Geoderma 148 (2008) 130–140

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Prediction of spatial soil property information from ancillary sensor data using ordinary linear regression: Model derivations, residual assumptions and model validation tests

S.M. Lesch [a,*], D.L. Corwin [b]

[a] Principal Statistician, Statistical Consulting Collaboratory, U.C. Riverside, 2683 Stat-Comp, 900 University Ave, Riverside, CA 92521, USA
[b] Soil Scientist, USDA-ARS, United States Salinity Laboratory, 450 W. Big Springs Rd., Riverside, CA 92507, USA

ARTICLE INFO

Article history:
Received 20 May 2008
Received in revised form 6 September 2008
Accepted 16 September 2008
Available online 25 October 2008

Keywords:
Linear models
Geostatistics
BLU estimation
BLU prediction
Soil salinity

ABSTRACT

Geospatial measurements of ancillary sensor data, such as bulk soil electrical conductivity or remotely sensed imagery data, are commonly used to characterize spatial variation in soil or crop properties. Geostatistical techniques like kriging with external drift or regression kriging are often used to calibrate geospatial sensor data to specific soil or crop properties. More traditional statistical methods such as ordinary linear regression models are also commonly used. Unfortunately, some soil scientists see these as competing and unrelated modeling approaches and are unaware of their relationship. In this article we review the connection between the ordinary linear regression model and the more comprehensive geostatistical mixed linear model and describe when and under what conditions ordinary linear regression models represent valid spatial prediction models. The formulas for the ordinary linear regression model parameter estimates and best linear unbiased predictions are derived from the geostatistical mixed linear model under two different residual error assumptions; i.e., strictly uncorrelated (SU) residuals and effectively uncorrelated (EU) residuals. The theoretically optimal (best linear unbiased) and computable (linear unbiased) predictions and variance estimates derived under the EU error assumption are examined in detail. Statistical tests for detecting spatial correlation in LR model residuals are also reviewed, in addition to three LR model validation tests derived from classical linear modeling theory. Two case studies are presented that highlight and demonstrate the various parameter estimation, response variable prediction and model validation techniques discussed in this article.

1. 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, 2005). 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 (Hendrickx et al., 2002; Rhoades et al., 1999). The need for such surveying work is expected to increase over time, as more agricultural land becomes degraded due to salinization.

Soil conductivity survey data represents just one type of ancillary sensor data that is commonly collected in order to help identify, quantify, and/or predict various soil or crop properties. Being spatial in nature (i.e., referenced across a spatial domain), it is both natural and often quite reasonable to consider some type of geostatistical modeling technique when attempting to calibrate such survey data to a specific (soil or crop) response variable. Numerous examples exist in the literature of geostatistical or spatial modeling approaches; the textbooks by Schabenberger and Gotway (2005), Schabenberger and Pierce (2002), Webster and Oliver (2001), Wackernagel (1998) and Isaaks and Srivastava (1989) are particularly relevant to the above mentioned calibration problem.

In addition to the commonly used geostatistical techniques like kriging with external drift or regression–kriging, ordinary linear regression models are also often employed when calibrating such data. Most statisticians and pedomatricians understand the inter-relationships between these two modeling approaches, or more specifically that ordinary regression models actually represent a special case of the external drift or regression–kriging approaches. However, it has been our experience that many soil scientists do not fully appreciate these relationships and in certain cases see these as either competing or unrelated modeling techniques. Additionally, we have personally encountered more than a few cases where a scientist professed a strong belief in one of these two particular modeling approaches (either for reasons of apparent statistical simplicity or perceived statistical rigor), irrespective of the appropriate modeling approach suggested by the actual data analysis.

* Corresponding author.
E-mail address: scott.lesch@ucr.edu (S.M. Lesch).

0016-7061/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2008.09.014

In the mainstream statistical literature, it is well known that 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 pedomatricians, such as universal kriging, kriging with external drift and/or regression–kriging, as well as standard statistical techniques like ordinary linear regression (LR) and analysis of covariance (ANOCOVA) models.

