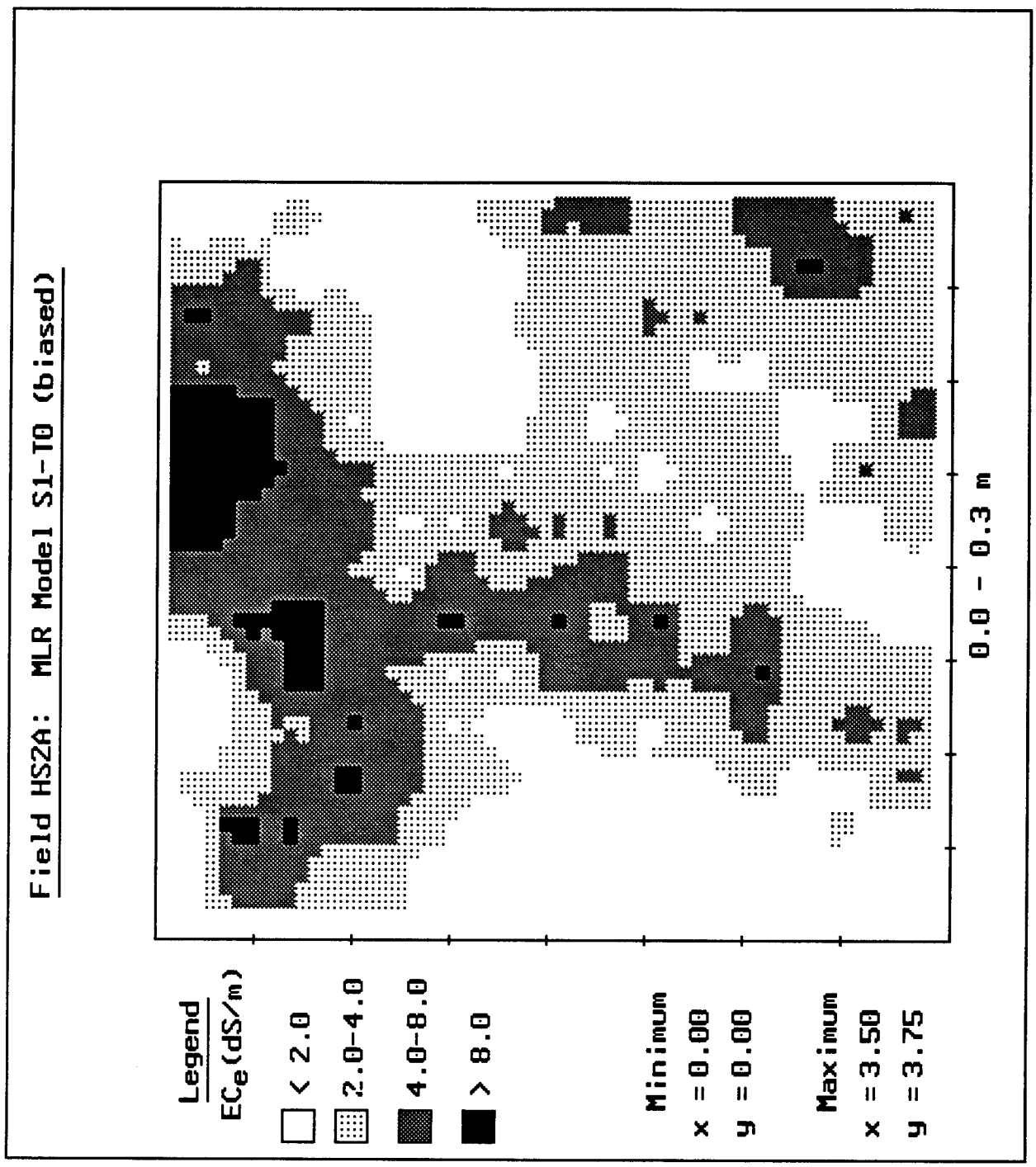


Figure 7.4 Printout of the predicted HS2A field salinity map for 0.0-0.3 m depth; input file is *HS2AFFIT.M10*. Note that the S1-T0 model used to predict the spatial salinity pattern is clearly biased.

The significant Moran test score and the residual drift apparent in the R/X and R/Y plots suggest that linear trend surface parameters should be included **in the fitted models.** Hence, you should now try fitting both the S3-Tx1y1 and S3i-Tx1y1 models to this survey data. Note that the residual diagnostics for the S3-Tx1y1 model are shown in the third columns of tables 7.3a-7.3d; these results suggest that the majority of problems apparent in the earlier diagnostics disappear once the trend surface parameters are incorporated into the S3-TO model.

Due to space considerations, we have not shown the results for the S3i-Tx1y1 model in tables 7.3a-7.3d. However, if you have not already done so, you should estimate and validate the S3i-Tx1y1 model now. Note that once the trend surface parameters are included into the equations, the interaction parameter estimates no longer appear to be highly significant (judging by the t-test scores). The prediction accuracy in the 0.0-0.3 m sample depth does appear to be better in this model (compared to the S3-Tx1y1 model); however, the residual Q-Q plot and skewness factor are considerably worse. Aside from the 0.0-0.3 m depth, the APVE, PRESS/n, and MSE estimates in both models are very similar. Most importantly, the final salinity estimates and predicted spatial maps are nearly equivalent.

