Lesch, S.M. 2012. Statistical models for the prediction of field scale, spatial salinity patterns from sol conductivity survey data. In: W.W. Wallender and K.K. Tanji (eds.) ASCE Manual and Reports on Engineering Practice No. 71 Agricultural Salinity Assessment and Management (2nd Edition). ASCE, Reston, VA. Chapter 14 pp:461-482.

CHAPTER 14

STATISTICAL MODELS FOR THE PREDICTION OF FIELD-SCALE AND SPATIAL SALINITY PATTERNS FROM SOIL CONDUCTIVITY SURVEY DATA

S. M. Lesch

INTRODUCTION

The collection of apparent soil electrical conductivity (EC_a) survey data for the purpose of characterizing various spatially referenced soil properties has received considerable attention in the soils literature over the last two decades (Corwin and Lesch 2005a,b). Although now commonly used in many precision agriculture survey applications, most of the original interest in EC_a survey data was motivated by the need to characterize and map soil salinity in a cost-effective manner (Rhoades et al. 1999; Hendrickx et al. 2002). The need for such surveying work is expected to increase over time, as more agricultural land becomes degraded due to salinization.

Apparent soil conductivity survey data often correlate reasonably well with various soil properties (salinity, soil texture, soil water content, etc.) under different field conditions (Corwin and Lesch 2005b; Lesch and Corwin 2003). However, as a general rule, EC_a survey readings tend to be strongly correlated with soil salinity levels. Thus, accurate salinity predictions can normally be constructed from EC_a survey data in semi-saline and saline fields using fairly simple statistical calibration techniques. Additionally, accurate maps of the field-scale salinity pattern can sometimes also be produced from EC_a survey data in marginally saline fields, provided that other important soil properties (such as soil texture and soil water content) exhibit fairly minimal spatial variation.

After EC_a survey data have been acquired in a field, calibration soil samples are normally collected at a certain number of EC_a survey locations. The measured salinity levels associated with these soil samples are then used (in conjunction with the co-located survey data) to estimate some type of spatial-statistical or geostatistical model. This statistical model is in turn used to predict the detailed spatial soil-salinity pattern from the full set of acquired survey data.

This chapter discusses the simplest and most frequently used statistical modeling approach for calibrating EC_a survey information with measured salinity data, such as ordinary regression. Ordinary linear regression models represent a special case of a much more general class of models commonly known as linear regression models with spatially correlated errors (Schabenberger and Gotway 2005), hierarchical spatial models (Banerjee et al. 2004), or geostatistical mixed linear models (Haskard et al. 2007). This broader class of models includes many of the geostatistical techniques familiar to soil scientists, such as universal kriging and kriging with external drift, as well as standard regression techniques—ordinary linear regression (LR) models and analysis of covariance models.

The remainder of this chapter is organized as follows. A technical review of the basic linear regression model estimation and validation techniques is presented in "Regression Models" and "Regression Model Validation Tests." Some suitable sampling strategies for calibrating linear regression equations are discussed in "Sampling Strategies," while the subsequent section presents a brief overview of the *ESAP* software package. Two salinity assessment examples are then presented in "Data Analysis Examples"; these data analyses demonstrate the statistical calibration and prediction techniques discussed in this chapter, along with some of the types of analysis output that the *ESAP* software can produce.

Regression Models: Estimation and Prediction Formulas

Site-specific prediction of diverse soil properties from EM survey data can be achieved using regression model estimation and prediction techniques. In the regression modeling approach advocated by Lesch and Corwin (2008), Lesch (2005), Rhoades et al. (1999), and Lesch et al. (1995), a suitable linear equation is specified that relates the target soil property of interest to a linear combination of conductivity signal data readings and (possibly) trend surface coordinates. One example of such an equation would be

$$y_1 = \beta_0 + \beta_1[\mathrm{EM}_{V,i}] + \beta_2[EM_{H,i}] + \beta_3[c_{x,i}] + \beta_4[c_{y,i}] + \varepsilon_i \qquad (14\text{-}1)$$

where the response variable (y) represents the soil property of interest (e.g., salinity, texture, water content) at the ith survey location; the predic-

variables (EM_V, EM_H, c_x, c_y) represent the corresponding EM38 vertical and horizontal signal readings and associated ith survey site coordinate locations, respectively; the β parameters represent empirical regression model coefficients; and ε represents the random error component associated with the model. Equation 14-1 relates the response variable (e.g., soil property) to both EM signal and trend surface components, and thus can be viewed as a "signal + trend" model. The trend surface components specified in Eq. 14-1 are optional and should only be included if they are found to be necessary (i.e., if the associated parameter estimates are statistically significant or if the inclusion of such components is needed to address an obvious spatial trend in a residual plot).

The optimal estimation of the aforementioned (or similar) regression model depends on the assumptions placed on the random error components. If the errors are assumed to be normally distributed and exhibit spatial correlation, then Eq. 14-1 is commonly called a spatial linear regression model in the statistical literature, or a kriging with external drift model in the geostatistical literature (Cressie 1991; Schabenberger and Gotway 2005). Such models can be efficiently estimated using maximum likelihood or restricted maximum likelihood fitting techniques (Littell et al. 1996). In contrast, if the errors can be assumed to be approximately uncorrelated, then ordinary least squares (OLS) fitting techniques can be used. In this latter case the model becomes identical to an ordinary linear regression equation, the only difference being that the predictions are spatially referenced.

The likelihood of the residual errors being approximately uncorrelated (as opposed to spatially correlated) depends primarily on (1) the method used to select the calibration sample sites, and (2) the degree to which the conductivity signal data correlates with the response variable of interest. When the signal data is strongly correlated with the target soil property and specialized sampling strategies are employed, the assumption of approximate residual independence is often satisfied. For detailed discussions concerning these issues, see Lesch and Corwin (2008) and Lesch (2005).

Although appropriate prediction statistics can be derived for either case, only the OLS results are presented here. Additionally, all of the following results are presented in matrix notation; a good review of matrix notation from a regression modeling viewpoint is given in Myers (1986). Following standard matrix notation, note that we can express Eq. 14-1 as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where y represents a $(n \times 1)$ vector of soil property measurements (collected across n sites), $\mathbf{X}$ represents the corresponding $(n \times p)$ regression model design matrix and $\mathbf{e}$ represents the $(n \times 1)$ vector of residual errors. Then, under the uncorrelated residual error assumption, the best linear unbiased estimate (BLUE) for β is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (14\text{-}2)$$

with a corresponding variance of

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{14-3}$$

where σ^2 represents the regression model mean square error (MSE) component.