The primary objective of this manuscript is to describe when and under what conditions (i.e., what set of assumptions) ordinary regression models represent valid linear spatial prediction models. To facilitate this discussion, we first present a review of the geostatistical mixed linear modeling approach. (Note that linear geostatistical models can typically be recast as "mixed" linear models, because the stochastic component usually contains more than one error term.) This review includes a brief discussion about the geostatistical mixed linear model (MLM) residual distribution assumptions, best linear unbiased estimates, and best linear unbiased predictions. The connection between an ordinary LR model and the geostatistical MLM is highlighted, specifically with respect to how the LR model can be derived from the geostatistical MLM under one of two more restrictive sets of residual distribution assumptions (i.e., the assumption of either strictly uncorrelated residuals or effectively uncorrelated residuals). Special emphasis is directed towards quantifying the practical effects that these residual distribution assumptions have on LR model parameter estimates and survey predictions. We also review some statistical tests that can be used to detect spatial correlation in LR model residuals (particularly the Moran test), and discuss three useful LR model validation tests that arise from classical linear modeling theory. Finally, two salinity survey case studies are presented that highlight and demonstrate the various model estimation, validation, and response variable prediction techniques discussed in this manuscript.

As alluded to above, we will argue that ordinary LR models can represent useful and valid (although not always technically optimal) spatial prediction models. However, a few caveats are in order. First, in the discussion that follows, we will assume that a sufficiently dense spatial grid of ancillary sensor data has been collected across the field or area of interest, such that the analyst need not worry about generating predictions "off-the-grid". Hence, linear prediction techniques such as cokriging or multi-stage hierarchical spatial models are not considered here. Second, we assume that all calibration soil samples are co-located with specific, identifiable survey locations. Third, while the modeling techniques discussed here can be extended to non-Gaussian (i.e., non-multivariate normal) error assumptions (Gotway and Stroup, 1997) and/or nonlinear models (Schabenberger and Pierce, 2002; Schabenberger and Gotway, 2005), we will limit the current discussion to linear Gaussian error models. Forth, we have chosen to use the geostatistical mixed linear modeling framework to motivate this discussion (rather than more traditional kriging derivations) because we believe that (i) the direct connection between the various linear prediction approaches is much easier to understand from this viewpoint, and (ii) the geostatistical mixed linear model can be defined to be equivalent to many (although certainly not all) commonly used kriging techniques.

Finally, we also wish to emphasize that the topic of optimal sampling strategies for estimating spatial linear prediction models is not discussed in any significant detail here. Both design-based (probabilistic) and model-based sampling strategies can be employed to estimate such models. Design-based sampling strategies have a well developed underlying theory and can be clearly useful in many spatial applications (Thompson, 1992; Brus and de Gruijter, 1993). Likewise, model-based sampling strategies 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), Warrick and Myers (1987), and Russo (1984); the estimation of spatially referenced LR 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). However, in this article we will assume that some type of valid sampling strategy already exists and instead focus primarily on model calibration and prediction issues.

2. Spatial linear models

2.1. The geostatistical mixed linear model

In a typical field survey where some type of ancillary sensor readings are collected, the general goal is to use this sensor data to help predict a specific, unobserved soil property. As just one example, assume that we have acquired a dense grid of electrical conductivity survey data across a particular field, collected soil samples at some of these survey locations, and then wish to use these sensor and calibration sample readings to estimate a model (that can in turn be used to predict the detailed spatial pattern of the soil property). Define the relationship between the soil property measurement, y, and sensor data, q, to be:

$$y = g(\mathbf{q}) + \xi \tag{1}$$

where $g(\mathbf{q})$ represents some unknown function of the vector of k-collocated sensor readings and corresponding survey locations, and ξ represents some type of spatial random error component. Now assume that Eq. (1) can be adequately approximated using the following geostatistical mixed linear model (Haskard et al., 2007):

$$\mathbf{y} = \mathbf{X}\beta + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}) \tag{2}$$