Given these results, we will limit the remainder of this discussion to a comparison of the S3-TO and S3-Tx1y1 models only. Note that the assessment and prediction diagnostics displayed in columns one and three of tables 7.3a-7.3d suggest that the addition of the trend surface parameters has not resulted in an increase in the prediction accuracy at any of the sampling depths. On the other hand, these additional parameters do not seem to seriously degrade the predictive capabilities at any of the depths, and the S3-Tx1y1 model appears to be significantly less biased (judging by the residual diagnostics). Based on these results, we would recommend using the S3-Tx1y1 model for salinity prediction purposes.

The trend surface parameter effect on the final salinity predictions can be best judged by producing salinity diagnostic reports and spatial prediction maps using both models. Table 7.4 displays the field median and range interval estimates produced by both models; note that there are only small differences in these estimates. However, the predictions differences between these two models show up more clearly when you compare the spatial salinity maps shown in Figures 7.5 and 7.6. The dominant effect of the trend surface parameters in the S3-Tx1y1 model appears to be a north/south bias adjustment to the predicted salinity distributions in the 0.6-0.9 m and 0.9-l .2 m depths. This is primarily why we would recommend using the S3-Tx1y1 model (in comparison to the S3-TO model): it appears to correct for some subtle prediction bias in the lower sample depths without significantly degrading the prediction accuracy in the near surface depths.

For the record, you may be wondering why we chose to display the predicted salinity maps using non-standard contour cutoff points. If you have not already done

Table 7.3a Detailed comparison of three different prediction models for CK44 survey data at the 0.0-0.3 meter sampling depth.

Depth level: 0.0-0.3 meters
Survey Code: 16-4r/4d (N = 139)

MODEL:	S3-T0	S3i-T0	S3-Tx1y1
LOF F prob level:	nonsignificant	nonsignificant	nonsignificant
MORAN score:	-0.820	-0.092	-0.620
Linearity of Q-Q Plot:	*fair (fracture)*	*fair (curvature)*	*ok*
Res Skew:	-0.384	-0.798	-0.415
Balance:	1.160	1.612	1.228
Abs > 2.5:	no	no	no
R/P Plot:	*random*	*random*	*random*
R/X Plot:	*drift (linear/mild)*	*drift (linear/mild)*	*random*
R/Y Plot:	*random*	*random*	*random*
Para #:	3	4	5
R^2:	0.853	0.872	0.861
Adj R^2:	0.826	0.838	0.811
MSE:	0.0735	0.0683	0.0798
Model F prob level:	< 0.001	< 0.001	< 0.001
Para |t| prob levels:	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 1	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 2	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 3
PRESS/n:	0.0907	0.0854	0.1191
APVE:	0.0941	0.0932	0.1165
Predicted min - max:	2.64, 44.3	2.88, 50.4	2.76, 46.4
Prd Var 50%:	0.0891	0.0847	0.1081
75%:	0.0959	0.0984	0.1241
100%:	0.1715	0.1925	0.2435
Linearity of Ln Plot:	*good*	*good*	*good*

Table 7.3b Detailed comparison of three different prediction models for CK44 survey data at the 0.3-0.6 meter sampling depth.

Depth level: 0.3-0.6 meters
Survey Code: 16-4r/4d (N = 139)

MODEL:	S3-T0	S3i-T0	S3-Tx1y1
LOF F prob level:	nonsignificant	nonsignificant	nonsignificant
MORAN score:	1.017	0.892	1.086
Linearity of Q-Q Plot:	*fair (fracture & outlier)*	*fair (fracture & outlier)*	*fair (fracture)*
Res Skew:	-0.520	-0.864	-0.491
Balance:	1.278	1.572	1.316
Abs > 2.5:	yes (-2.50)	yes (-2.68)	yes (-2.63)
R/P Plot:	*random*	*random*	*random*
R/X Plot:	*linear drift or outlier*	*linear drift or outlier*	*random*
R/Y Plot:	*linear drift or outlier*	*linear drift or outlier*	*random*
Para #:	3	4	5
R^2:	0.827	0.838	0.865
Adj R^2:	0.794	0.794	0.817
MSE:	0.0341	0.0341	0.0303
Model F prob level:	< 0.001	< 0.001	< 0.001
Para |t| prob levels:	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 1	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 2	< .01: 1 .01-.05: 0 .05-.10: 1 > .10: 3
PRESS/n:	0.0415	0.0441	0.0454
APVE:	0.0436	0.0465	0.0442
Predicted min - max:	5.52, 29.0	4.97, 28.1	5.81, 24.8
Prd Var 50%:	0.0413	0.0422	0.0410
75%:	0.0445	0.0491	0.0471
100%:	0.0796	0.0960	0.0924
Linearity of Ln Plot:	*good*	*good*	*good*

Table 7.3c Detailed comparison of three different prediction models for CK44 survey data at the 0.6-0.9 meter sampling depth.