Likewise, the residuals (i.e., empirical model errors) for Eq. 14-1 are defined to be

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\beta} \tag{14-4}$$

and these residuals provide an unbiased estimate of the MSE component, that is

$$\hat{\sigma}^2 = (\mathbf{r}^T\mathbf{r})/(n-p).$$

Now, let $\mathbf{y}_z$ represent the (unknown) vector of soil property values at all of the remaining survey locations and define $\mathbf{X}_z$ to be the corresponding design matrix associates with these sites. Then, again under the uncorrelated residual error assumption, the best linear unbiased prediction (BLUP) of these soil property values can be shown to be

$$\hat{\mathbf{y}}_z = \mathbf{X}_z\hat{\beta} \tag{14-5}$$

with a corresponding variance estimate of

$$Var(\mathbf{y}_z - \hat{\mathbf{y}}_z) = \sigma^2(\mathbf{I} + \mathbf{X}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_z^T). \tag{14-6}$$

Corresponding formulas for both individual and field average prediction estimates can also be immediately derived from standard linear modeling theory. For example, individual survey site predictions (and their corresponding variance estimates) become

$$\begin{aligned} \hat{y}_0 &= \mathbf{x}_z\hat{\beta} \\ Var\{y_0 - \hat{y}_0\} &= \sigma^2(1 + \mathbf{x}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T) \end{aligned} \tag{14-7}$$

where $\mathbf{x}_z$ represents the $(1 \times p)$ design vector associated with a specific prediction site. Likewise, the average prediction associated with the entire nonsampled survey grid can be computed as

$$\hat{y}_{ave} = \mathbf{x}_{ave}\hat{\beta}$$
$$Var\{y_{ave} - \hat{y}_{ave}\} = \sigma^2(1/(N-n) + \mathbf{x}_{ave}(\mathbf{X}^T\mathbf{X}) \quad (14\text{-}8)$$

where $\mathbf{x}_{ave}$ represents the average of the $N - n$ design vectors associated with the nonsampled survey positions. Note that all of these results are exactly identical to ordinary linear regression model parameter estimation and prediction formulas presented in standard regression model textbooks (Myers 1986).

In many practical survey applications, determining the probability that a new prediction exceeds some specific threshold value is also of interest. Although not commonly discussed in most classical linear modeling textbooks, regression models can also be used to produce such probability estimates. More specifically, upon adopting a Bayesian perspective, the probability that an unobserved y_0 lies within the interval (a, b) can be computed as

$$\pi_i[a,b] = Prob(a \le y_0 \le b) = \int_g^h t_{(n-p)}dt \quad (14\text{-}9)$$

where $t_{(n-p)}$ represents a central t-distribution having $n - p$ degrees of freedom (i.e., the regression model residual degrees of freedom), $g = (a - \hat{y}_0)/\sqrt{Var\{\hat{y}_0\}}$, and $h = (b - \hat{y}_0)/\sqrt{Var\{\hat{y}_0\}}$ (Press 1989: assuming vague prior distributions on the model parameters). These latter probability predictions can in turn be used to calculate a range interval estimate (RIE) defined as

$$RIE[a,b] = \frac{100}{N-n}\sum_{i=1}^{N-n}\pi_i[a,b] \quad (14\text{-}10)$$

which represents a prediction of the percentage of nonsampled sites (on the survey grid) that exhibit soil property values falling within the interval (a, b). For example, one might be interested in predicting the percentage of survey sites in a field with salinity levels in excess of 4 dS/m. Equations 14-9 and 14-10 can be used to calculate this value, while simultaneously adjusting out the "shrinkage-effect" inherent in the associated regression model predictions. Lesch et al. (2005) discuss the above estimates in more detail and show multiple examples of their application.

Regression Model Validation Tests

If an ordinary linear regression model is to be successfully used in place of the geostatistical or spatial linear model, then more-restrictive modeling assumptions need to be met. In addition to the assumption of a normally distributed error process, the critical assumption in the linear regression model is the uncorrelated residual assumption. A formal test for spatial correlation in the residual pattern can be carried out using either a nested likelihood ratio test or via the Moran residual test statistic (Upton and Fingleton 1985; Haining 1990; Tiefelsdorf 2000; Schabenberger and Gotway 2005). The likelihood ratio test can only be performed after first estimating a suitable geostatistical or spatial linear model [see pages 343–344 of Schabenberger and Gotway (2005) for more discussion of this topic]. In contrast, the Moran test can be carried out directly on the ordinary regression model residuals.

As originally introduced by Brandsma and Ketellapper (1979), the Moran test statistic was designed to detect spatially correlated residuals in conditionally and/or simultaneously specified spatial autoregressive models (Schabenberger and Gotway 2005). However, it can also be used to assess the uncorrelated residual assumption in a general linear modeling framework. The Moran residual test statistic (δ_M) is defined as

$$\delta_M = \frac{\mathbf{r}^T \mathbf{W} \mathbf{r}}{\mathbf{r}^T \mathbf{r}} \tag{14-11}$$

where $\mathbf{r}$ is defined in Eq. 14-4, $\mathbf{W}$ represents a suitably specified proximity matrix, and $\hat{\beta}$ is calculated using Eq. 14-2. While the specification of W can be application-specific, in most soil survey applications it is generally reasonable to specify $\mathbf{W}$ as a scaled inverse distance squared matrix. Under such a specification, where d_{ij} represents the computed distance between the ith and jth sample locations, the $\{w_{ij}\}$ elements associated with the ith row of the $\mathbf{W}$ matrix are defined as

$$w_{ii} = 0 \quad \text{and} \quad w_{ij} = d_{ij}^2 \Big/ \sum_{j=1}^{n} d_{ij}^{-2}, \tag{14-12}$$

respectively.

Brandsma and Ketellapper (1979) describe how to calculate the first two moments of δ_M, i.e., $E(\delta_M)$ and $Var(\delta_M)$ [see also Lesch and Corwin (2008) and Lesch (2005)]. The corresponding Moran test score can then be computed as

$$z_M = (\delta_M - E(\delta_M)) / \sqrt{Var(\delta_M)} \tag{14-13}$$

compared to the upper (one-sided) cumulative standard normal probability density function. A test score in excess of 1.65 ($\alpha \approx 0.05$) is normally interpreted as being statistically significant. Provided that the regression model has been correctly specified, such a test score implies that the model residuals exhibit significant spatial correlation. In this situation, the parameter estimates and survey predictions may be highly inefficient and the mean square error estimate and parameter test statistics may be substantially biased. If sufficient data are available (or additional data can be collected), then a suitable spatial or geostatistical linear modeling approach should instead be employed.

In addition to the uncorrelated residual assumption, one must also verify that the model residuals satisfy the usual standard normal error assumption and that the hypothesized model is correctly specified. Fortunately, most well-known residual analysis techniques used in an ordinary regression analysis are just as useful when applied to a spatially referenced linear regression model. These include assessing the assumption of residual normality using quantile (QQ) plots and the Shapiro-Wilk test (Shapiro and Wilk 1965), detecting outliers and/or high leverage points (plots of internally or externally studentized residuals), and detecting model specification bias (residual versus prediction plots, partial regression leverage plots, influence plots, etc.).