where y represents an $(n\times 1)$ vector of observed soil property data, s represents the corresponding vector of paired (s_x, s_y) survey location coordinates, $\mathbf{X}$ represents an $(n\times p)$ fixed data matrix that includes observed functions of sensor readings and possibly also the survey location coordinates, β represents a $(p\times 1)$ vector of unknown parameter estimates, $\eta(\mathbf{s})$ represents a 0-mean, second order stationary spatial Gaussian error process, and $\varepsilon(\mathbf{s})$ represents a vector of jointly independent Normal$(0, \sigma_n^2)$ random variables. Typical stationary spatial structures for $\eta(\mathbf{s})$ are well documented in the spatial-statistical and geostatistical literature; examples in 2-dimensions include the isotropic and anisotropic exponential and spherical covariance structures, as well as the Matérn class of covariance functions (Haining, 1990; Cressie, 1991; Wackernagel, 1998; Webster and Oliver, 2001; Schabenberger and Gotway, 2005). Note also that the second $\varepsilon(\mathbf{s})$ error component is usually referred to as the "nugget" effect in the geostatistical literature (Webster and Oliver, 2001).

Eq. (2) represents a versatile spatial linear prediction model that can incorporate various types of modeling assumptions. The deterministic component of the model $(\mathbf{X}\beta)$ can be defined to include trend surface parameters and/or fixed blocking effects, in addition to various hypothesized soil property/sensor relationships. As noted above, the stochastic error terms $(\eta(\mathbf{s})+\varepsilon(\mathbf{s}))$ can be parameterized to match the geostatistical covariance functions commonly used in spatial prediction and kriging. Indeed, ordinary kriging (OK), universal kriging (UK), and kriging with external drift (KED) and/or regression kriging (RK) models can all be derived as special cases of Eq. (2) (Schabenberger and Gotway, 2005; Haskard et al., 2007). Likewise, if the errors are assumed to be spatially uncorrelated (i.e., $\eta(\mathbf{s})=0$), then Eq. (2) reduces to an ordinary linear model. Depending upon how the design matrix $\mathbf{X}$ is specified, in this latter scenario one can obtain an ordinary regression model, a trend surface model, or an ANOCOVA model, etc.

In the most general case, $\mathbf{X}\beta$ may contain multiple fixed effects and the residual errors are assumed to exhibit some form of spatially correlated structure. Assume that the corresponding residual errors follow a Gaussian (e.g., multivariate Normal) distribution defined as

$$\begin{aligned} &\eta(\mathbf{s}) \sim G(0, \sigma_s^2 C(\theta)) \\ &\varepsilon(\mathbf{s}) \sim G(0, \sigma_n^2 \mathbf{I}) \\ &\mathrm{Cov}\{\eta(\mathbf{s}), \varepsilon(\mathbf{s})\} = 0 \\ &\Rightarrow \\ &\mathrm{Var}\{\eta(\mathbf{s}) + \varepsilon(\mathbf{s})\} = \sigma_s^2 \mathbf{C}(\theta) + \sigma_n^2 \mathbf{I} = \boldsymbol{\Sigma} \end{aligned} \tag{3}$$

where $\boldsymbol{\Sigma}$ is assumed to be positive definite and $\mathbf{C}(\theta)$ represents the correlation function of a second order stationary error process (for example, $\mathbf{C}(\theta)$ could represent an isotropic exponential correlation function with range parameter θ). Then using standard mixed linear modeling theory (Harville, 1990; Cressie, 1991; Harville and Jeske, 1992) one can show that the best linear unbiased estimator (BLUE) for β is

$$\hat{\beta} = \left(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \tag{4}$$

with a corresponding variance of

$$\mathrm{Var}\left(\hat{\beta}\right) = \left(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^{-1}. \tag{5}$$