Depth level: 0.6-0.9 meters
Survey Code: 16-4r/4d (N = 139)

MODEL:	S3-T0	S3i-T0	S3-Tx1y1
LOF F prob level:	nonsignificant	nonsignificant	nonsignificant
MORAN score:	-1.348	-1.126	-1.252
Linearity of Q-Q Plot:	*fair (curvature)*	*fair (curvature)*	*fair (curvature)*
Res Skew:	0.062	0.237	-0.201
Balance:	1.363	1.384	1.125
Abs > 2.5:	no	no	no
R/P Plot:	*random*	*random*	*random*
R/X Plot:	*random*	*random*	*random*
R/Y Plot:	*random*	*random*	*random*
Para #:	3	4	5
R^2:	0.404	0.448	0.419
Adj R^2:	0.292	0.301	0.211
MSE:	0.0474	0.0468	0.0528
Model F prob level:	< 0.05	0.050	< 0.15
Para |t| prob levels:	< .01: 0 .01-.05: 1 .05-.10: 1 > .10: 1	< .01: 0 .01-.05: 0 .05-.10: 2 > .10: 2	< .01: 0 .01-.05: 0 .05-.10: 1 > .10: 4
PRESS/n:	0.0594	0.0601	0.0761
APVE:	0.0606	0.0638	0.0771
Predicted min - max:	5.82, 15.1	6.21, 16.8	5.68, 14.5
Prd Var 50%:	0.0574	0.0579	0.0715
75%:	0.0618	0.0673	0.0821
100%:	0.1105	0.1316	0.1611
Linearity of Ln Plot:	*fair (weak)*	*fair (weak)*	*fair (weak)*

Table 7.3d Detailed comparison of three different prediction models for CK44 survey data at the 0.9-1.2 meter sampling depth.

Depth level: 0.9-1.2 meters
Survey Code: 16-4r/4d (N = 139)

MODEL:	S3-T0	S3i-T0	S3-Tx1y1
LOF F prob level:	nonsignificant	nonsignificant	nonsignificant
MORAN score:	1.797	2.091	1.471
Linearity of Q-Q Plot:	*fair (fracture)*	*ok*	*ok*
Res Skew:	-0.288	-0.152	0.426
Balance:	1.086	1.131	1.443
Abs > 2.5:	no	no	no
R/P Plot:	*random*	*random*	*random*
R/X Plot:	*random*	*random*	*random*
R/Y Plot:	*drift (linear)*	*drift (linear)*	*random*
Para #:	3	4	5
R^2:	0.383	0.471	0.511
Adj R^2:	0.268	0.329	0.337
MSE:	0.0499	0.0457	0.0452
Model F prob level:	< 0.05	< 0.05	0.051
Para |t| prob levels:	< .01: 0 .01-.05: 0 .05-.10: 2 > .10: 0	< .01: 0 .01-.05: 1 .05-.10: 1 > .10: 2	< .01: 0 .01-.05: 0 .05-.10: 3 > .10: 2
PRESS/n:	0.0639	0.0605	0.0701
APVE:	0.0639	0.0624	0.0660
Predicted min - max:	4.08, 11.0	3.18, 10.1	4.11, 13.2
Prd Var 50%:	0.0605	0.0566	0.0613
Prd Var 75%:	0.0651	0.0658	0.0703
Prd Var 100%:	0.1165	0.1288	0.1379
Linearity of Ln Plot:	*fair (weak)*	*fair (weak)*	*fair (weak)*

so, initiate the *SALTMAP* program and read in either the CK44FFIT.M40 or CK44FFIT.M43 data files. Once the 0.6-0.9 m and 0.9-1.2 m prediction maps are displayed, you will see that nearly all the meaningful pattern is lost when the standard cutoff points (2,4,8,16) are used. When this happens, you will need to experiment with different cutoff points in order to create more visually interpretable maps (unless your chosen cutoff points are agronomically meaningful, in which case they shouldn't be changed regardless of what the predicted maps look like). If you do decide to use a new set of cutoff points, remember that you also need to produce a second salinity diagnostic report which reflects these new cutoff values, if you wish to compute a matching set of range interval estimates and classification accuracy scores.

Table 7.4 Predicted CK44 field average Ln salinity and range interval estimates from the S3-T0 and S3-Tx1y1 models.