The standard jack-knifing techniques commonly used to assess the predictive capability of an ordinary regression model are also directly applicable. Most standard statistical software packages can readily produce jack-knifed residual and/or prediction estimates in a computationally efficient manner. Cook and Weisberg (1999) and Myers (1986) offer good reviews of many relevant regression model diagnostic and assessment techniques.

Sampling Strategies for Spatially Referenced Linear Regression Models

Space limitations preclude a detailed discussion of the numerous sampling strategies one can employ to estimate spatially referenced regression models. Broadly speaking, the most common strategies currently employed can be classified as either (1) probability-based (design-based) sampling, (2) prediction-based (model-based) sampling, and (3) grid sampling. Brief descriptions of each of these approaches are given here.

In general, probability-based sampling strategies tend to be commonly employed in most spatial research problems. Probability-based sampling strategies include techniques, such as simple random sampling, stratified random sampling, cluster sampling, capture-recapture techniques, and line transect sampling. Thompson (1992) provides a good review of multiple types of probability-based sampling strategies.

Probability-based sampling strategies have a well-developed underlying theory and are clearly useful in many spatial applications (Thompson 1992; Brus and de Gruijter 1993). However, they are not designed specifically for estimating models per se. Indeed, most probability-based sampling strategies explicitly avoid incorporating any parametric modeling assumptions, relying instead upon randomization principles (which are built into the design) for drawing statistical inference.

Prediction-based sampling strategies represent an alternative approach for developing sampling designs that are explicitly focused toward model estimation. The underlying theory behind this approach for finite population sampling and inference is discussed in detail in Valliant et al. (2000). More generally, response surface design theory and optimal experimental design theory represent two closely related statistical research areas that also study sampling designs specifically from the model estimation viewpoint (Atkinson and Donev 1992; Myers and Montgomery 2002). Techniques from these latter two subject areas have been applied to the optimal collection of spatial data by Müller (2001), the specification of optimal designs for variogram estimation by Müller and Zimmerman (1999), the estimation of spatially referenced regression models by Lesch (2005) and Lesch et al. (1995), and the estimation of geostatistical linear models by Brus and Heuvelink (2007), Minasny et al. (2007), and Zhu and Stein (2006). Conceptually similar types of nonrandom sampling designs for variogram estimation have been introduced by Bogaert and Russo (1999), Warrick and Myers (1987), and Russo (1984).

Grid sampling represents another form of nonrandom sampling that has been used for many years in the soil sciences. Grid sampling has historically been recommended for accurately mapping soil boundaries and/or as a precursor to an ordinary kriging analysis (Burgess et al. 1981; Burgess and Webster 1984).

Theoretically, any of these sampling approaches can be used for the purposes of estimating a regression model, although each approach exhibits various strengths and weaknesses. Lesch (2005) compares and contrasts probability-based and prediction-based sampling strategies in more detail, and highlights some of the strengths of the prediction-based sampling approach.

Overview of the *ESAP* Software Package

Many types of diverse software programs can be utilized for the assessment and quantification of soil salinity inventory information via soil conductivity survey data. The more common types of software applications include spatial mapping software, GIS software, statistical software, and geophysical software (when appropriate). Nonetheless, the need for a stand-alone, comprehensive salinity assessment software package was recognized some years ago by the technical staff at the U.S. Salinity Labo-

...ware. The *ESAP* software package was specifically developed to handle ...field scale salinity inventorying and assessment work, primarily in ...response to this need (Lesch et al. 2000).

The *ESAP* software package contains a series of integrated, comprehensive software programs, designed for the Windows XP (or equivalent) operating system. This software can be used for the prediction of field-scale, spatial soil salinity information from conductivity survey data and has specifically been designed to facilitate the use of cost-effective, technically sound soil salinity assessment and data interpretation techniques. The current publicly available shareware version of *ESAP* (version 2.35) contains three data processing programs designed to guide the analyst through the entire survey process: *ESAP-RSSD*, *ESAP-SaltMapper*, and *ESAP-Calibrate*. The *ESAP-RSSD* program can be used to generate optimal model-based soil sampling designs from conductivity survey data. The *ESAP-SaltMapper* program may be used to generate 1-D transect plots and/or 2-D raster maps of either raw soil conductivity or predicted soil salinity data. Additionally, the *SaltMapper* software can be used to identify and locate tile line positions in fields suffering drainage-related salinity problems. The *ESAP-Calibrate* program is normally used to convert raw conductivity data into estimated soil salinity data, via either statistical or deterministic calibration modeling techniques. However, this latter program can also be used to estimate other soil properties from conductivity survey data and/or analyze various soil property/conductivity relationships.

ESAP version 2.35 contains two additional utility programs: *ESAP-SigDPA* and the *DPPC-Calculator*. The *SigDPA* program can be used to perform various conductivity data preprocessing chores, such as screening out negative conductivity readings and/or assigning row numbers to transect conductivity survey data. The *DPPC-Calculator* can be used to convert insertion four-probe readings into calculated soil salinity levels (in conjunction with measured or estimated soil temperature, texture, and water content information).

The *ESAP-RSSD* and *ESAP-Calibrate* programs contain the bulk of the model-based sampling and statistical modeling algorithms within the *ESAP* software package. As discussed, the *ESAP* software package represents an integrated, self-contained salinity assessment software system. All of the data analysis examples presented in the next section were performed using the version 2.35 *ESAP* software components (i.e., *RSSD*, *Calibrate*, and *SaltMapper*).

Data Analysis Examples

Example 1: A survey of a marginally saline lettuce field in Indio, California

An electromagnetic induction (EMI) survey was performed by the Coachella Valley Resource Conservation District in June 2003 within a

14-ha lettuce field located in Indio, California. The primary goals of this survey were threefold: (1) to construct an accurate soil salinity inventory for the field, (2) to determine if this field should be leached before the fall cropping season, and (3) to construct relevant yield-loss projections based on the predicted field soil salinity conditions. A total of 2,040 Geonics EM38 vertical (EM_V, mS/m) and horizontal (EM_H, mS/m) signal readings were collected across 29 north-south survey transects within this field and then processed through the USDA-ARS *ESAP* software package. This software selected 12 survey locations for soil sampling using a prediction-based *ESAP* sample design (Lesch et al. 2000). Soil samples were collected from 0 to 0.6 m and 0.6 m to 1.2 m depths and analyzed for soil salinity (EC_e, dS/m), soil saturation percentage (SP, %), and gravimetric water content (θ_g, %). Table 14-1 lists the univariate summary statistics (mean, standard deviation, minimum, and maximum) for the EM38 survey and soil sample data, respectively. Figure 14-1 shows the interpolated EM_V signal map for this field, along with the spatial positions of the 12 sampling locations. Note also that some advanced statistical aspects concerning this specific data analysis are discussed in Lesch and Corwin (2008).

