1307

WATER RESOURCES RESEARCH, VOL. 31, NO. 2, PAGES 387-398, FEBRUARY 1995

Spatial prediction of soil salinity using electromagnetic induction techniques 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation

Scott M. Lesch

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

David J. Strauss

Department of Statistics, University of California, Riverside

James D. Rhoades

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

Abstract. In our companion paper we described a regression-based statistical methodology for predicting field scale salinity (EC,) patterns from rapidly acquired electromagnetic induction (EC,) measurements. This technique used multiple linear regression (MLR) models to construct both point and conditional probability estimates of soil salinity from EC, survey data. In this paper we introduce a spatial site selection algorithm designed to identify a minimal number of calibration sites for MLR model estimation. The algorithm selects sites that are spatially representative of the entire survey area and simultaneously facilitate the accurate estimation of model parameters. Additionally, we introduce two statistical criteria that are useful for selecting optimal MLR variable combinations, describe a technique for identifying faulty signal data, and explore some of the differences between our recommended model-based sampling plan are some more commonly used design-based sampling plans. Survey data from two of the fields analyzed in the previous paper are used to demonstrate these techniques.

1. Introduction

In the companion paper *[Lesch et al.,* this issue] we described a regression-based statistical methodology suitable for predicting field-scale spatial salinity (EC,) conditions from rapidly acquired electromagnetic induction (EC,) data. This approach, suggested as an alternative to cokriging, was used to produce both point and conditional probability estimates at new EC, survey locations. While both regression and cokriging have certain advantages and disadvantages, a very attractive feature of the former approach is its cost-effectiveness: regression models can be fitted with significantly reduced calibration sample sizes.

In this paper we describe a spatial site selection algorithm specifically designed to identify calibration sites that are well suited for multiple linear regression (MLR) models. As described in the previous paper, we work with EC, data observed at N survey sites. Our task is to choose a "good" subset of n calibration sites at which to sample the soil salinity. Our proposed algorithm selects a limited set of calibration sites ($n \leq 20$) with desirable spatial and statistical characteristics by combining survey site location information with response surface design techniques. Our algorithm ensures that the selected set of calibration sites is (1) spatially representative of the entire survey region and (2) suitable from a statistical design viewpoint, in that the EC, instrument data corresponding to these calibration sites permit efficient estimation of the regression parameters. We note that the proposed algorithm satisfies these two conditions in a quantitative manner but does not find a unique solution to any formal optimization problem.

The sampling algorithm requires that the EC, signal data first be transformed and decorrelated. We show how this can be accomplished using a principal components analysis, explain how such a transformation can be used in conjunction with a response surface design to identify a statistically efficient set of calibration sites, and discuss how the response surface design can be modified in order to optimize the spatial locations of the final calibration sites. We also describe two variable selection criteria that are useful for selecting optimal combinations of regression variables in the MLR model and show how the principal components analysis can be used to detect faulty signal data. We demonstrate these different techniques using the salinity survey and calibration data already introduced in the previous paper.

The details concerning our analysis are, by necessity, specific to the MLR salinity modeling approach already described. However, the techniques we employ are quite general and

Paper number 94WR02180.
0043-1397/95/94WR-02180$05.00

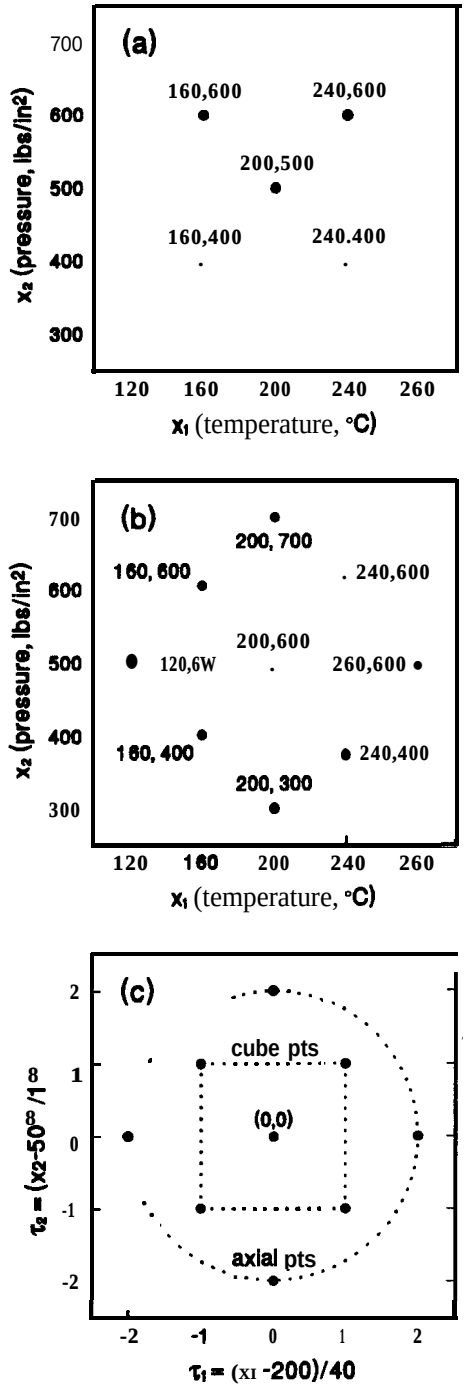

Figure 1. Various types of two-parameter, central composite response surface designs: (a) first-order design, (b) second-order design, and (c) coded second-order design, depicting the center point, four cube points, and four axial points. 1 pound per square inch (lb/in*) equals 6895 Pa.

hence should be applicable to other types of similar environmental calibration (regression) problems.

2. Statistical Methodology

This section is divided into four parts. In the first part we briefly review response surface designs and describe how these designs can be modified to handle spatial survey data. An algorithm for constructing a spatial response surface sampling design is discussed in the second part, along with techniques for detecting faulty signal data. In the third part we derive the general parametric form for a MLR salinity prediction model, and in the fourth we discuss two variable selection criteria useful for identifying optimal variable combinations within the prediction model.

2.1. Site Selection Criteria

We begin our discussion by introducing a series of techniques which are useful for the development of a spatial sampling plan suitable for estimating a regression model. These techniques include response surface designs, principal components analysis, and an algebraic formula for measuring the spatial uniformity of a set of points distributed within a two-dimensional region. We introduce each of these techniques using the hypothetical example outlined below.

Suppose we wish to predict the level of an attribute, y , given knowledge of two correlated covariates, x_1 and x_2. Suppose further that a MLR equation can be assumed a priori to represent a reasonable prediction model. If we can only afford to collect a total of n samples, then at what (x_1, x_2) covariate levels should these n samples be observed at in order to accurately estimate the regression model parameters? To answer this question, we must first specify whether or not the x_1 and x_2 covariates are controllable; e.g., can we actively manipulate (or set) the covariates to one or more prespecified levels before observing y . Sampling strategies specifically designed for estimating regression models under controlled, experimental conditions are known as response surface designs. (An introduction to response surface designs can be found in the work by ***Montgomery*** [1984]; a more thorough treatment of the subject is given by Box and ***Draper [1987].)*** In a response surface design, the response attribute, y, is observed at a limited number of n predetermined covariate levels, where these covariate levels (commonly referred to as the design levels) are chosen using some type of optimality criterion. The usual criterion is to minimize the mean square error (MSE) of the expected regression model; however, other criteria can also be employed [Box ***and Draper, 1987].***

Assume for the moment that x_1 is temperature, x_2 is pressure, and y is the reaction time in a hypothetical chemical process. Two response surface designs commonly used for estimating a regression model relating y to x_1 and x_2 are shown in Figures la and lb. Both designs are known as central composite (CC) designs. Figure la displays a "first-order CC design," since it is used for estimating a first-order (linear) relationship:

$$\hat{y} = \mathrm{PO} + \beta_1 x_1 + \beta_2 x_2. \qquad (1)$$

Figure lb displays a "second-order CC design," useful for estimating a quadratic relationship with interaction between the covariates:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2. \qquad (2)$$