Likewise, one can show that the best linear unbiased prediction (BLUP) for $\mathbf{y}_z$ (where $\mathbf{y}_z$ represents the remaining (non-sampled) survey locations can be expressed as

$$\hat{\mathbf{y}}_z = \mathbf{X}_z \hat{\beta} + \boldsymbol{\Sigma}_{yz} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) \tag{6}$$

where X_z represents the design matrix associated with y_z and Σ_{yz} represents the model based covariance matrix between y_z and the observed sample data y. Additionally, the corresponding variance estimate associated with this prediction vector is

$$\begin{aligned} \mathrm{Var}\left(\mathbf{y}_z - \hat{\mathbf{y}}_z\right) = {} & \boldsymbol{\Sigma}_z - \boldsymbol{\Sigma}_{yz} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{yz}^T \\ & + \left[\mathbf{X}_z - \boldsymbol{\Sigma}_{yz} \boldsymbol{\Sigma}^{-1} \mathbf{X}\right] \left(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^{-1} \left[\mathbf{X}_z - \boldsymbol{\Sigma}_{yz} \boldsymbol{\Sigma}^{-1} \mathbf{X}\right]^T \end{aligned} \tag{7}$$

where $\boldsymbol{\Sigma}_z$ represents the model based variance matrix of $\mathbf{y}_z$ (Goldberger, 1962; Cressie, 1991).

When the covariance structure is known up to a proportionality constant in the geostatistical mixed linear model (i.e., $\boldsymbol{\Sigma} = \tau^2 \mathbf{V}$, where $\mathbf{V}$ is assumed known *a priori*), the β parameter vector in Eq. (2) can be estimated using generalized least squares (Graybill, 1976; Rao and Toutenburg, 1999). However, the specific $\boldsymbol{\Sigma}$ parameter values are rarely known *a priori*. In practice, the β parameter vector and $\boldsymbol{\Sigma}$ variance structure must be jointly estimated from the sample data, typically using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation techniques (Littell et al., 1996). In such situations it is generally necessary to collect a fairly large amount of sample data in order to reasonably estimate the parameters associated with the covariance structure when even the simplest isotropic covariance functions are employed (Irvine et al., 2007).

2.2. Ordinary linear regression (LR) models

Consider the structure of the variance components shown in Eq. (3). This variance structure (in the full geostatistical mixed linear model) can be expressed as

$$\begin{aligned} &\mathrm{Var}\{\eta(\mathbf{s}) + \varepsilon(\mathbf{s})\} = \sigma_s^2 \mathbf{C}(\theta) + \sigma_n^2 \mathbf{I} = \Sigma = \tau^2 \mathbf{V} \\ &\text{for } \tau^2 = \sigma_s^2 + \sigma_n^2, \quad \mathbf{V} = \alpha \mathbf{C}(\theta) + (1-\alpha)\mathbf{I}, \text{ and } \alpha = \sigma_s^2 / \tau^2. \end{aligned} \tag{8}$$

The degree of spatial correlation structure in the residual error distribution is determined by σ_s^2 (the partial sill) and the assumed structure for $\mathbf{C}(\theta)$, where (for isotropic functions) the scalar value of θ controls the range for an isotropic process. Clearly, as either $\sigma_s^2 \to 0$ or $\theta \to 0$, the residual errors become spatially uncorrelated and V becomes proportional to an identity matrix. More importantly, if/when the effective (or absolute) range for $\mathbf{C}(\theta)$ is less than the minimum separation distance between the calibration sampling locations, the residual errors again become approximately (or strictly) spatially uncorrelated.

This latter point is particularly relevant when calibrating soil property information to many types of remotely sensed survey data. In properly planned salinity surveys where the auxiliary sensor data are expected to be well correlated with the response variable of interest, it is often possible to sample beyond the expected effective range of spatial correlation in the residual pattern. For example, field-scale soil salinity patterns can often be mapped with very high precision using bulk soil electrical conductivity survey data and ordinary LR models, since the residual error distribution typically exhibits only short range spatial correlation (Lesch et al., 1995; Corwin & Lesch, 2005; Lesch et al., 2005). Therefore, a simpler LR model can be used in place of the full geostatistical MLM to generate a map with a high degree of prediction accuracy, provided that an appropriate sampling strategy is employed (Lesch, 2005). This is especially advantageous in commercial applications, since LR models can be estimated using far less sample data (i.e., typically 10 to 15 sites).