The results from an exploratory regression modeling analysis performed in *ESAP* suggested that the following natural log(EC_e)/log(EM

TABLE 14-1. Basic EM38 and Soil Sample Summary Statistics: Indio Lettuce Field

Variable	Units	N	Mean	Std. Dev.	Min	Max
EM_V	mS/m	2040	63.67	13.87	36.25	119.25
EM_H	mS/m	2040	38.02	10.28	17.63	81.75

Variable	Units	Depth	N	Mean	Std. Dev.	Min	Max
EC_e	dS/m	0–0.6 m	20	1.86	1.18	0.72	4.22
		0.6 m–1.2 m	20	1.93	1.28	0.26	3.92
SP	%	0–0.6 m	20	36.95	4.09	32.20	45.57
		0.6 m–1.2 m	20	32.92	5.14	26.35	44.10
θ_g	%	0–0.6 m	20	16.76	3.41	9.85	21.10
		0.6 m–1.2 m	20	16.64	5.33	10.60	25.20

EC_e = soil salinity
EM_H = EM38 horizontal signal
EM_V = EM38 vertical signal
SP = soil saturation percentage
θ_g = gravimetric water content

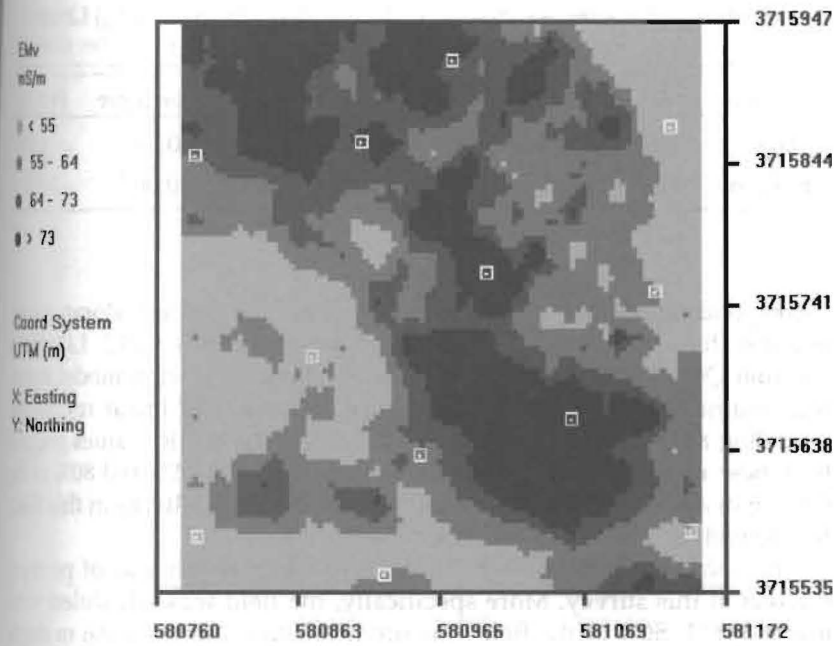

FIGURE 14-1. Survey of a marginally saline lettuce field in Indio, California, showing the interpolated EM_V signal map for this field, along with the spatial positions of the 12 sampling locations.

regression equation should be used to describe the soil salinity/signal conductivity relationship in this field:

$$\ln(EC_{ij}) = \beta_{0j} + \beta_{1j}(z_{1i}) + \beta_{2j}(z_{2i}) + \beta_{3j}(z^2_{1i}) + \varepsilon_{ij} \qquad (14\text{-}14)$$

where

$$z_{1i} = \ln(EM_{V,i}) + \ln(EM_{H,i}), \text{ and}$$
$$z_{2i} = \ln(EM_{V,i}) - \ln(EM_{H,i}) \qquad (14\text{-}15)$$

In Eq. 14-14, the subscript $j = 1, 2$ corresponds to the two sampling depths, $i = 1, 2, \ldots 2{,}040$ corresponds to the EM38 sampling locations, β_{0j} through β_{4j} represent the two sets of regression model parameters (which define the two depth-specific prediction functions), and the residual errors for each sampling depth are assumed to be spatially uncorrelated. Table 14-2 presents the key summary statistics for each estimated regression function; these statistics include the R^2, root mean square error (RMSE) estimate, overall model F-score and associated p-value, and the

TABLE 14-2. Summary Statistics for Depth-Specific ln(EC_e) Linear Regression Models: Indio Lettuce Field

Depth	R^2	RMSE	F-score	Pr > F	Moran Score	Pr > Z_M
0–0.6 m	0.922	0.196	31.37	<0.001	0.652	0.25
0.6–1.2 m	0.798	0.490	10.54	0.004	−0.067	>0.5

corresponding Moran test score and p-value. These latter Moran scores suggest that the uncorrelated residual assumption is valid. Likewise, residual QQ plots (not shown) confirm that the regression model errors follow a normal distribution and, hence, the ordinary linear regression modeling approach can be adopted. Additionally, the R^2 values suggest that these regression models can be used to describe 92% and 80% of the 0 to 0.6 m and 0.6 to 1.2 m observed spatial log(EC_e) patterns in this field, respectively.

The spatial salinity pattern in the 0 to 0.6 m depth was of primary interest in this survey. More specifically, the field was scheduled to be leached if (1) 50% of the field was predicted to exhibit 0 to 0.6 m depth salinity levels >2 dS/m and/or the field average ln(EC_e) level exceeded ln(2) = 0.693, or (2) 25% of the field was predicted to exhibit 0 to 0.6 m depth salinity levels >3 dS/m. Table 14-3 presents the predicted field average ln(EC_e) levels (and corresponding 95% confidence intervals), as well as the range interval estimates for both sampling depths. These predictions can be automatically calculated in the *ESAP* software package (using Eqs. 14-8 through 14-10, respectively). Figure 14-2 shows the corre-

TABLE 14-3. Regression Model Predicted Field Average ln(EC_e) Levels and Range Interval Estimates: Indio Lettuce Field

	0–0.6 m Depth	0.6–1.2 m Depth
Field average ln(EC_e)	0.494	0.548
95% confidence interval	(0.35, 0.64)	(0.19, 0.91)
Range Interval Estimates (% Area of Field Classified into RIEs)		
<2.0 dS/m	66.5	54.8
2.0–3.0 dS/m	22.9	19.9
3.0–6.0 dS/m	10.1	20.0
>6.0 dS/m	0.5	5.3