The minimum sample size in the first-order design is $n = 5$, while the second-order design requires at least nine samples. In general, the minimum required sample size will increase as either the expected order of the regression model or the total number of covariates increases. Note also that both CC designs are "orthogonal." Orthogonality implies that the design levels of x_1 and x_2 are statistically independent. This is achieved by

choosing x_1 and x_2 design levels which, when multiplied together, sum to 0; i.e.,

$$\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \approx 0$$

When employing a response surface design, it is common practice to scale the independent regressor variables and work instead with "coded" design levels. An example of this technique is shown in Figure 1c, where we again display our hypothetical second-order CC design using the corresponding coded levels τ_1 and τ_2. Typically, all response surface designs are shown in this coded manner. In Figure lc the coded design level (0, 0) is referred to as the "center point," while the remaining design levels are referred to as either "cube points" or "axial points." In a two-parameter, second-order CC design, the cube points are all the design levels which appear as ($\pm$a, $\pm a$), while the axial points are those levels which appear as either (tb, 0) or (0, $\pm b$), where b > a. (In our example, a = 1 and ***b*** = 2.) In a k-parameter, second-order CC design, there will be one center point, 2^k cube points, written as ($\pm$a, $\pm a, \cdots, \pm a$), and $2k$ axial points, written as ($\pm b$, 0, $\cdots$, 0), (0, $\pm b, \cdots$, 0), $\cdots$, (0, 0, $\cdots, \pm b$). Henceak-parameter, second-order CC design will require at least $2^k + 2k + 1$ sample observations.

In an experimental setting, where all the design variables are controllable, response surface designs represent an effective (model based) sampling strategy for estimating regression models. However, response surface designs have received little, if any, attention in the sampling literature. This is probably due to two reasons: (1) in a typical survey the potential covariates $(x_1, x_2, \cdots, x_k)$ will rarely be controllable or independent and (2) aside from a few types of geostatistical sampling strategies [Russo, 1984], "design-based" sampling plans are usually employed to select the samples. The reader should recognize the two different uses of the term "design" being used here. A design-based sampling strategy describes a type of sampling plan which makes no parametric assumptions regarding the relationship of the response attribute, y, to the x_1, $x_2, \cdots, x_k$ covariates ***[de Gruijter and ter Braak, 1990; Brus and de Gruijter, 1993],*** whereas a model-based sampling strategy can be used when a specific parametric model is assumed to describe the response/covariate relationship [Thompson, 1992]. Examples of design-based sampling strategies include simple random sampling, stratified random sampling, and cluster sampling. Response surface designs represent one example of model-based sampling strategies.

A model-based sampling strategy will be quite advantageous under the right circumstances. For example, an approximate response surface design can be used to greatly reduce the number of sample sites needed to ensure efficient parameter estimates in our salinity regression model. We use the term "approximate" because obviously the EC, covariate readings are not controllable. Hence we first need to observe all of the EC, signal readings before selecting the sample sites. Then, after collecting all of the signal data, we can search through these data and select a subset of survey sites with signal levels that most closely match some theoretical set of response surface design levels. In this paper we will refer to this type of sampling technique as a "pseudo response surface" (PRS) design.

One immediate problem arising from this type of design has already been mentioned; the covariates will likely be correlated. In our case the EC, covariate information often tends to be highly correlated, making it impossible to choose approximately orthogonal PRS design points using the raw signal data. However, this problem can be circumvented by performing a principal components transformation on the signal data [Johnson ***and Wichern,*** 1988]. The EC, instrument readings should first be centered and scaled before applying the principal components transformation (i.e., subtract off the observed mean signal level from each reading and then divide this quantity by the observed signal standard deviation). The transformation procedure can then be performed on these normalized data to produce orthogonal (uncorrelated) principal component data, where each vector of transformed signal observations will have 0 mean and unit variance (after dividing each vector of observations by the square root of the corresponding eigenvalue). Hence these centered and scaled principal component scores can then be directly compared to a set of preselected, coded response surface design levels.

Figure 2a displays a second-order coded CC design (similar to the design shown in Figure lc), with a set of centered and scaled principal component data (κ_1 and κ_2) overlaid on the design plot. In Figure 2a the (τ_1, τ_2) cube and axial design levels have been chosen by superimposing an ellipsoid defined as $\tau_1^2 + \tau_2^2 = 3.84$ onto the graph. We have used a value of 3.84 in order to generate axial design levels equal to ($\pm$ 1.96, 0) and (0, 21.96). (Note that the interval of (-1.96, 1.96) will define the marginal 95% probability interval for κ_1 or κ_2 when κ_1 and κ_2 are assumed to be normally distributed with mean 0 and unit variance.) The nine (κ_1, κ_2) bivariate observations circled in Figure 2a generate the "optimal" PRS design, in that these principal component scores most closely match the (τ_1, τ_2) design levels. In this example, our optimality criterion is to minimize

$$\mathrm{DLS}_j = (\kappa_{1i} - \tau_{1j})^2 + (\kappa_{2i} - \tau_{2j})^2 \qquad (3)$$

where (τ_{1j}, τ_{2j}) is the jth CC design level, (κ_{1i}, κ_{2i}) represents the ith bivariate principal component score, and DLS is an abbreviation for "design level similarity."

A PRS sampling approach like the one outlined above can be used to select a set of calibration sites which should allow for accurate regression model parameter estimates. The principal component scores associated with these sites will also be well "balanced" in a statistical sense; e.g., the observed mean levels of the nine κ_1 and κ_2 observations should be approximately equal to the κ_1 and κ_2 population means. (In Figure 2a the sample κ_1 and κ_2 means are -0.057 and -0.040, while both population means are identically equal to 0.) However, if we were to use such a sampling design to collect spatial data, it is clear that a set of sample sites with very undesirable spatial characteristics could be selected. Figure 2b displays a hypothetical plot of the nine physical sample locations which could have been generated by our PRS design; note that the majority of the sites are clustered in one corner of the field. Intuitively, we should strive to select sample locations in a fairly uniform manner across the entire field, because a uniform (systematic) sampling strategy will result in data which are more representative of the entire survey area. (The analyst should keep in mind that the regression model will only be as representative as the data it is based on.) Additionally, spatially well balanced samples will facilitate the efficient estimation of trend surface parameters in the regression model, if such parameters prove to be necessary. Figure 2c displays a much more uniform distribution of sample site locations. Ideally, we would like to be

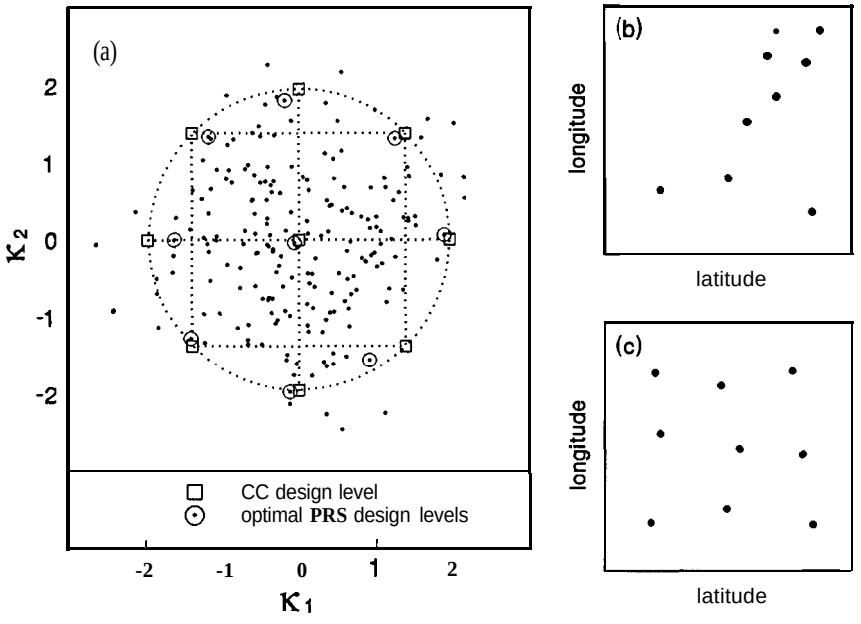

Figure 2. (a) Plot of principal component data overlaid on a second-order central composite design, with optimal PRS sites circled. (b) Plot of hypothetical sample site locations displaying poor spatial uniformity. (c) Plot of hypothetical sample site locations displaying good spatial uniformity.

able to adjust a PRS design to produce a distribution of sample sites which looks something like Figure 2c, rather than Figure 2b.