	Model:	S3-T0		
Ln Salinity Estimate	0.0-0.3 m	0.3-0.6 m	0.6-0.9 m	0.9-1.2 m
w/ 95% CI	2.395 2.26, 2.53	2.498 2.41, 2.59	2.265 2.16, 2.37	1.869 1.76, 1.98
Range Interval Estimates	0.0-0.3 m	0.3-0.6 m	0.6-0.9 m	0.9-1.2 m
0.0 - 6.0:	19.7%	4.9%	5.5%	42.3%
6.0 - 9.0:	19.9%	18.8%	35.7%	43.6%
9.0 - 13.5:	21.3%	35.4%	45.8%	12.5%
> 13.5:	39.1%	40.9%	13.0%	1.6%

	Model:	S3-Tx1y1		
Ln Salinity Estimate	0.0-0.3 m	0.3-0.6 m	0.6-0.9 m	0.9-1.2 m
w/ 95% CI	2.401 2.26, 2.54	2.495 2.41, 2.58	2.261 2.14, 2.38	1.867 1.76, 1.97
Range Interval Estimates	0.0-0.3 m	0.3-0.6 m	0.6-0.9 m	0.9-1.2 m
0.0 - 6.0:	19.8%	4.9%	7.1%	43.9%
6.0 - 9.0:	19.2%	18.0%	35.1%	40.4%
9.0 - 13.5:	21.6%	37.2%	43.4%	13.2%
> 13.5:	39.4%	39.9%	14.4%	2.5%

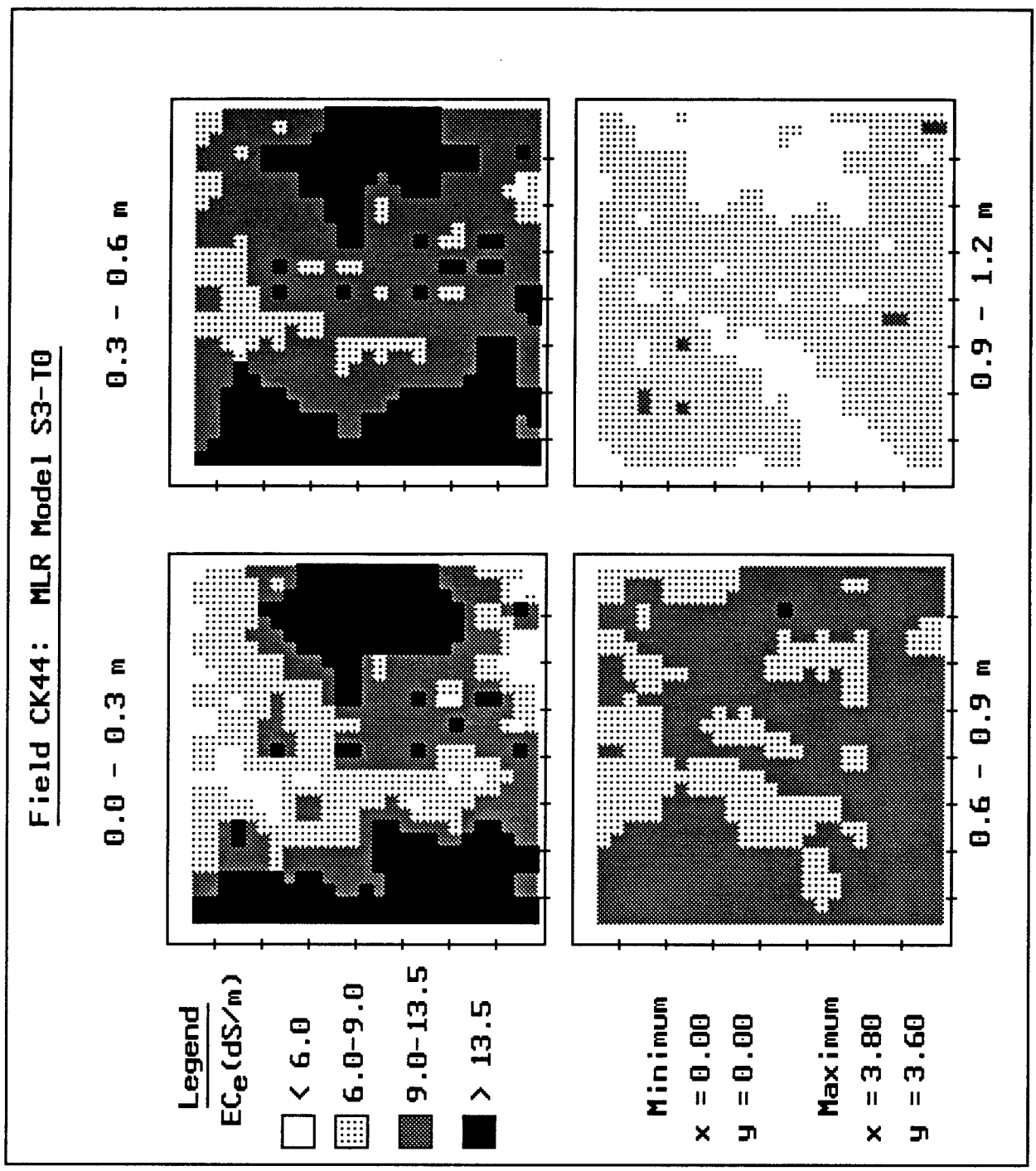


Figure 7.5 Composite printout of the predicted CK44 field salinity maps for the 0.0-0.3 m, 0.3-0.6 m, 0.6-0.9 m, and 0.9-1.2 m depths; input file is *CK44FFIT.M40*.