In principal, an LR model can be used in place of the more elaborate geostatistical MLM whenever the model is assumed to exhibit either (i) strictly uncorrelated (SU) residual errors, or (ii) effectively uncorrelated (EU) residual errors. Note that the SU residual assumption implies that $\sigma_s^2 = 0$ and thus $\Sigma = \sigma_n^2 \mathbf{I}$. Under such an assumption, the geostatistical model collapses exactly into an ordinary LR model and all of the usual results for ordinary regression models can be applied in this situation. Such models are referred to as "*spatially referenced*" regression models by Lesch (2005), and as "*aspatial*" models by Fotheringham et al. (2002). In contrast, the (typically more reasonable) EU residual assumption implies that a set of calibration sample sites have been acquired where all of the minimum nearest-neighbor distances between sampling locations exceeds the effective range of residual spatial correlation. In this latter situation the geostatistical model does not reduce exactly into an ordinary LR model; for example, we can still use the available sample data to compute the BLUE for β, but not the empirical BLUP for $\mathbf{y}_z$. More specifically, both the parameter estimates and *computable* survey predictions turn out to be the same under either residual error assumption, but their statistical properties change. These estimation and prediction concepts are discussed in detail below.

2.2.1. Parameter estimate and prediction formulas: strictly uncorrelated residual assumption

Under the SU residual assumption, one can easily verify that $\boldsymbol{\Sigma} = \sigma_n^2 \mathbf{I}$, $\boldsymbol{\Sigma}_z = \sigma_n^2 \mathbf{I}_z$, and $\boldsymbol{\Sigma}_{yz} = 0$. Thus, the BLUE for β becomes

$$\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \tag{9}$$

with a corresponding variance of

$$\mathrm{Var}\left(\hat{\beta}\right) = \sigma_n^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}. \tag{10}$$

Additionally, the BLUP for $\mathbf{y}_z$ reduces to

$$\hat{\mathbf{y}}_z = \mathbf{X}_z \hat{\beta} \quad (\text{since } \boldsymbol{\Sigma}_{yz} = 0) \tag{11}$$

with a corresponding variance estimate of

$$\mathrm{Var}\left(\mathbf{y}_z - \hat{\mathbf{y}}_z\right) = \sigma_n^2 \left(\mathbf{I}_z + \mathbf{X}_z \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}_z^T\right). \tag{12}$$

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

$$\hat{y}_0 = \mathbf{x}_z\hat{\beta}$$
$$\mathrm{Var}\{y_0-\hat{y}_0\} = \sigma_n^2\left(1+\mathbf{x}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T\right) \quad (13)$$

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

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

where $\mathbf{x}_{\mathrm{ave}}$ represents the average of the $N-n$ design vectors associated with the non-sampled survey positions. Note that all of these results are exactly identical to ordinary LR model parameter estimation and prediction formulas (Myers, 1986).

2.2.2. Estimate and prediction formulas: effectively uncorrelated residual assumption

Under the EU residual assumption, we find that $\boldsymbol{\Sigma}\approx(\sigma_n^2+\sigma_s^2)\mathbf{I}$ where this approximation is exact if the effective range of the spatial residual correlation structure is both finite and less than the minimum nearest-neighbor sampling distance. When the above relationship is exact, we again find the BLUE for β to be equal to (9), but now with a corresponding variance of

$$\mathrm{Var}\left(\hat{\beta}\right) = \tau^2(\mathbf{X}^T\mathbf{X})^{-1}, \quad (15)$$

for $\tau^2=\sigma_n^2+\sigma_s^2$. In practice, the calculated mean square error (MSE) estimate can be used to yield an unbiased estimate of τ^2, but this pooled variance estimate can not be partitioned into it's individual components without further assumptions (or prior information). Furthermore, one can not necessarily assume that $\boldsymbol{\Sigma}_z=\sigma_n^2\mathbf{I}_z$ or $\boldsymbol{\Sigma}_{yz}=0$. (These covariance matrices are instead determined by the assumed $\eta(s)$ spatial structure and the density of the survey grid.) Thus, rather than Eq. (11), the BLUP for z under the EU residual assumption reduces to a simplified version of Eq. (6); i.e.,

$$\hat{\mathbf{y}}_z = \mathbf{X}_z\hat{\beta}+\alpha\mathbf{C}_{yz}(\theta)\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right) \quad (16)$$