In order to develop a criterion for measuring the spatial uniformity of a set of sample site locations, we can use a well-known result from geostatistics. Optimal sampling strategies, with respect to minimizing the maximum kriging error, are discussed by McBratney et *al.* [1981] and *Webster and Oliver [1990].* They show that the maximum kriging error within a rectangular region will be minimized when an equilateral triangular survey grid that "fills" the region is employed. Actually, for any given sampling design of size N, this type of survey grid will minimize both the maximum and average distance between a prediction (interpolation) point and the nearest survey point.

This result can be used to develop a measurement of spatial uniformity. Define Ω to be a two-dimensional rectangular region and Y to be a dense, centric systematic grid of survey sites of size N within Ω. Define ψ to be a subset of survey sites from Y of size $n, n <$ N. For example, ψ could represent the set of calibration sites chosen by a PRS design. Define d_i, i = 1, N, to be the physical distance from the i th survey site in Y to the nearest calibration site in ψ. (Note that exactly n of these d_i values will be 0, since n of the survey sites must also be calibration sites.) Then the average distance from a survey site in Y to the nearest calibration site in ψ is

$$\mathrm{AD}(\$) = (1/N)\sum_{i=1}^{N} d_i \qquad (4)$$

where AD is an abbreviation for "average distance."

Equation (4) represents a convenient algebraic formula for measuring the spatial uniformity of a set of calibration sites scattered across n points of a centric, systematic survey grid. From *McBratney et al.* [1981] we know that (4) is minimized when the calibration sites are distributed systematically across the survey grid. Hence the following two "uniformity criteria" can be used for assessing the degree of spatial uniformity inherent in a set of calibration sites:

1. Let $\psi 1$ and $\psi 2$ represent two subsets of calibration sites, both of size n. If $\mathrm{AD}(\psi 1) < \mathrm{AD}(\psi 2)$, then the sample site locations in $\psi 1$ exhibit a more representative distribution within Ω.

2. Let $\psi 1$ represent a subset of calibration sites of size n, and let A represent the remaining survey sites in Y which are not included within \$1. Suppose we wish to add exactly one site from A into $\psi 1$ such that the new set of $n+1$ sites in $\psi 1$ results in the most representative distribution within Ω. Then the "optimal" site to add into $\psi 1$ is the site in A which results in the minimum $\mathrm{AD}(\psi 1)$ score, out of all possible N $-n$ scores.

The above criteria, together with (4), perform two useful functions: (1) to differentiate, with respect to spatial uniformity, between two potential subsets of calibration sites, and (2) to decide on the best location for an additional calibration site, subject to the existing spatial distribution of calibration sites already selected.

We are now in a position to describe how these two uniformity criteria can be combined with the previously discussed techniques for selecting a PRS design, in order to generate a spatial sampling plan with desirable spatial and statistical properties.

2.2. Development of a Spatial Response Surface Sampling Design

Our site selection algorithm identifies an efficient subset of calibration sites for estimating a regression model by incorporating spatial uniformity criteria into a PRS design. The specific assumptions employed in the selection algorithm are outlined below.

We assume that the EC, instrument data have been col-

lected on a centric, systematic grid at N sites ($N \geq 100$) across the entire survey area. For this discussion we will also assume that these EC, data consist of either three or four different signal readings at each site; for example, two EM-38 signal readings (horizontal and vertical mode) and one or two additional signal readings from some other electrical conductivity device (such as an insertion four probe or surface array). (Mention of trademark or proprietary products in this paper does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable.) Let κ_1, κ_2, and κ_g represent the first three centered and scaled principal component scores, computed from the natural log-transformed signal data. (When there are four instrument readings, we assume that the fourth principal component score can be discarded without any significant loss in prediction accuracy, due to the high correlation between signal readings.) Let CC, represent a three-parameter central composite response surface design without a center point, and where the eight cube and six axial design levels of CC,, shown in Table 1, satisfy the following relationship: $\tau_1^2 + \tau_2^2 + \tau_3^2 = 3.84$. Let $\psi 1$ represent the 14 survey sites selected by this CC, design; i.e., $\psi 1$ represents the optimal PRS design, where the optimality criterion employed at each design level is to minimize $DLS_j = (\kappa_{1i} - \tau_{1j})^2 + (\kappa_{2i} - \tau_{2j})^2 + (\kappa_{3i} - \tau_{3j})^2$. Additionally, let $\psi 2$ and $\psi 3$ represent the second and third best PRS designs, such that the principal component scores associated with each survey site in $\psi 2$ produce the second smallest DLS_j values, and each site in $\psi 3$ produces the third smallest values. Let Ψ represent a potential combination of 14 survey sites, where these 14 sites are subsampled only from $\psi 1, \psi 2$, and/or \$3, and such that each site in Ψ is associated with one and only one of the 14 distinct design levels in CC,. Finally, let AD(W) represent the average distance value associated with Ψ.

Given the above definitions the following algorithm can be used to select a subset of calibration sites:

1. Compute the first three principal component scores, compare these scores to the 14 CC, design levels, and identify the three subsets of survey sites for inclusion into the $\psi 1, \psi 2$, and $\psi 3$ PRS designs. Do not allow any survey site to be contained in more than one PRS design or to be associated with more than one CC, design level.
2. Assign $\Psi = \psi 1$ and compute $AD(\Psi)$.
3. Forj = 1 to 14, temporarily interchange (swap) the site in Ψ associated with the jth CC, design level with the site from $\psi 2$ (associated with the same design level). After each swap, recompute $AD(\Psi)$. If AD(V) decreases, permanently interchange these two sites.
4. Repeat step 3, now comparing each site in Ψ with the corresponding site from \$3.
5. Iterate by repeating steps 3 and 4 until no further swapping of sites occurs; i.e., until AD(W) no longer decreases.

Once step 5 is completed, Ψ will contain 14 calibration sites which are spatially distributed across the survey area in a reasonably uniform manner. A limited number of additional sites can then be added to Ψ, one at a time, using our second uniformity criterion. (In steps 6 and 7, Ψ is now allowed to contain more than 14 sites, and n equals the total number of calibration sites contained within Ψ after each iteration.)

6. For $i = 1$ to $N - n$, $(14 \leq n \leq 20)$, temporarily add the ith survey site to Ψ, recompute $AD(\Psi)$, and then remove this site. After computing the $AD(\Psi)$ scores associated with all potential $N - n$ new Ψ subsets, identify the survey site which produced the minimum average distance score, add this site to Ψ, and set $n = n + 1$.
7. Repeat step 6 until the prespecified calibration sample size, $n = n_0$, is reached.

Table 1. Fourteen Response Surface Levels of the Three-Parameter, Second-Order CC Design (Without a Center Point) Used in the Site Selection Algorithm

Design Levels		
τ_1	τ_2	τ_3
1.96	0.00	0.00
-1.96	0.00	0.00
0.00	1.96	0.00
0.00	-1.96	0.00
0.00	0.00	1.96
0.00	0.00	-1.96
1.13	1.13	1.13
1.13	1.13	-1.13
1.13	-1.13	1.13
1.13	-1.13	-1.13
-1.13	1.13	1.13
-1.13	1.13	-1.13
-1.13	-1.13	1.13
-1.13	-1.13	-1.13

A flowchart outlining steps l-7 in this algorithm is shown in Figure 3.

The algorithm outlined above essentially selects a set of calibration sites by sequentially minimizing two optimality criteria. First, three sets of potential calibration sites ($\psi 1$, \$2, and $\psi 3$) are identified by selecting survey sites which minimize the DLS scores. An iterative swapping procedure is then used to select the subset of calibration sites (Ψ) from $\psi 1, \psi 2$, and $\psi 3$ with the minimal AD(V) score. A few additional calibration sites are then added to Ψ, one at a time, again by minimizing the new $AD(\Psi)$ score after each iteration. The end result is a set of calibration sites which are both statistically and spatially well balanced.