RIE = range interval estimate

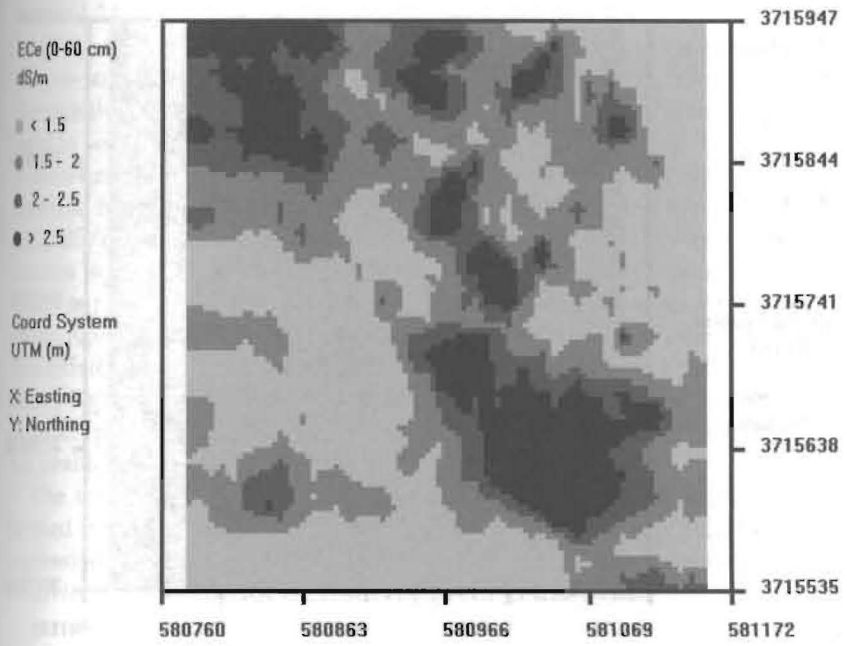

FIGURE 14-2. The corresponding predicted spatial salinity map for the field in Fig. 14-1. This map was produced within the ESAP SaltMapper *program by interpolating the back-transformed, individual ln(EC_e) predictions onto a fine-mesh grid using an adjustable smoothing kernel.*

sponding predicted spatial salinity map for this field. This map was produced (within the *ESAP SaltMapper* program) by interpolating the back-transformed, individual ln(EC_e) predictions onto a fine-mesh grid using an adjustable smoothing kernel.

The results shown in Table 14-3 and Fig. 14-2 suggest that this field does not need to be leached. The 0 to 0.6 m field average ln(EC_e) estimate is 0.494, and 66.5% of the individual 0 to 0.6 m depth predictions are calculated to be below 2 dS/m. Additionally, only 10.6% of these predictions are calculated to exceed 3 dS/m. Thus, none of the specified criteria for implementing a leaching process are met in this field.

Within the preceding 5 years, the landowner had grown alternating winter vegetable crops of romaine lettuce and broccoli in this field. The *ESAP* software can be used to convert the Fig. 14-2 salinity map into projected relative yield loss maps for these crops, using standard salt-tolerance equations published for these vegetables (Mass and Hoffman 1977). Figure 14-3 shows the projected relative yield loss map for romaine lettuce, based on a threshold of 1.3 dS/m, a slope estimate of 13% yield loss

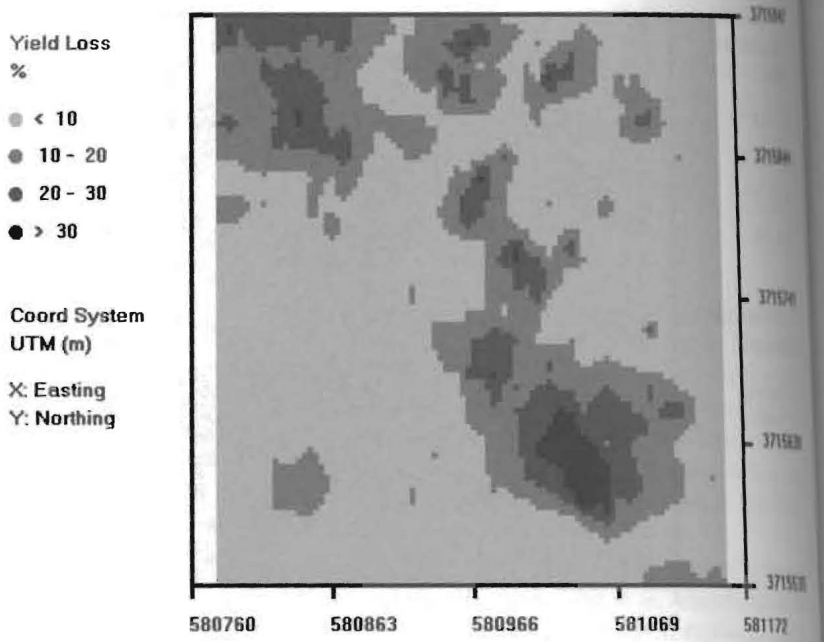

FIGURE 14-3. For the field shown in Figs. 14-1 and 14-2, the projected relative yield loss map for romaine lettuce, based on a threshold of 1.3 dS/m, a slope estimate of 13% yield loss per one unit increase in EC_e (beyond the threshold), and a 80:20 root-weighting distribution (for the 0-m to 0.6-m and 0.6-m to 1.2-m depths, respectively).

per one unit increase in EC_e (beyond the threshold), and a 80% to 20% root-weighting distribution (for the 0 to 0.6-m and 0.6- to 1.2-m depths, respectively). The calculated field average romaine lettuce yield loss in this field is 8.7%. The corresponding calculated field average yield loss for broccoli is <1% (using a threshold of 2.8, a slope of 9.2%, and a 70% to 30% root-weighting distribution). These additional yield loss estimates also suggest that a full-scale leaching of this field is currently unwarranted, particularly if broccoli is the next scheduled crop in the rotation.

Example 2: Pre- and postleaching surveys of a Coachella Valley vegetable field

Pre- and postleaching EM surveys were performed by U.S. Salinity Laboratory personnel in July and October 2003 within a 13-ha vegetable field located in Thermal, California. The main goal of this survey was to spatially quantify the leaching process and determine the percent reduction in the post- versus preleaching median salinity levels in the field. A

total of 1,243 and 1,288 Geonics EM38 vertical (EM_V, mS/m) and horizontal (EM_H, mS/m) signal readings were collected within this field during the pre- and postleaching survey processes, respectively, and processed through the USDA-ARS *ESAP* software package. This software was again used to select 12 locations for soil sampling in each survey, using a prediction-based *ESAP* sample design (Lesch et al. 2000). Soil samples were collected from the 0 to 0.6 m sample depth and analyzed for soil salinity (EC_e, dS/m), soil saturation percentage (SP, %), and gravimetric water content (θ_g, %). Table 14-4 lists the univariate summary statistics for the EM38 survey and 0 to 0.6 m sample data associated with each survey event. Note that one soil sample in the preleaching survey event had to be discarded due to contamination during the laboratory analysis procedures. Figures 14-4 and 14-5 show the interpolated July (preleaching) and October (postleaching) EM_H signal maps for this field, along with the spatial positions of the sampling locations.