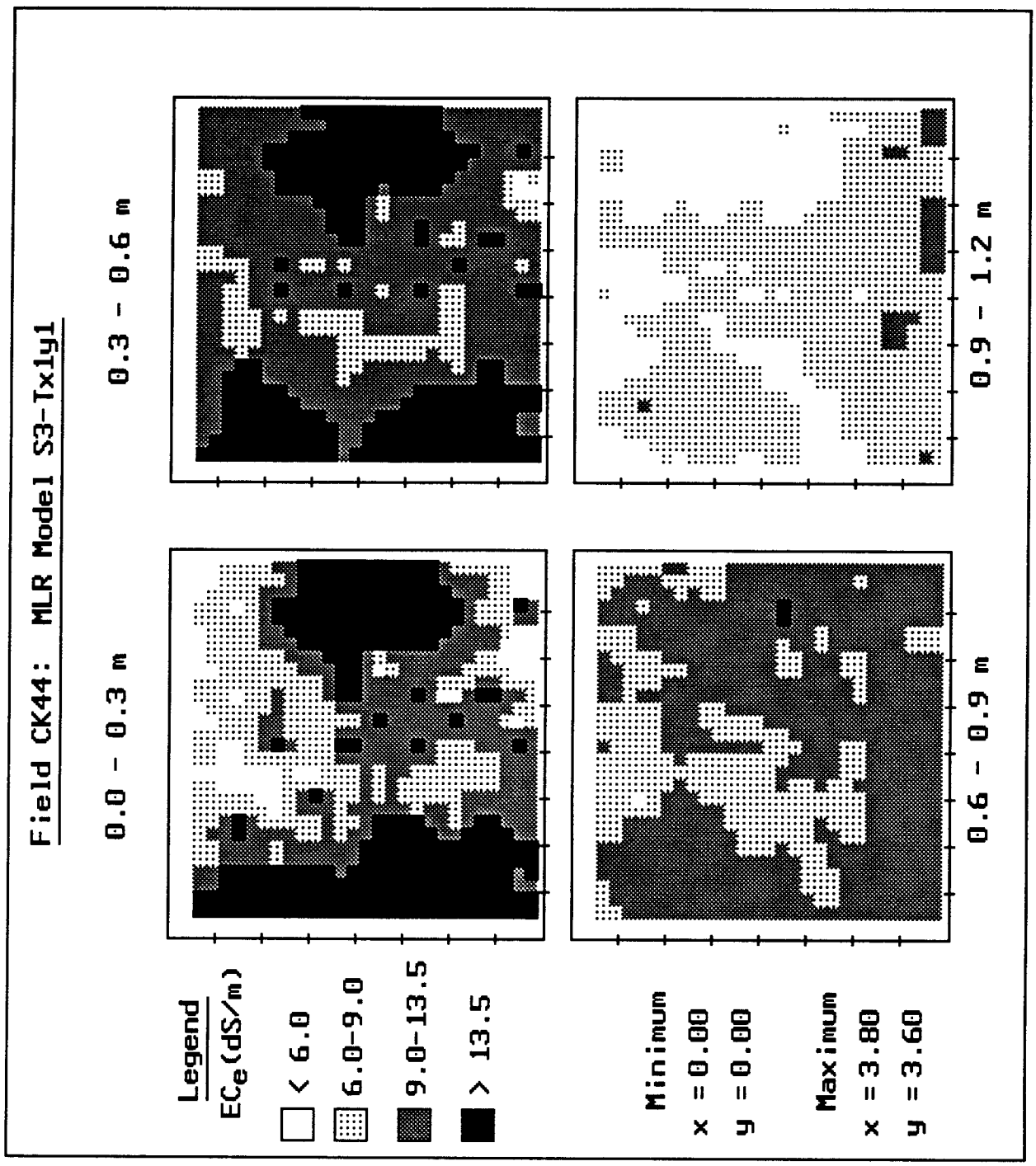


Figure 7.6 Composite printout of the predicted CK44 field salinity maps for the 0.0-0.3 m, 0.3-0.6 m, 0.6-0.9 m, and 0.9-1.2 m depths; input file is *CK44FFIT.M43.*

7.4 Analysis of the AZ09 Survey Data

The ESAP software programs documented in this manual have been found to work well across a wide range of typical field conditions. However, there are a limited number of fundamental assumptions which must be satisfied in order to generate reliable soil salinity predictions. Probably the most important assumption which must be satisfied in practice is that the EM survey readings have not been grossly confounded by significant changes in the soil texture. The AZ09 project represents an example of a survey where this assumption was clearly inappropriate; e.g., the EM survey readings were severely corrupted by chaotic soil texture variations. This data will now be used to demonstrate how such signal/texture confounding can be detected during the model validation analysis.