Four points concerning this algorithm are worth highlighting. First, with regard to the actual optimality criteria, the ultimate goal is to identify a set of survey sites associated with principal component scores which closely match a set of theoretical response surface design levels, while simultaneously trying to make sure that these sites are distributed as uniformly as possible throughout the survey area. Thus the algorithm selects a "good" subset of sites which will facilitate the accurate estimation of a MLR regression model (which may or may not include all three principal component scores and/or additional trend surface terms). It does not formally minimize the expected MSE of any specific regression equation, nor does it guarantee that the final selected calibration sites will represent the absolute "best" (i.e., optimal) subset of sites, with respect to any residual error and/or model variance criteria.

Second, the algorithm clearly selects sites in a deterministic (i.e., nonrandom) manner. It represents a model-based sampling strategy where the calibration sites are specifically selected to reduce the potential bias and improve the ultimate prediction accuracy of the regression model, assuming that a regression model can indeed explain the response/covariate relationship. This is a very important point. Unlike a design-based sampling strategy, the information gathered from this set of calibration sites cannot, by itself, be used to construct any

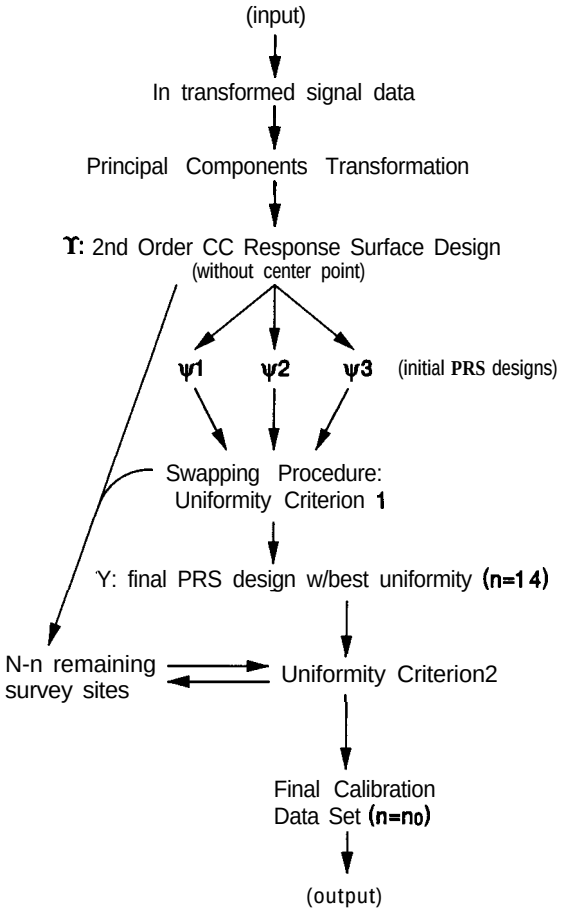

Figure 3. Flowchart depicting the various stages of our sample site selection algorithm (the statistical symbols and terms are explained in the text).

unbiased salinity estimates (such as the field mean salinity level, range intervals, etc.). It is through the fitted regression model that all our salinity estimates arise, and the regression model ultimately dictates the final prediction accuracy and/or inherent bias in each of these estimates. We will expand on this topic in the discussion section; however, for now we simply note that this site selection algorithm should only be used in conjunction with a regression modeling approach, and only when one can be reasonably certain a priori that a regression model will be appropriate.

Third, this algorithm can be used for generating a small set of additional survey sites which will represent the full range of salinity variability present within the survey area and, like the calibration sites, tend to be spread throughout the survey area in a uniform manner. These additional sites can be used in two ways: (1) they can serve as a representative set of monitoring sites, where soil samples could be acquired at some future point in time and then used to test for a change in the field average salinity level, or (2) they can serve as validation sites. In the latter case, soil samples at these sites would be acquired during the initial sampling phase, but not included in the calibration data set. The estimated MLR equation could then be used to predict the EC, levels at these sites, the goal being to assess the prediction accuracy and/or bias using observed salinity data (hence the term "validation" sites). Regardless of their ultimate use, these additional sites can be selected by simply reapplying the site selection algorithm on the $N - n_0$ remaining survey points, using a restricted set of design levels. For example, after identifying the calibration sites, our algorithm will choose eight additional validation/monitoring sites by first selecting three new PRS designs from the remaining survey data using only the cube design levels of the three-parameter, second-order CC design, and then repeating steps 2-5 to identify the final eight sites.

Fourth, the algorithm can be easily modified to incorporate a different type of response surface design. For example, the algorithm could employ a four-parameter, first- or second-order CC design if we wished to systematically sample across all four principal component scores. Likewise, if we only acquired two signal readings (for example, just EM-38 data), the algorithm could rely on a two-parameter CC design. When necessary, the "target" design levels can also be adjusted (by using a suitably chosen constant, different from 3.84). The ability to easily employ different response surface designs makes this type of sampling approach very versatile; one can easily customize the algorithm for different surveying scenarios.

One other important advantage of this algorithm, specifically with regard to the principal components transformation, is that faulty and/or unusual EC, signal data become very easy to detect. As already stated, the centered and scaled principal component scores have means equal to 0, variances equal to 1, and are jointly uncorrelated. From multivariate normality theory *[Johnson and Wichem, 1988]* it can be shown that if X is a p **X N** vector of multivariate normal observations, then the solid ellipsoid of x values satisfying

$$(x - \mathbf{u})'\mathbf{\Sigma}^{-1}(\mathbf{x} - u) \leq \&a) \qquad (5)$$

has a coverage probability of $1 - \alpha$. In (5), x represents a p **X 1** vector of observations from X, **u** represents the mean of X, $\mathbf{\Sigma}$ represents thep **X** p variance-covariance matrix, and χ^2_p represents a chi-square distribution with p degrees of freedom. If we assume that the principal component scores are approximately normally distributed, then (5) reduces to $\kappa_1^2 + \kappa_2^2 + \kappa_3^2 \leq \chi^2_3(\alpha)$. For example, if we set $\alpha = 0.001$, then an ellipsoid defined as $\kappa_1^2 + \kappa_2^2 + \kappa_3^2 \leq 16.27$ should contain 99.9% of the principal component data. Note also that the "statistical" distance from the ith trivariate set of principal component observations $(\kappa_{1i}, \kappa_{2i}, \kappa_{3i})$ to the known distribution mean of (0, 0, 0) issimply $(\kappa_{1i}^2 + \kappa_{2i}^2 + \kappa_{3i}^2)^{1/2}$. Hence, 99.9% Of the trivariate principal component observations should not occur more than $(16.27)^{1/2} \approx 4.03$ units away from (0, 0,0). In the results section we will show some examples of bivariate principal component plots which immediately reveal unusual EC, signal data, based on the above derivations.

2.3. Identification of Potential MLR Model Variables

The calibration site selection algorithm ensures that linear, quadratic, and interaction terms associated with the first three principal components can be estimated in the MLR model. However, most of these parameters are not needed for typical salinity survey data. For example, the relationship between log-transformed soil salinity and the log-transformed EC, signal data is almost always linear. Therefore by applying natural log transformations to the EC, readings (before constructing the principal component scores), we eliminate the need for quadratic terms in the model. Additionally, the multiple-EC,

readings tend to be highly correlated; the first two principal component scores usually account for more than 95% of the total signal variability. We have consistently found that only the interaction between the first and second principal components need be considered. This implies the following MLR salinity model:

$$\ln(EC,) = \beta_0 + \beta_1\kappa_1 + \beta_2\kappa_2 + \beta_3\kappa_1\kappa_2 + \beta_4\kappa_3 + \varepsilon \quad (6)$$

where $\varepsilon \sim N(0, \sigma^2\mathbf{I})$.

Strictly speaking, the spatially homogeneous equation (6) is only appropriate when both the soil texture and soil water content levels can be assumed to be approximately constant across the field. Additional trend surface variables can be included in the MLR model, when gradual fluctuations in either the soil texture and/or water content occur across the survey area. Generally, a second-order trend surface model, when used in conjunction with the principal component variables shown in (6), is adequate for most practical situations. This model would be written as

$$\ln(EC,) = \beta_0 + \beta_1\kappa_1 + \beta_2\kappa_2 + \beta_3\kappa_1\kappa_2 + \beta_4\kappa_3 + \beta_5 x + \beta_6 y + \beta_7 xy + \beta_8 x^2 + \beta_9 y^2 + \varepsilon \quad (7)$$

where the variables x and y are now used to represent the physical (x, y) locations of the sample data and the remaining variables are the same as before.