The results from an exploratory regression modeling analysis performed in *ESAP* confirmed that the following simple log(EC_e)/log(EM) regression equation could be used to describe the soil salinity/signal conductivity relationship for each survey event in this field:

$$\ln(EC_{ij}) = \beta_{0j} + \beta_{1j}(z_{1ij}) + \varepsilon_{ij} \qquad (14\text{-}16)$$

TABLE 14-4. Basic EM38 and Soil Sample Summary Statistics: Coachella Valley Vegetable Field[a]

Variable	Units	Date	N	Mean	Std.Dev.	Min	Max
EM_H	mS/m	July	1243	23.25	9.12	10.63	79.75
EM_V	mS/m	July	1243	44.35	13.29	27.25	124.63
EM_H	mS/m	October	1288	30.99	13.10	15.25	121.88
EM_V	mS/m	October	1288	48.26	18.69	27.75	175.38
EC_e	dS/m	July	11	1.83	0.99	0.75	3.69
SP	%	July	11	32.53	2.36	29.44	37.33
θ_g	%	July	11	0.12	0.03	0.06	0.16
EC_e	dS/m	October	12	0.98	0.39	0.63	1.94
SP	%	October	12	34.07	5.88	28.63	46.33
θ_g	%	October	12	0.24	0.10	0.11	0.44

[a]All soil samples acquired from 0–0.6 m sampling depth
EC_e = soil salinity
EM_H = EM38 horizontal signal
EM_V = EM38 vertical signal
SP = soil saturation percentage
θ_g = gravimetric water content

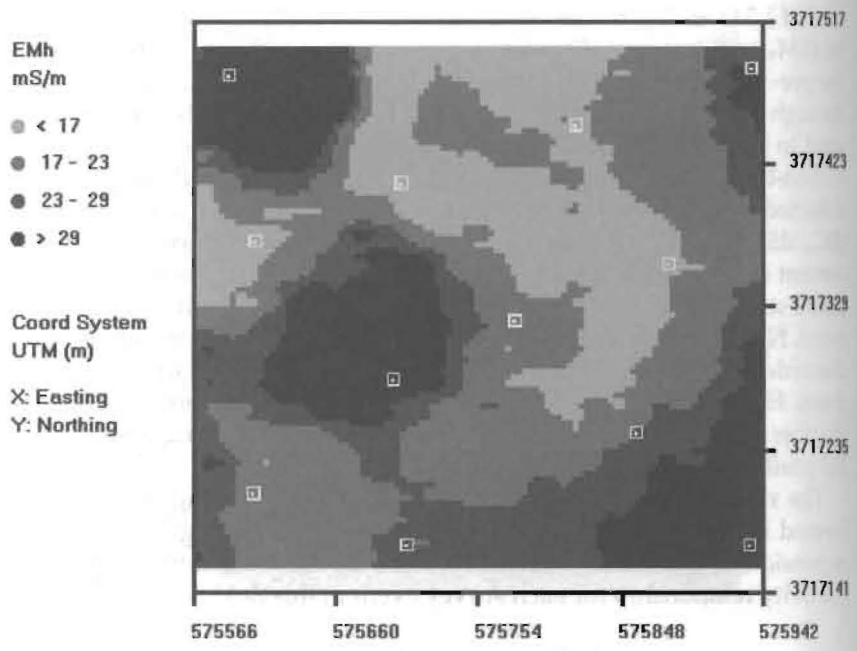

FIGURE 14-4. The interpolated July (preleaching) EM_H signal maps of a Coachella Valley, California, vegetable field, along with the spatial positions of the sampling locations.

where $z_{1ij} = \ln(EM_{V,i,j}) + \ln(EM_{H,i,j})$. In Eq. 14-16, the subscript $j = 1, 2$ now corresponds to the two sampling dates, the i subscript correspond to the EM38 sampling locations acquired during each survey process, $\{\beta_{01}, \beta_{11}\}$ and $\{\beta_{02}, \beta_{12}\}$represent the two sets of regression model parameters (which define the two time-dependent prediction functions), and the residual errors for each sampling depth are again assumed to be spatially uncorrelated. Table 14-5 presents the key summary statistics for each estimated regression function; these statistics again include the R^2, root mean square error (RMSE) estimate, overall model F-score and associated p-value, and the corresponding Moran test score and p-value. The Moran scores and residual QQ plots (not shown) suggest that the normally distributed, uncorrelated residual assumption is valid. The RMSE and R^2 values suggest that the postleaching LR model is more accurate; this increase in prediction accuracy is most likely due to the presence of higher and more uniform soil moisture conditions during the post-leaching survey process.

In September 2003, a total of 64 cm of Colorado River water was applied to this field over a seven-day leaching cycle. The leaching was performed using 25 m-wide ponding basins laid out across the field,

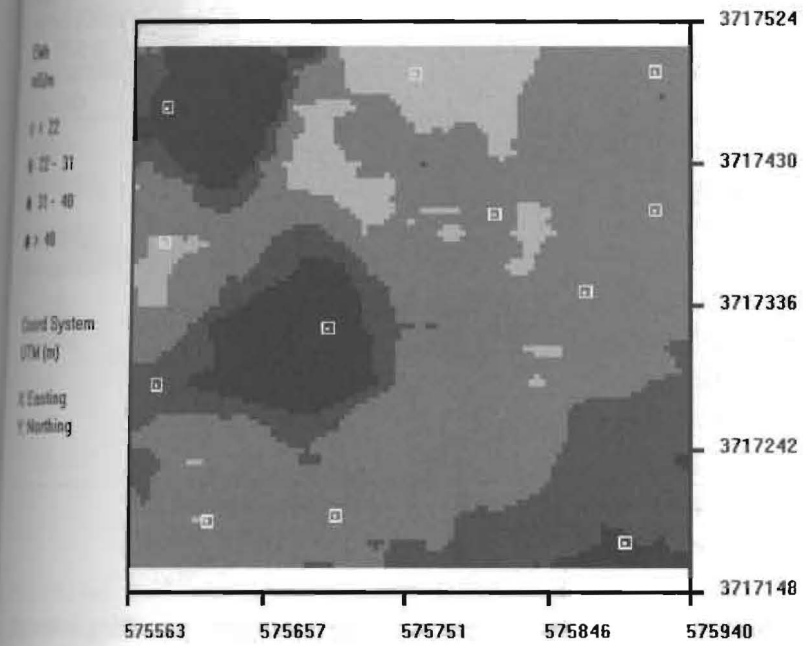

FIGURE 14-5. *The same field as in Fig. 14-4, in October (postleaching).*

after the soil had been deep-chiseled, plowed, and land-planed. The basins were laser-leveled and the water was released from a standpipe located within the northwest corner of the field (a head channel along the north edge of the field was used to deliver the water to each basin). Calculations from the flow and volume measurements performed during the leaching process suggested that approximately 55 cm of water infiltrated the soil and that the distribution uniformity of the basin system was 93%.