The above mentioned Arizona survey/salinity data resides in the c:\emsurvey\data subdirectory in a file entitled ***AZ09.DAT.*** Upon reading this file into the ***EMSMLR*** program, you should find that there are 114 survey sites, ***25*** calibration sites, 5 replication sites, and that the salinity data has been collected at 4 sample depths (0.0-0.3, 0.3-0.6, 0.6-0.9, and 0.9-I .2 m). Note also that the Ln salinity data appears to be rather poorly correlated with the 1st principal component score across all four sampling depths.

When you use the Model identification option to compute the APVE and PRESS scores for all 50 models, you should find that the S2i-TO model appears to be the "best". The S2i-TO model estimation and validation results for all four sampling depths are shown in table 7.5. You should immediately note the following two types of residual violations; (1) the LOF F test probability levels are significant at all sampling depths, and (2) there appear to be at least two outliers present in this data (one at the 0.0-0.3 m depth, and another at the 0.6-0.9 m depth). You should also note that the model R^2 values are very low across all sampling depths and the MSE estimates are quite high in the lower two depths.

The results shown in table 7.5 confirm that the S2i-TO regression equations are highly unreliable and clearly biased. Hence, these equations should not be used for spatial salinity prediction purposes.

When there are obvious outliers present in a data set (as is the case here), it is worthwhile to remove the questionable data points and refit the model. You can perform such an analysis by reading in the ***AZ0923.DAT*** data file (note that the sample data associated with sites #7 and #88 have been removed from this file). Invoking the Model Identification option on this data produces a different set of APVE and PRESS scores; the "best" model now appears to be the S2i-Ty1 parameter combination. The R^2 levels, MSE estimates, LOF F test probability levels, and Moran scores for this model are shown in table 7.6 for all 4 sampling depths. Note that the

Table 7.5 Detailed listing of the S2i-T0 model summary statistics across all four sampling depths; input file is *AZ09.DAT.*

Model: S2i-T0
Survey Code: 25-5r/4d (N = 114)

DEPTH:	0.0-0.3 m	0.3-0.6 m	0.6-0.9 m	0.9-1.2 m
LOF F prob level:	0.0376	0.0212	0.0076	0.0057
MORAN score:	-0.088	-0.299	-1.312	-1.284
Linearity of Q-Q Plot:	*poor (outlier)*	*ok*	*poor (outlier)*	*ok*
Res Skew:	-1.322	-0.533	-0.778	-0.384
Balance:	2.351	1.406	1.564	1.474
Abs > 2.5:	yes (-3.4)	yes (-2.6)	yes (-3.0)	no
R/P Plot:	*outlier*	*random*	*outlier*	*random*
R/X Plot:	*outlier*	*random*	*outlier*	*random*
R/Y Plot:	*outlier*	*random*	*outlier*	*random*
Para #:	3	3	3	3
R^2:	0.502	0.485	0.530	0.602
Adj R^2:	0.444	0.426	0.475	0.556
MSE:	0.0714	0.1664	0.3055	0.3014
Model F prob level:	< 0.001	< 0.001	< 0.001	< 0.001
Para |t| prob levels:	< .01: 2 .01-.05: 0 .05-.10: 0 > .10: 1	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 1	< .01: 1 .01-.05: 1 .05-.10: 0 > .10: 1	< .01: 1 .01-.05: 0 .05-.10: 0 > .10: 2
PRESS/n:	0.0788	0.1965	0.3516	0.3399
APVE:	0.0821	0.1912	0.3509	0.3462
Predicted min - max:	0.50, 1.87	0.54, 3.62	0.45, 6.19	0.46, 6.67
Prd Var 50%:	0.0770	0.1793	0.3292	0.3248
75%:	0.0804	0.1873	0.3439	0.3393
100%:	0.1936	0.4509	0.8277	0.8166
Linearity of Ln Plot:	*fair (outlier)*	*fair (weak)*	*poor (outliers)*	*poor (outliers)*

Table 7.6 Pertinent summary statistics from the S2i-Tyl model, using the AZ0923.DA ***T*** input file.

Deoth	R^2	MSE	LOF F test prob level	Moran score
0.0-0.3 m	0.580	0.0423	0.1103	-0.878
0.3-0.6 m	0.622	0.1075	0.0538	-2.379
0.6-0.9 m	0.699	0.1997	0.0194	-0.089
0.9-l .2 m	0.738	0.2054	0.0134	-1.220

spatial bias in the predictions is still present (judging by the LOF F test probability levels). Indeed, the Moran scores have become even more negative, suggesting a possible cyclic residual autocorrelation effect induced from strong textural variations.

In this survey data set, there are actually no physical reasons to justify the removal of the two sites producing the earlier outliers. However, temporarily removing such data points and refitting the regression models can reveal important characteristics about the data set. In this case, we have seen that it is not the outliers that corrupt the model fitting process so much as the short scale spatial autocorrelation inherent in the residuals. In all likelihood, there is nothing "wrong" with the salinity data from sites #7 and #88. Rather, it is simply the extreme textural variation corrupting the EM signal data (and hence invalidating the regression modeling assumptions) which causes the data from these sites to appear so unusual.