Different terms in (7) can be systematically removed in order to form simpler MLR prediction models. A listing of possible principal component and trend surface variable subsets is shown in Table 2. There are five different combinations of principal component variables that can be used in a prediction model, assuming that the ith principal component is restricted from entering the model unless the first $i-1$ components are also included. Likewise, there are 10 different sets of possible trend surface variables. Any of these 10 sets of trend surface variables can be combined with the five different subsets of principal component scores, yielding a total of 50 possible parameter combinations in the MLR model.

2.4. Decision Rules for Selecting the Final MLR Model Variables

Variable selection is a critical part of MLR model estimation. A model missing important variables will produce biased predictions, while the inclusion of unnecessary parameters can inflate the mean square error estimate and degrade prediction accuracy. Since the ultimate goal of our MLR model is accurate salinity estimation, rather than parameter inference, variable selection should be based on predictive criteria. Two prediction criteria that are useful for variable selection are sequential cross validation, based on the prediction sum of squares (PRESS) residuals, and the average prediction variance estimate (APVE) associated with the N – n prediction sites.

The ith PRESS residual is generated by removing the ith observation from the calibration data set, refitting the MLR model, estimating the deleted observation, and then computing the corresponding prediction error, e_{-i} [Myers, 1986; *Weisberg,* 1985]. This process is performed for each observation and the PRESS statistic is defined as:

$$\text{PRESS} = \sum_{i=1}^{n} (e_{-i})^2 \quad (8)$$

Table 2. Hierarchical Listing of Five Possible Combinations of Principal Component Variables and 10 Possible Combinations of Trend Surface Components

Abbreviation	Variables
Valid Combinations of Principal Component Variables	
S3i-	$\kappa_1, \kappa_2, \kappa_{12}, \kappa_3$
S2i-	$\kappa_1, \kappa_2, \kappa_{12}$
s3-	$\kappa_1, \kappa_2, \kappa_3$
s2-	κ_1, κ_2
SI-	κ_1
Valid Combinations of Trend Surface Variables	
Tquad	x, y, xy, x^2, y^2
Tx2y2	x, y, x^2, y^2
Tx2yl	x, y, x^2
Txly2	x, y, y^2
TX2	x, x^2
TY2	y, y^2
Tx1y1	x, Y
Txl	x
Ty1	Y
TO	none

The PRESS statistic is a model validation measurement; a small PRESS statistic implies small prediction errors, and hence a better model.

In practice, it is not necessary to estimate the MLR model n times to compute the PRESS residuals. The PRESS residuals can be computed from the ordinary residuals with the following relationship [Myers, 1986]:

$$e_{-i} = e_i/(1 - h_i), \quad (9)$$

where $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. Hence, PRESS statistics for each parameter combination can be computed after the model is estimated, and the model (parameter combination) with the smallest PRESS can be identified.

Recall that the prediction variance for thejth survey site was [Myers, 1986]

$$v_j^2 = s^2[1 + \mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j] = s^2(1 + \text{h,}) \quad (10)$$

We define the average prediction variance estimate (APVE) for the $N - n$ survey sites as

$$\text{APVE} = 1/(N-n)\sum_{j=1}^{N-n} s^2(1 + \text{h,}) = s^2(1 + H)$$

$$H = 1/(N-n)\sum_{j=1}^{N-n} h_j \quad (11)$$

The APVE represents an estimate of what the average prediction variance should be, assuming that the model parameterization is correct. MLR models with small average prediction variance estimates are preferable, since they will theoretically be the most accurate.

The APVE can be computed for each model and used in conjunction with the PRESS statistics to identify the final "best" combination of model variables. Sometimes both statistics will clearly identify one model as superior. However, sometimes these two statistics may not agree, and/or may not consistently select the same model parameterization across multiple depths. We have found these two criteria to be most effective when they are used as a guideline for identifying a few

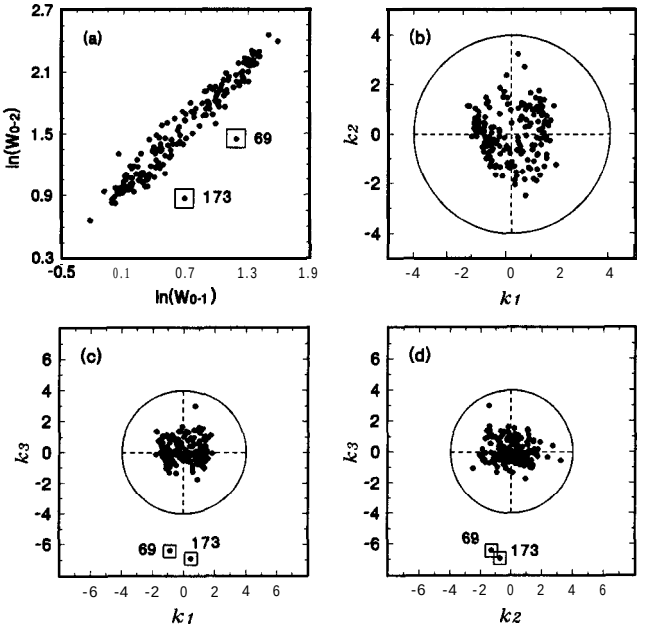

Figure 4. Various plots of EC, signal data from field S2A: (a) In Wenner data recorded at 1- and 2-m spacings, where sites 69 and 173 appear unusual; (b) plot of κ_1/κ_2, with no outliers present; (c) plot of κ_1/κ_3, with two significant outliers (sites 69 and 173); and (d) plot of κ_2/κ_3, with two significant outliers (sites 69 and 173).

good "candidate" models. More comprehensive statistical criteria can then be used to select the final MLR parameterization.

3. Application Examples

Examples of EC, signal verification, calibration site selection, and MLR variable identification techniques are discussed below using field survey data introduced in the previous paper. Survey data from field WWD-1 are used to demonstrate both the detection of faulty signal data and the calibration site selection algorithm. An example of MLR variable selection and model identification is demonstrated with the data from field S2A.

The original survey data in field WWD-1 were collected on a systematic grid consisting of 14 rows with 13 survey sites per row. The row spacing was 50 m, while sites within a row were spaced 55 m apart. The total number of survey sites was 14 × 13 = 182. However, no survey data was collected at the last site in row 1 because an evaporation pond had been installed in the northeast corner of the field. Additionally, the Wenner instrument data from the first site in row three had to be discarded because one of the probes failed to make contact with the soil (the furrow at this site had a noticeable tractor tire rut). Hence usable survey data were acquired at 180 sites only.

All bivariate plots of the four EC, signal readings (horizontal and vertical mode EM-38 data and 1- and 2-m spacing Wenner array data) appeared quite correlated. Additionally, the bivariate plot of the Wenner data appeared to contain at least two unusual observations: sites 69 and 173. A plot of this log-transformed data is shown in Figure 4a. At both sites it appeared that the Wenner readings from the 2-m spacing were considerably lower than they should have been, given the l-m readings.

A principal component transformation was applied to both the EM-38 and Wenner survey data, and bivariate plots of the first three scaled and centered principal component scores were constructed. As previously discussed, if the principal component scores are assumed to be distributed as independent N(0, 1) random variables, no data should lie more than about 4 units from the mean. In Figures 4b, 4c, and 4d, the three bivariate plots are shown with circles of radius 4 overlaid on the principal component data. Sites 69 and 173 clearly appeared to be outliers in both the κ_1/κ_3 and κ_2/κ_3 plots; note that each fell more than 6.4 units (standard deviations) away from the means of each bivariate distribution. Hence the Wenner data were judged to be unreliable, and both of these sites were removed from the survey data.

A second principal component transformation was applied to the log survey data from the remaining 178 sites, after removing sites 69 and 173. Only one site (site 99) appeared to lie more than 4 units away from the bivariate means in any of the plots. Since the bivariate principal component data for site 99 just marginally exceeded the 4-unit threshold in both the κ_1/κ_3 and κ_2/κ_3 plots, we chose not to delete this site from the survey data. However, we did remove this site from possible consideration before running the sample site selection algorithm, so that this marginal outlier would not inadvertently be chosen as one of the calibration sites.