The temporal change in the spatial salinity pattern in the 0 to 0.6 m depth was of primary interest in this survey. Table 14-6 shows the *ESAP*-predicted pre- and postleaching salinity summary statistics for this field. The postleaching median salinity level is estimated to be 0.91 dS/m,

TABLE 14-5. Summary Statistics for Time-Specific $\ln(EC_e)$ Linear Regression Models: Coachella Valley Vegetable Field

Date	R^2	RMSE	F-Score	Pr > F	Moran Score	Pr > Z_M
July	0.600	0.340	13.51	0.005	−1.37	>0.5
October	0.837	0.148	51.37	>0.001	−0.55	>0.5

TABLE 14-6. Regression Model Predicted Field Average ln(EC_e) Levels and Range Interval Estimates: Coachella Valley Vegetable Field

	July	October
Field average ln(EC_e)	0.513	−0.098
95% confidence interval	(0.26, 0.76)	(−0.19, 0.00)
Range Interval Estimates (% Area of Field Classified into RIEs)		
<1.0 dS/m	16.6	68.0
1.0–1.5 dS/m	27.2	25.5
1.5–2.0 dS/m	21.4	4.6
2.0–3.0 dS/m	20.9	1.7
<3.0 dS/m	13.9	0.2

RIE = range interval estimate

which represents about a 46% decrease over the pre-leaching level (1.67). The *ESAP-Calibrate* software can perform a *t*-test on the difference between two field median (log mean) estimates; the corresponding *t*-score is this example is –5.14 ($p < 0.0001$). Additionally, 68% of the field is estimated to exhibit postleaching salinity levels below 1 dS/m, and less than 2% of the field exceeds 2 dS/m. These estimates imply a substantial leaching effect, given that the corresponding preleaching estimates were 16.7% (<1 dS/m) and 34.8% (>2 dS/m), respectively.

The predicted pre- and postleaching 0 to 0.6 m salinity maps for this field are shown in Figs. 14-6 and 14-7. A pronounced leaching effect can be clearly seen in the postleaching salinity map, and the near-surface salinity levels across the entire field appear to be significantly reduced. These results are perhaps not that surprising, given the large volume of water used during the leaching process (≈ 8.3 ha-m).

Finally, it is worthwhile to observe that the raw October (postleaching) EM38 signal data exhibited a higher average level than the July (preleaching) data (see Table 14-4 and Figs. 14-4 and 14-5). The general increase in the EM signal response was again most likely due to the elevated near-surface soil moisture conditions. The top 30 cm of the soil profile was particularly dry during the July survey; these dry surface conditions undoubtedly depressed the EM38 signal response. These results demonstrate why a direct interpretation of EM38 signal data is often misleading. Note that the median near-surface soil salinity level in this field decreased by nearly 46%, even though the average horizontal EM signal reading increased from 23.3 mS/m to 31.0 mS/m.

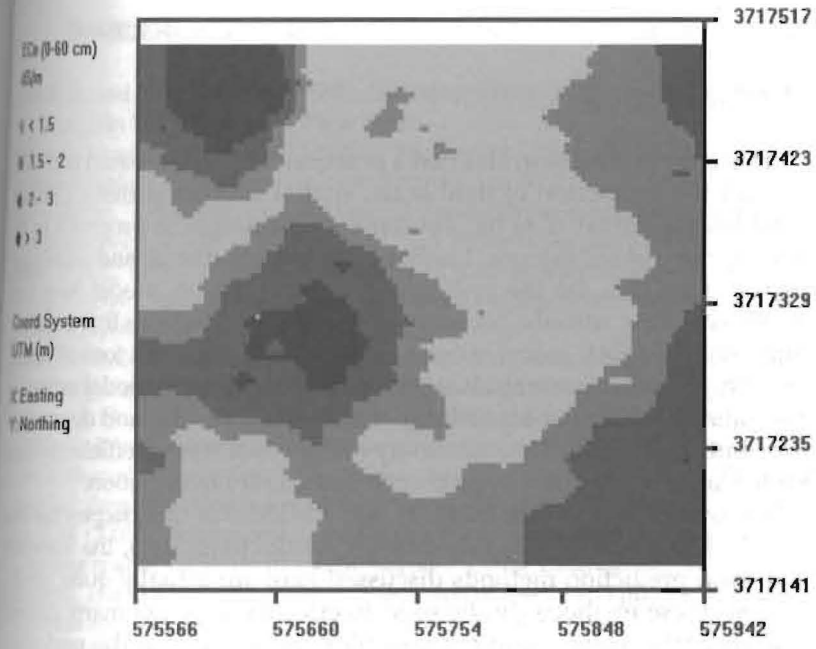

FIGURE 14-6. *The predicted preleaching 0- to 0.6-m salinity map for the field shown in Figs. 14-4 and 14-5.*

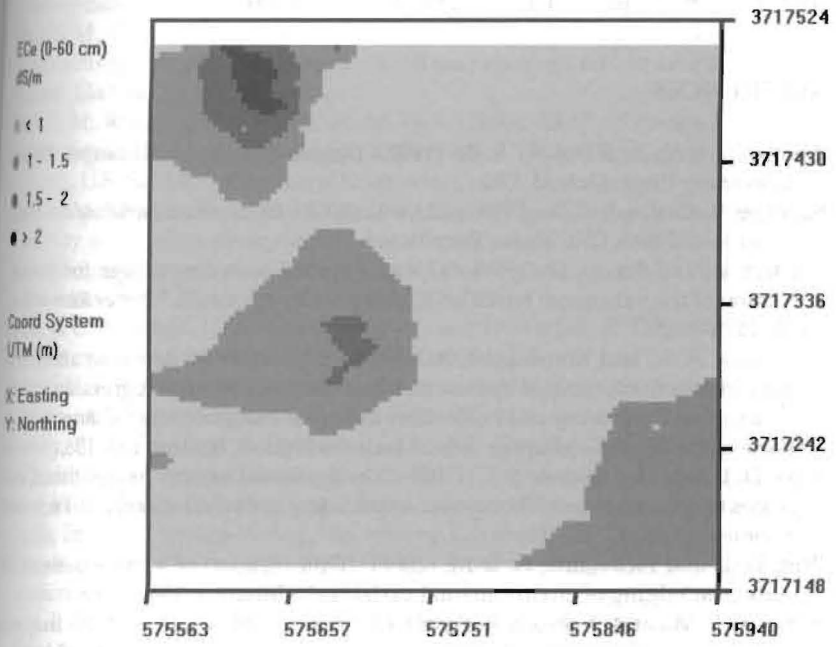

FIGURE 14-7. *The predicted postleaching 0- to 0.6-m salinity map for the field shown in Fig. 14-6. A pronounced leaching effect can be clearly seen here and the near-surface salinity levels across the entire field appear to be significantly reduced.*

SUMMARY

This chapter demonstrates that a practical, regression-based methodology for the prediction of field-scale, spatial salinity patterns from soil conductivity survey data has substantial advantages in programs for the management of soil salinity. The basic parameter estimate and salinity prediction formulas for the ordinary linear regression model have been reviewed, along with the necessary modeling assumptions that have been built into the *ESAP* model, which also provides guidance for soil salinity sampling. The two case studies presented highlight the model estimation and salinity prediction capabilities of the *ESAP* software and demonstrate how bulk soil electrical conductivity survey data can be efficiently interpreted and used to quantify field-scale soil salinity information.