Figure 7.7 displays depth-distribution schematic plots of the soil SP measurements associated with the HS2A, WWDI, CK44, and AZ09 calibration samples. These soil SP measurements have been converted into rough texture measurements using the following SP/texture scale:

[SP < 25%]:	sand
[25% < SP < 37.5%]:	sandy loam
[37.5% c SP < 50%]:	loam
[50% < SP < 75%]:	loamy clay
[SP > 75%]:	clay

Note that these texture classes are defined using the dotted vertical lines in Figure 7.7.

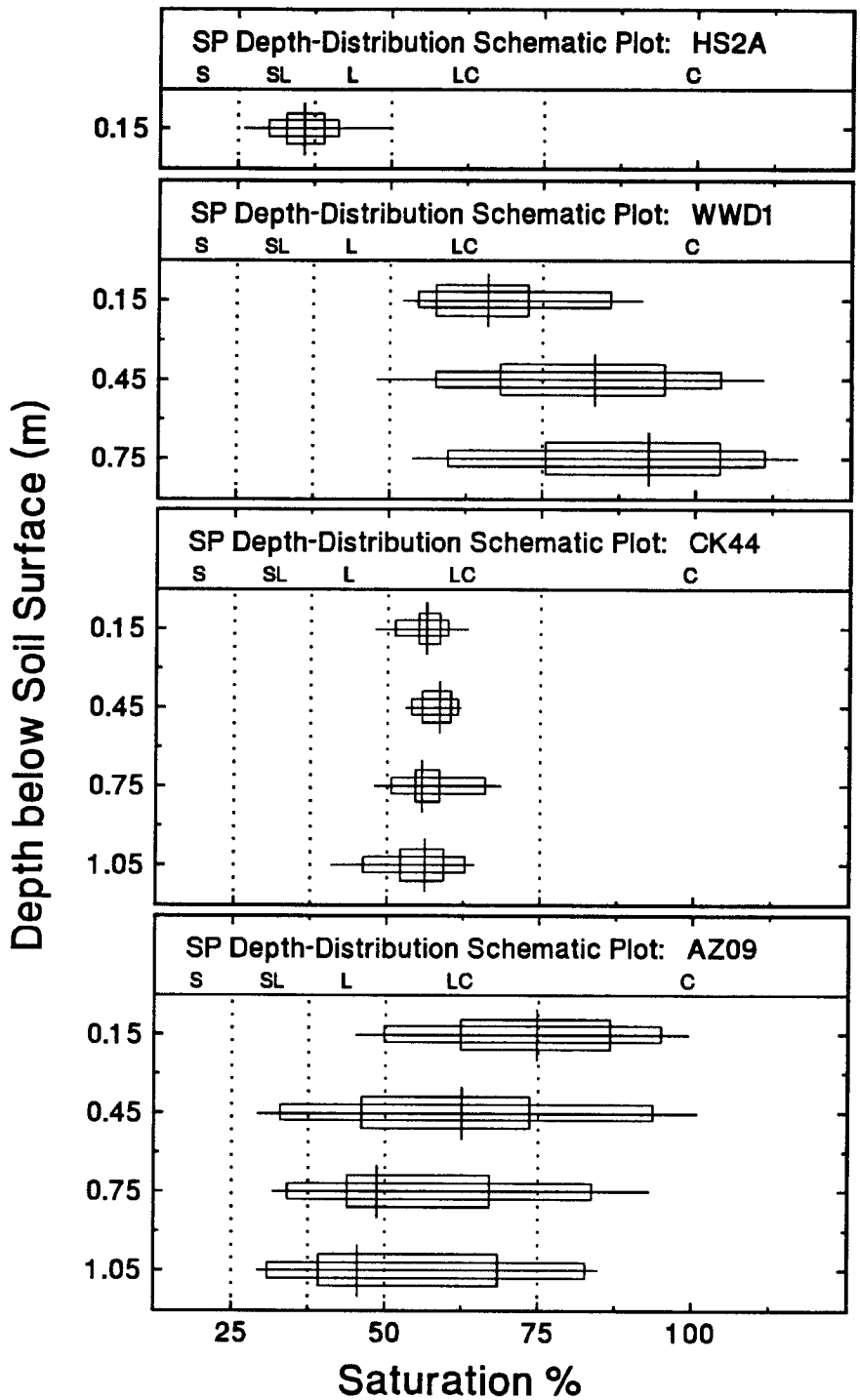


Figure 7.7 Soil SP depth-distribution schematic plots for the sample data from the HS2A, WWD1, CK44, and AZ09 surveys.