Principal component transformation statistics (correlation matrices, eigenvalues, and eigenvectors) for both the full (N = 180) and reduced (N = 178) survey data sets are shown in Table 3. Deletion of sites 69 and 173 resulted in only minimal changes in the eigenvalues and eigenvectors. Note also that the estimated coefficients in the final eigenvectors were physically meaningful. The first eigenvector, which explained the dominant proportion of the log EC, variability, was basically a simple average of the four individual signal readings. The second eigenvector represented a contrast between the EM-38

Table 3a. Principal Component Transformation Statistics on Log-Transformed WWD-1 Survey Data (Correlation Matrix, Eigenvalues, and Eigenvectors) for Full Data Set (N = 180)

Correlation	ln (EM_V)	ln (EM_H)	ln ($W_{0\text{-}1}$)	ln ($W_{0\text{-}2}$)
ln (EM_V)	1.0000	0.9927	0.8565	0.9209
ln (EM_H)		1.0000	0.8990	0.9492
ln ($W_{0\text{-}1}$)			1.0000	0.9741
ln ($W_{0\text{-}2}$)				1.0000

Eigenvalues	Percent Variability	Total Percent Variability
3.79676	0.94919	0.94919
0.18494	0.04624	0.99543
0.01503	0.00377	0.99920
0.00327	0.00080	1 .ooooo

Variable	First Eigenvector	Second Eigenvector	Third Eigenvector
ln (EM_V)	0.4966	-0.5794	-0.0850
ln (EM_H)	0.5059	-0.3738	-0.1859
ln ($W_{0\text{-}1}$)	0.4910	0.6572	-0.5543
ln ($W_{0\text{-}2}$)	0.5063	0.3044	0.8068

and Wenner data, which is essentially a contrast between deep (O-l.2 m) and shallow (O-O.6 m) signal information. The third eigenvector seemed to be primarily a contrast between the two Wenner readings, which represents a contrast between two shallow readings: O-O.6 m versus O-O.3 m.

After validating the survey data from field WWD-1,178 sites were left in the data set, of which 177 represented potential calibration site locations. The site selection algorithm was then used to select 16 calibration sites and eight additional validation sites; these computations were carried out in the field with a portable 386 personal computer and customized site selection software developed at the U.S. Salinity Laboratory.

Figures 5a-5d show various stages of the calibration and prediction site selection process. Figure 5a shows the 42 sites selected to be in the three initial PRS designs (ψ1, $2, and $3). While these sites appeared to be spread approximately throughout the field, note that they tended to be distributed in clusters. Figure 5b shows the final 14 sites contained in Ψ which most closely matched the design levels, after adjusting for spatial location. Note that the final 14 sites were distributed almost systematically throughout the field. The two sites shown as double circles were added after the 14 design sites had been selected, in order to bring the sample size up to 16. The algorithm based the selection of these last two sites solely on their spatial locations. Figure 5c shows the original locations of eight additional validation sites. These sites were selected to correspond to the eight CC, "cube" design levels, $[\kappa_1, \kappa_2, \kappa_3] = (\pm 1.13, \pm 1.13, \pm 1.13)$, again after adjusting for spatial location. These design levels were used so that the validation data set would represent the full range of soil salinity levels throughout the field. We manually decided to move one of the sites (as shown in Figure 5c) to achieve a more uniform spatial pattern. Figure 5d displays the final locations of the 16 calibration and eight prediction sites. The five calibration sites shown as double squares represent the five calibration locations where duplicate soil core samples were extracted.

After the calibration sites within a field have been selected, soil samples from each site can be extracted and returned to the laboratory for analysis. Upon determining the soil EC, levels, the MLR model-building stage begins. The first step to successful model estimation is the identification of an optimal combination of MLR variables. MLR variable combinations that resulted in the five lowest PRESS and APVE statistics for field S2A are shown in Table 4. Note that four models were common to both lists: S2-Tx2y1, S3-Tx2y1, S3-Tyl, and S3-Ty2.

Some pertinent summary statistics for these four MLR models are shown in Table 5. These include the number of parameters, R^2, adjusted R^2, the model mean square error (MSE), a jackknifed estimate of the MSE (computed from the PRESS statistics), and the APVE statistic. On the basis of both these statistics and the residual diagnostic plots (not shown), we decided to use the S3-Tx2yl parameterization for prediction purposes. Although this model contained the highest number of variables (6) it had the lowest MSE and jackknifed MSE estimates, the highest adjusted R^2, and nearly the lowest APVE statistic. The final parameter estimates for this model are given in the footnote to Table 5.

For comparative purposes we computed prediction summary statistics using all four model parameterizations; these results are shown in Table 6. The predicted field average ln (EC,) levels were all within 0.023 units of each other, and all four confidence intervals contained the observed ln (EC,) level. The range interval estimates produced by each model were also nearly equivalent. The prediction correlation matrix shown in Table 6b confirmed that even the individual predictions tended to be quite similar. Although we chose to use the S3-Tx2yl model, the results in Table 6 confirm that any one of the other three MLR model parameterizations could have been employed without seriously altering the prediction statistics and/or the estimated spatial salinity map.

Table 3b. Principal Component Transformation Statistics on Log-Transformed WWD-1 Survey Data (Correlation Matrix, Eigenvalues, and Eigenvectors) for Data Set After Removal of Two Unusual Sites (N = 178)

Correlation	ln (EM_V)	ln (EM_H)	ln ($W_{0\text{-}1}$)	ln ($W_{0\text{-}2}$)
ln (EM_V)	1.0000	0.9924	0.8573	0.9114
ln (EM_H)		1.0000	0.8997	0.9373
ln ($W_{0\text{-}1}$)			1.0000	0.9627
ln ($W_{0\text{-}2}$)				1.0000

Eigenvalues	Percent Variability	Total Percent Variability
3.78090	0.94522	0.94522
0.18549	0.04637	0.99159
0.03017	0.00754	0.99913
0.00344	0.00087	1.00000

Variable	First Eigenvector	Second Eigenvector	Third Eigenvector
ln (EM_V)	0.4976	-0.5806	-0.0430
ln (EM_H)	0.5065	-0.3800	-0.1981
ln ($W_{0\text{-}1}$)	0.4917	0.6395	-0.5692
ln ($W_{0\text{-}2}$)	0.5040	0.3311	0.7968

4. Discussion

The initial field EC, data should always be validated before the calibration site locations are chosen. During rapid field survey operations, "bad" instrument data can occur. For example, the probes in the automated Wenner array system (mounted underneath a mobile assessment vehicle) may occasionally fail to make good contact with the soil. One or more probes can get inadvertently inserted into large soil cracks and hence fail to transmit or receive a conductance signal, or a physical obstruction (such as a stone) can block a probe from reaching its penetration depth, thus resulting in faulty data. Likewise, EM-38 readings can be seriously distorted if small metallic objects are buried near the sample site. Plotting the bivariate principal component data is an easy and effective way to identify questionable survey readings. Faulty readings can then be removed immediately after the survey data have been collected, before the soil sampling begins.

The calibration site selection algorithm described in this paper has some very definite advantages and disadvantages in comparison with other types of sampling plans. As we have pointed out, it employs a model-based, nonrandom identification strategy to select a set of calibration sites with desirable spatial and statistical properties. Furthermore, it has been specifically created for estimating a regression model, as opposed to directly estimating any field salinity statistics. Hence it has been designed for use with (and only with) the MLR modeling approach described by ***Lesch et al.*** [this issue].