It is worthwhile to note that although the focus of this chapter has been on predicting soil salinity from survey conductivity data, the associated statistical prediction methods discussed here are actually quite general. Indeed, these methods can be used to effectively model many different soil property/sensor data relationships, provided that the underlying modeling assumptions are satisfied. For a review of these more general calibration techniques, see Lesch and Corwin (2003) and/or the references contained in Table 1 of Corwin and Lesch (2005a).

REFERENCES

Atkinson, A. C., and Donev, A. N. (1992). *Optimum experimental designs*, Oxford University Press, Oxford, UK.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*, CRC Press, Boca Raton, Fla.

Bogaert, P., and Russo, D. (1999). "Optimal spatial sampling design for the estimation of the variogram based on a least squares approach." *Water Resour. Res.*, 35, 1275–1289.

Brandsma, A. S., and Ketellapper, R. H. (1979). "Further evidence on alternative procedures for testing of spatial autocorrelation amongst regression disturbances," in *Exploratory and explanatory statistical analysis of spatial data*, C. P. A. Bartels and R. H. Ketellapper, eds., Martinus Nijhoff, Boston, 113–136.

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

Brus, D. J., and Heuvelink, G. B. M. (2007). "Optimization of sample patterns for universal kriging of environmental variables." *Geoderma*, 138, 86–95.

Burgess, T. M., and Webster, R. (1984). Optimal sampling strategy for mapping soil types. I. Distribution of boundary spacings." *J. Soil Sci.*, 35, 641–654.

Burgess, T. M., Webster, R., and McBratney, A. B. (1981). "Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy." *J. Soil Sci.*, 32, 643–654.

Cook, R. D., and Weisberg, S. (1999). *Applied regression including computing and graphics*, John Wiley and Sons, New York.

Corwin, D. L., and Lesch, S. M. (2005a). "Apparent soil electrical conductivity measurements in agriculture." *Comp. Electron. Ag.*, 46, 11–43.

——— (2005b). "Characterizing soil spatial variability with apparent soil electrical conductivity. I. Survey protocols." *Comp. Electron. Ag.*, 46, 103–134.

Cressie, N. A. C. (1991). *Statistics for spatial data*, John Wiley and Sons, New York.

Ebdon, R. (1990). *Spatial data analysis in the social and environmental sciences*, Cambridge University Press, Cambridge, UK.

Haskard, K. A., Cullis, B. R., and Verbyla, A. P. (2007). "Anisotropic Matèrn correlation and spatial prediction using REML." *J. Ag. Bio. Environ. Stats.*, 12, 147–160.

Hendrickx, J. M. H., Das, B., Corwin, D. L., Wraith, J. M., and Kachanoski, R. G. (2002). "Indirect measurement of solute concentration," in *Methods of soil analysis, Part 4: Physical methods*. Soil Science Society of America Book Series, J. H. Dane and G. C. Topp, eds., Soil Science Society of America, Madison, Wisc., 1274–1306.

Lesch, S. M. (2005). "Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties." *Comp. Electron. Ag.*, 46, 153–179.

Lesch, S. M., and Corwin, D. L. (2008). "Prediction of spatial soil property information from ancillary sensor data using ordinary linear regression: Model derivations, residual assumptions and model validation tests." *Geoderma* 148, 130–140.

Lesch, S. M., and Corwin, D. L. (2003). "Using the dual-pathway parallel conductance model to determine how different soil properties influence conductivity survey data." *Agron. J.*, 95, 365–379.

Lesch, S. M., Corwin, D. L., and Robinson, D. A. (2005). "Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils." *Comp. Electron. Ag.*, 46, 351–378.

Lesch, S. M., Rhoades, J. D., and Corwin, D. L. (2000). *ESAP-95 version 2.10R: User manual and tutorial guide*, Research Report 146, USDA-ARS, George E. Brown, Jr., ed., U.S. Salinity Laboratory, Riverside, Calif.

Lesch, S. M., Strauss, D. J., and Rhoades, J. D. (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." *Water Resour. Res.*, 31, 387–398.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS system for mixed models*, SAS Institute Inc., Cary, N.C.

Maas, E. V., and Hoffman, G. J. (1977). "Crop salt tolerance: Current assessment." *J. Irrig. and Drainage Div., ASCE*, (103 IR2), 115–134.

Minasny, B., McBratney, A. B., Walvoort, D. J. J. (2007). "The variance quadtree algorithm: Use for spatial sampling design." *Comput. Geosci.* 33, 383–392.

Müller, W. G. (2001). *Collecting spatial data: Optimum design of experiments for random fields*, 2nd ed., Physica-Verlag, Heidelberg, Germany.

Müller, W. G, and Zimmerman, D. L. (1999). "Optimal designs for variogram estimation." *Environmetrics*, 10, 23–37.

Myers, R. H. (1986). *Classical and modern regression with applications*, Duxbury Press, Boston.

Myers, R. H., and Montgomery, D. C. (2002). *Response surface methodology: Process and product optimization using designed experiments*, 2nd ed., John Wiley and Sons, New York.

Press, S. J. (1989). *Bayesian statistics: Principles, models, and applications*, John Wiley and Sons, New York.

Rhoades, J. D., Chanduvi, F., and Lesch, S. M. (1999). *Soil salinity assessment: Methods and interpretation of electrical conductivity measurements*, FAO Irrigation and Drainage Paper No. 57, Food and Agriculture Organisation of the United Nations, Rome.

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

Schabenberger, O., and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*, CRC Press, Boca Raton, Fla.

Shapiro, S. S., and Wilk, M. B. (1965). "An analysis of variance tests for normality (complete samples)." *Biometrika*, 52, 591–611.

Thompson, S. K. (1992). *Sampling*, John Wiley and Sons, New York.

Tiefelsdorf, M. (2000). *Modeling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I*, Springer-Verlag, New York.

Upton, G., and Fingleton, B. (1985). *Spatial data analysis by example*, John Wiley and Sons, New York.

Valliant, R., Dorfman, A. H., and Royall, R .M. (2000). *Finite population sampling and inference: A prediction approach*, John Wiley and Sons, New York.

Warrick, A. W., and Myers, D. E. (1987). "Optimization of sampling locations for variogram calculations." *Water Resour. Res.*, 23, 496–500.

Zhu, Z., and Stein, M. L. (2006). "Spatial sampling design for prediction with estimated parameters." *J. Ag. Bio. Environ. Stats.*, 11, 24–44.

NOTATION

BLUE = best linear unbiased estimate
EC_e = soil salinity
EM_H = EM38 horizontal signal
EMI = electromagnetic induction
EM_V = EM38 vertical signal
$\mathbf{e}$ = $(n \times 1)$ vector of residual errors
RIE = range interval estimate
SP = soil saturation percentage
$\mathbf{X}$ = $(n \times p)$ regression model design matrix
$\mathbf{Y}$ = $(n \times 1)$ vector of soil property measurements
β = $(p \times 1)$ parameter vector
δ_M = Moran residual test statistic
θ_g = gravimetric water content