A general rule of thumb is that if the majority of SP data covers more than one soil texture class, then the EM signal data may become confounded by texture variations. Furthermore, this confounding can become quite serious if the field EC, levels are relatively low. From Figure 7.7, it is apparent that the soil textural variation in the AZ09 survey data is far too great to successfully employ a regression modeling approach (especially when we consider that 85% of the sample soil salinity data are less than 4.0 dS/m). This may not always hold true; sometimes the soil salinity and soil texture levels change in a closely related manner, as is the case in the WWD1 survey data. However, chaotic textural variations can cause serious regression model bias, and thus we generally don't recommend using this prediction approach under such conditions.

There are other, more sophisticated spatial and/or geostatistical prediction techniques which could be used to model the AZ09.DAT survey data. However, none of these more advanced techniques will be of much practical use unless the calibration salinity sample size is also significantly increased. The interested reader should refer to the references in Lesch, et. al., 1995a for a partial listing of these alternative techniques.

8. REFERENCES

Atkinson, A.C. 1985. Plots, transformations and regression. An introduction to graphical methods of diagnostic regression analysis. Oxford Science Publications, Oxford.

Cook, P.G. and G.R. Walker. 1992. Depth profiles of electrical conductivity from linear combinations of electromagnetic induction measurements. Soil Sci. Soc. Am. J. 56:1015-1022.

Diaz, L. and J. Herrero. 1992. Salinity estimates in irrigated soils using electromagnetic induction. Soil Sci. 154: 151-157.

Lesch, S.M., J.D. Rhoades, L.J. Lund, and D.L. Corwin. 1992. Mapping soil salinity using calibrated electromagnetic measurements. Soil Sci. Soc. Am. J. 56:540-548.

Lesch, S.M., D.J. Strauss and J.D. Rhoades. 1995a. Spatial prediction of soil salinity using electromagnetic induction techniques: 1. Statistical prediction models: A comparison of multiple linear regression and cokriging. Water Resour. Res. 31:373-386.

Lesch, S.M., D.J. Strauss and J.D. Rhoades. 1995b. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resour. Res. 31:3387-398.

McKenzie, R.C., W. Chomistek, and N.F. Clark. 1989. Conversion of electromagnetic inductance readings to saturated paste extract values in soils for different temperature, texture, and moisture conditions. Can. J. Soil Sci. 69:25-32.

McNeil, J.D. 1980. Electromagnetic terrain conductivity measurement at low induction numbers. Tech. Note TN-6, Geonics Limited, Ontario, Canada.

McNeil, J.D. 1986. Rapid, accurate mapping of soil salinity using electromagnetic ground conductivity meters. Tech. No. TN-18, Geonics Limited, Ontario, Canada.

Myers, R.H. 1986. Classical and modern regression with applications. Duxbury Press, Boston, MA.

Rhoades, J.D. 1992. Instrumental field methods of salinity appraisal, pp. 231-48, in Advances in measurement of soil physical properties: Bring theory into practice. G.C. Topp, W.D. Reynolds, and R.E. Green (Eds.). SSSA Special Publ. no.30. ASA, CSSA, and SSA, Madison, WI.

Rhoades, J.D. 1994. Soil salinity assessment: Recent advances and findings. ISSS Sub-Commission A Meetings, Acapulco, Mexico, July 1 O-l 6, 1994.

Rhoades, J.D. and D.L. Corwin. 1990. Soil electrical conductivity: effects of soil properties and application to soil salinity appraisal. Commun. Soil Sci. Plant Anal. 21:837-860.

Rhoades, J.D., D.L. Corwin, and S.M. Lesch. 1991. Effect of soil EC,-depth profile pattern on electromagnetic induction measurements. Research Report #1 25, 108 p.

Rhoades, J.D., N.A. Manteghi, P.J. Shouse, and W.J. Alves. 1989. Estimating soil salinity from saturated soil-paste electrical conductivity. Soil Sci. Soc. Am. J. 53:428-433.

Rhoades, J.D., P.J. Shouse, W.J. Alves, N.A. Manteghi, and S.M. Lesch. 1990. Determining soil salinity from soil electrical conductivity using different models and estimates. Soil Sci. Soc. Am. J. 54:46-54.

Slavich, P.G. 1990. Determining EC_a-depth profiles from electromagnetic induction measurements. Aust. J. Soil Res. 28:443-452.

Weisberg, S. 1985. Applied linear regression. Second Ed. John Wiley & Sons, NY.

Williams, B.G., and G.C. Baker. 1982. An electromagnetic induction technique for reconnaissance surveys of soil salinity hazards. Aust. J. Soil Res. 20:107-I 18.

Wu, L., J.B. Swan, R.R. Allmaras, and S.D. Logsdon. 1995. Tillage and traffic influences on water and solute transport in corn-soybean systems. Soil Sci. Soc. Am. J. 59:185-191.

Yates, S.R., R. Zhang, P.J. Shouse, and M. Th. van Genuchten. 1993. Use of geostatistics in the description of salt-affected lands, pp. 283-304, in Water Flow and Solute Transport in Soils: Developments and Applications. Advanced Series in Agriculture, Series no. 20. D. Russo and G. Dagan (Eds.). Springer Verlag, NY.