Brus and de Gruijter [1993] and ***de Gruijter and ter Braak [1990]*** offer some convincing arguments in favor of using a

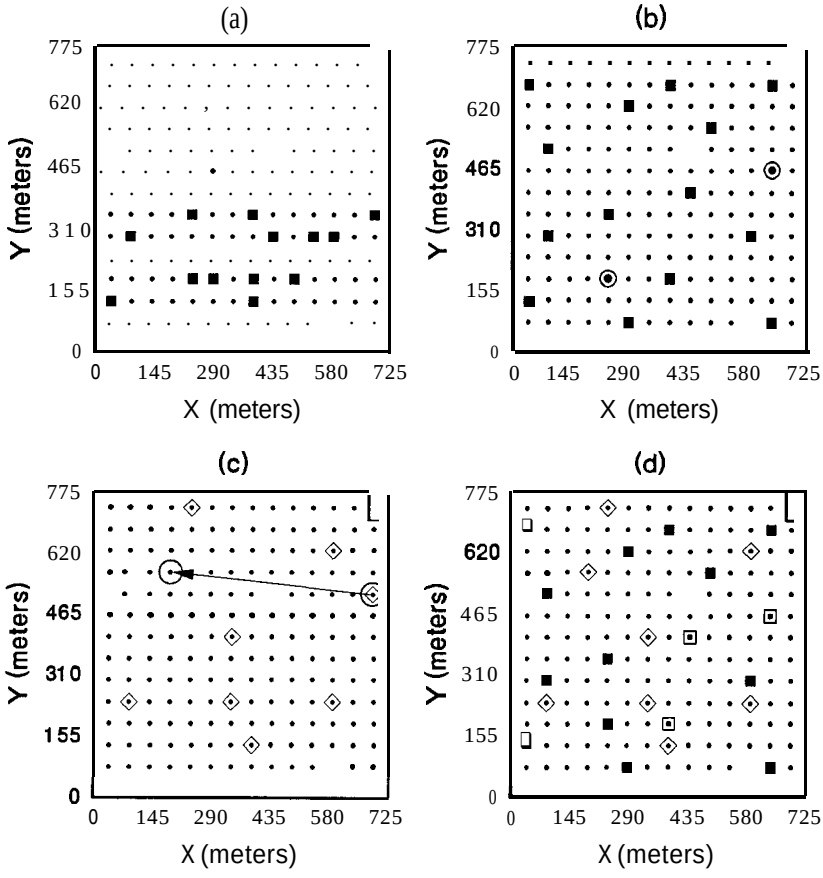

Figure 5. Various stages of the calibration and validation site selection process in WWD-1: (a) initial 42 sites chose by the response surface design; (b) final 14 calibration sites, with two additional sites added to improve spatial uniformity; (c) location of the eight validation sites, with one site manually relocated to improve spatial uniformity; and (d) locations of final 16 calibration and eight validation sites.

design-based approach (which includes design-based sampling plans) for spatial inference. They point out that model-based approaches, such as geostatistical techniques like kriging, make a number of parametric assumptions (i.e., model assumptions) which may or may not be satisfied in practice. Furthermore, the concept of unbiasedness, with respect to some population estimate (or prediction) has a different meaning under the two approaches.

Design-based sampling plans and methods of inference have some advantages in that they can be more objective (i.e., they require no subjective estimation of parametric model parameters) and hence may be more appropriate for some types of contamination and/or assessment studies. However, design-based methods of inferences cannot produce point estimates of a response variable at any nonsampled locations, nor can they incorporate covariate information in any direct, efficient manner. On the other hand, model-based sample inference methods are specifically designed for point prediction. Both cokriging and MLR models effectively incorporate covariate information which can be used to increase the prediction accuracy at nonsampled locations throughout the survey area.

Nonetheless, while both cokriging and regression are model-based approaches, the analyst must keep in mind that some of their parametric assumptions are quite different. In a cokriging approach a fundamental assumption is made that approximately unbiased estimates of the variogram functions are obtainable and, additionally, that all the variables are stochastic. Theoretically, one cannot choose sampling locations which in any way depend on the expected level of the response variable, since such a sampling design would induce unintended bias into the shapes of the variograms. However, in a regression model the covariates are considered to be deterministic; only the residual error term is considered stochastic. Additionally, no variogram estimates are required, nor will the MLR model predictions depend in any direct way on the physical locations of the calibration sample sites. Hence it is not only permissible, but often preferable, to select the calibration sites in a non-random manner, provided the selection procedure is designed to minimize the error associated with the parameter estimates.

In a regression model, prediction accuracy will be degraded and/or bias may be induced when either (1) the model is misspecified or (2) the parameters are poorly estimated. Model misspecification can occur either because one or more important parameters are missing from the model or when the residual error assumptions are grossly violated. Likewise, parameters can be poorly estimated because (1) one or more independent regressor variables are highly correlated, (2) the range between the highest and lowest response levels for one or more regressors is too narrow, and hence deterministic effects caused by changing the regressor response level(s) are drowned out by random noise, and/or (3) the sampling design is poorly balanced with respect to the regressor levels on which

Table 4. Five Best Models (MLR Variable Combinations) for Field S2A, as Determined by PRESS and APVE Statistics

PRESS Statistics			APVE Statistics		
Rank	Model	Value	Rank	Model	Value
1	S3-Tx2yl	2.06	1	S2-Tx2yl	0.106
2	S3-Tyl	2.09	2	S3-Tx2yl	0.107
3	S3-Tx2y2	2.12	3	Sl-Tx2yl	0.111
4	S3-Ty2	2.12	4	S3-Tyl	0.112
5	S2-Tx2yl	2.31	5	S3-Ty2	0.116

the predictions will be based (i.e., the sample regressor levels are not representative of the population levels, and hence the estimated regression model is not based on data which are representative of the population to be predicted.)

In our companion paper we dealt with one aspect of potential model misspecification, i.e., the independent error assumption. We stressed the use of the Moran residual test, and more importantly, sampling designs which facilitate the construction of a lack-of-fit test. Intuitively, if one or more covariates are being corrupted by some secondary (unknown) spatial effect, then the regression model will be biased, and the residuals associated with duplicate samples should be highly correlated. On the other hand, an unbiased regression model should produce a MSE estimate which is approximately equivalent to the pure error estimate, which in a spatial setting is equivalent to the observed "nugget" variance of the response variable. This is why we stress constructing a lack-of-fit test; it supplies the analyst with a direct means of testing for model bias.

In this paper we have dealt with the remaining issue of model misspecification, i.e., regression parameter identification techniques. We have proposed two statistical criteria for identifying important parameters, the PRESS and APVE statistics. These statistics can be used in conjunction with other statistical (and often nonstatistical) criteria to help identify a worthwhile set of regressor covariates for prediction purposes, provided the underlying assumptions regarding the applicability of a regression relationship are approximately satisfied.

We have also described in detail how a sampling plan can be designed to ensure efficient parameter estimates. Note that our algorithm effectively performs three functions. First, it decorrelates the signal data, hence removing the problems associated with multicollinear regressor variables. Second, by employing a suitable underlying response surface design, it effectively maximizes the range between the sampled response levels of each regressor variable (principal component score) associated with the signal data. Third, by employing these same response surface design techniques in conjunction with spatial uniformity criteria, it significantly increases the probability of choosing calibration sites which are well balanced in both a spatial and statistical sense (i.e., both the independent regressor levels and calibration sample locations are representative of the survey population). Hence our algorithm effectively deals with the three main issues which determine the precision of the parameter estimates and thus, in turn, serves to maximize the prediction accuracy and minimize the bias inherent in the fitted model.

Table 5. Model Summary Statistics for Final MLR Variable Combinations: p (Number of Parameters), R^2, Adj R^2 (Adjusted R^2), MSE, JMSE (Jackknifed MSE), and APVE

Model	p*	R^2	Adj R^2†	MSE	JMSE‡	APVE
S2-Tx2yl	5	0.920	0.892	0.0787	0.1155	0.106
S3-Tx2yl	6	0.933	0.903	0.0708	0.1030	0.107
S3-Tyl	4	0.907	0.883	0.0852	0.1045	0.112
S3-Ty2	5	0.916	0.886	0.0829	0.1060	0.116

Parameter estimates for S3-Tx2yl are as follows: standard deviations are shown in parentheses: E[ln (EC,)] = [0.730 ± (0.12)] + [0.982 ± (0.08)]κ_1 – [0.088 ± (0.07)]κ_2 – [0.155 ± (0.09)]κ_3 – [0.024 ± (0.06)]x + [0.159 ± (O.O7)]y + [0.219 ± (0.10)]x^2

*Not including intercept.

†Adj $R^2 = 1 - [(n - 1)/(n - p - 1)](1 - R^2)$.

‡JMSE = PRESS/n.

Table 6a. Prediction Summary Statistics for Final MLR Parameter Combinations: Field Average ln (EC,) and Range Interval Estimates (Observed Values Also Shown)

	MLR Prediction Models				
	Observed Data	S2-Tx2yl	S3-Tx2yl	S3-Ty1	S3-Ty2
G	1.017	0.967 (0.84, 1.10)	0.957 (0.83, 1.08)	0.951 (0.82, 1.08)	0.944 (0.81, 1.08)
Θ[0, 2]	0.425	0.427	0.430	0.422	0.419
Θ[2, 4]	0.231	0.236	0.247	0.275	0.270
Θ[4, 8]	0.161	0.190	0.179	0.161	0.168
Θ[8,16]	0.134	0.100	0.092	0.087	0.089
Θ[>16]	0.049	0.047	0.052	0.055	0.054

G denotes field average ln (EC,); @[a, b] denotes range interval estimate. Values in parentheses denote 95% confidence intervals.

Two other points deserve some discussion. First, some authors in the nonspatial sampling literature have been rather critical of applying regression models to survey data. For example, ***Hansen et al.*** [1983] suggested that a regression modeling approach can lead to substantial bias in the predicted population mean within a sampling survey. However, their conclusion was refuted by ***Cumberland and Royal1*** [***1988],*** who showed that the reason for the bias in their study was due primarily to their reliance on a (design based) simple random sampling scheme for choosing the calibration data. Cumberland and Royal1 showed that a model-based estimation technique requires a sampling scheme which is well balanced (i.e., the sample and population means of the regressor variables must be approximately equal) and that simple random sampling does not ensure this, regardless of the sample size. Additional comments concerning the need for adequate balance in model-based estimation techniques are given by ***Thompson [1992], who*** discusses the pros and cons of design- and model-based sampling inferences in a more general setting.

Second, in the spatial sampling literature, *Brus* ***and de*** *Gruijter*

Table 6b. Prediction Summary Statistics for Final MLR Parameter Combinations: Correlation Matrix for the Four Sets of Prediction Data

	s2-Tx2yl	S3-Tx2yl	S3-Tyl	S3-Ty2
S2-Tx2yl	1.000	0.986	0.928	0.923
S3-Tx2yl		1.000	0.975	0.970
S3-Tyl			1.000	0.996
S3-Ty2				1.000

[1993] refer to *Borgman and Quimby [1988],* who state that one of the greatest shortcomings of the geostatistical (i.e., model based) approach is the inherent difficulty in validating the model assumptions. While this is, to a certain extent, a valid point, we feel that in this particular example our regression modeling approach is less subject to such a criticism. There is a battery of residual diagnostic techniques available to the analyst when fitting a regression model. These techniques, when used in conjunction with duplicate sampling (for constructing residual lack-of-fit tests) and the acquisition of independent validation sites (to test for correct parameter specification and/or model prediction bias), should prove to be more than adequate for assessing the inherent regression model assumptions.

In conclusion, our suggested sampling algorithm will prove to be worthwhile when a strong correlation exists between the target response level and one or more covariates and the covariates themselves are not corrupted by additional, unknown spatially dependent attributes. Provided the fitted MLR model can be validated using the assessment techniques described above, the predictions should prove to be approximately unbiased and considerably more accurate than any predictions and/or estimates produced by other types of model based (geostatistical) or design-based inference methods employing the same, limited calibration sample size. Furthermore, the highly nonrandom sampling plan we employ serves to significantly reduce, rather than inflate, the prediction bias. On the other hand, if the modeling assumptions are grossly violated, then the resulting regression equation will more than likely be biased and produce unreliable predictions. Furthermore, the samples themselves cannot be used to construct unbiased estimates of the salinity population parameters, because these samples have not been collected in a random manner. In this latter scenario we suggest that the regression approach be discarded and replaced with some other appropriate model- or design-based inference technique and sampling plan.

The site selection software used for these surveys is available from the authors on request. It is designed for use on IBM compatible personal computers (386 microprocessor or higher strongly recommended) and can be used in conjunction with the salinity estimation software described in the companion paper.

5. Conclusion

A deterministic, model-based spatial site selection algorithm incorporating both classical and geostatistical selection criterion has been described. This algorithm uses a response surface design to select three subsets of appropriate sites for soil sampling and then iteratively selects sites from these subsets to produce a final calibration set with a spatially uniform sampling pattern. The algorithm can also add additional sites to increase the sample size and/or uniformity of the sampling pattern and generate a second set of site locations suitable for sampling at some later date. We have discussed the differences between this algorithm and other types of sampling plans and demonstrated why such a deterministic scheme should be preferred, provided the underlying regression modeling assumptions are appropriate.

Two statistical criteria useful for MLR variable selection have also been discussed: the PRESS and APVE statistics. Both of these statistics can help identify the final model parameters that will minimize the MLR prediction errors and maximize the prediction accuracy. Additionally, a technique for detecting faulty or questionable survey data has been described, based on the magnitude of the principal component scores. This screening technique will help prevent corrupted survey data from inadvertently influencing either the sampling plan and/or final regression model predictions.

The site selection algorithm, signal validation techniques, and variable selection criteria have been designed to be used in conjunction with the MLR modeling and prediction techniques described by *Lesch et al.* [this issue]. Together these techniques represent a comprehensive salinity monitoring and assessment methodology. The number of soil samples can be minimized while still retaining the prediction accuracy inherent in statistical calibration techniques, hence facilitating an assessment methodology that can be applied in a rapid, practical, and cost-effective manner.

Acknowledgments. We would like to thank two anonymous referees whose help improved earlier versions of these manuscripts by providing a number of constructive comments and suggestions. We are also indebted to William Alves, Robert LeMert, and Nahid Manteghi for their assistance with the field surveying work and laboratory analysis of the salinity data discussed in these manuscripts.

References

Borgman, L. E., and W. F. Quimby, Sampling for tests of hypothesis when data are correlated in space and time, in *Principles of Environmental Sampling,* edited by L. H. Keith, pp. 25-43, American Chemical Society, Washington, D. C., 1988.

Box, G. E. P.. and N. R. Draper. *Empirical Model-Building and Response Surfaces,* John Wiley,New York, 1987.

Brus, D. J., and J. J. de Gruijter, Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science, *Environmetrics, 4, 123-152, 1993.*

Cumberland, W. G., and R. M. Royall, Does simple random sampling provide adequate balance?, *J. R. Stat. Soc. B, 50, 118-124, 1988.*

de Gruijter, J. J., and C. J. F. ter Braak, Model-free estimation from spatial samples: A reappraisal of classical sampling theory, *Math. Geol., 22, 407-415, 1990.*

Hansen, M. H., W. G. Madow, and B. J. Tepping, An evaluation of model-dependent and probability-sampling inferences in sampling surveys, *J. Am. Stat. Assoc., 78, 776-793, 1983.*

Johnson, R. A., and D. W. Wichern, *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice-Hall, Englewood Cliffs, N. J., 1988.

Lesch, S. M., D. J. Strauss, and J. D. Rhoades, Spatial prediction of soil salinity using electromagnetic induction techniques, 1, Statistical prediction models: A comparison of multiple linear regression and cokriging, *Water Resour. Res.,* this issue.

McBratney, A. B., R. Webster, and T. M. Burgess, The design of optimal sampling schemes for local estimation and mapping of regionalized variables, I, Theory and methods, *Comput. Geosci., 7, 331-334, 1981.*

Montgomery, D. C., *Design and Analysis of Experiments,* 2nd ed., John Wiley, New York, 1984.

Myers, R. H., *Classical and Modem Regression With Applications,* Duxbury Press, Boston, Mass., 1986.

Russo, D., Design of an optimal sampling network for estimating the variogram, *Soil Sci. Soc. Am. J., 48, 708-716, 1984.*

Thompson, S. K., *Sampling,* John Wiley, New York, 1992.

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

Weisberg, S., *Applied Linear Regression,* 2nd ed., John Wiley, New York, 1985.

S. M. Lesch and J. D. Rhoades, U.S. Salinity Research Laboratory, 4500 Glenwood Drive, Riverside, CA 92501.

D. J. Strauss, Department of Statistics, University of California, Riverside, CA 92502.

(Received January 19, 1994; revised August 1, 1994;
accepted August 18, 1994.)