where $\mathbf{C}_{yz}$ represents the model based correlation matrix between $\mathbf{y}_z$ and the observed sample data $\mathbf{y}$ and as before $\alpha=\sigma_s^2/\tau^2$. Via direct substitution of the above variance components, the corresponding variance estimate associated with this BLUP can be easily shown to be

$$\mathrm{Var}\left(\mathbf{y}_z-\hat{\mathbf{y}}_z\right) = \tau^2\left[\Omega_z-\Omega_{yz}\Omega_{yz}^T+(\mathbf{X}_z-\Omega_{yz}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}_z-\Omega_{yz}\mathbf{X})^T\right] \quad (17)$$

where $\boldsymbol{\Omega}_z=\alpha\mathbf{C}_z(\theta)+(1-\alpha)\mathbf{I}_z$ and $\Omega_{yz}=\alpha\mathbf{C}_{yz}(\theta)$.

Clearly, Eq. (16) is not of much practical use, since by assumption the sample data will yield no information about the short-range spatial correlation structure of the residuals. Thus, in the absence of prior information, one can not estimate either $\mathbf{C}_{yz}$ or α in Eq. (16) and hence the empirical BLUP for $\mathbf{y}_z$ can not be constructed. To circumvent this problem, we can instead use the less efficient (but computable and still unbiased) linear predictor given in (11); i.e., $\hat{\mathbf{y}}_z=\mathbf{X}_z\hat{\beta}$. After some straight-forward matrix algebra (see Proof 1A in the Appendix), we find that the corresponding variance estimate associated with this linear unbiased predictor (LUP) is

$$\mathrm{Var}\left(\mathbf{y}_z-\hat{\mathbf{y}}_z\right) = \tau^2\left[\Omega_z-2\cdot\Omega_{yz}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_z^T+\mathbf{X}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_z^T\right]. \quad (18)$$

Likewise, the corresponding variance estimate associated with a single new prediction on the survey grid becomes

$$\mathrm{Var}\left(\mathbf{y}_0-\hat{\mathbf{y}}_0\right) = \tau^2\left[1-2\cdot\omega_{yz}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T+\mathbf{x}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T\right] \quad (19)$$

where $\mathbf{x}_z$ is again the $(1\times p)$ design vector associated with the new prediction site and ω_{yz} represents the spatial correlation vector between the calibration data ($\mathbf{y}$) and y_0; i.e., $\omega_{yz}=\alpha\cdot\mathbf{c}_{yz}(\theta)$.

It is interesting to note that even when we resort to using the computable LUP (under the EU residual assumption), we still can not calculate the corresponding variance estimates. For this reason, practitioners instead typically use the ordinary variance formulas shown in Eqs. (12) and (13). Fortunately, provided that the ancillary sensor data exhibits a sufficiently smooth spatial structure, these ordinary (SU) variance formulas tend to be conservative. To understand intuitively how Eq. (19) can produce a variance estimate that is less than Eq. (13), consider the following scenario. Suppose that the new prediction site is arbitrarily close to the ith existing calibration site and assume that the ancillary sensor data exhibits a strictly continuous (i.e., no nugget) spatial structure, such that $\mathbf{x}_z\rightarrow\mathbf{x}_i$ as the separation distance between the two sites becomes arbitrarily small. Additionally, the ith component of the ω_{yz} correlation vector must converge to α (for $0<\alpha\leq 1$) and all remaining components converge to 0 (under the EU residual assumption). Substituting these limiting values into Eq. (19) yields

$$\begin{aligned}\mathrm{Var}\left(y_0-\hat{y}_0\right) &= \tau^2\left[1-2\alpha\cdot\mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T+\mathbf{x}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T\right]\\ &\approx\tau^2\left[1+(1-2\alpha)\cdot\mathbf{x}_z(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_z^T\right] \quad (\text{since } \mathbf{x}_i\approx\mathbf{x}_z)\end{aligned} \quad (20)$$

which must clearly be smaller than Eq. (13) $\forall\alpha>0$.

In summary, both the SU and EU residual assumptions lead to the same set of best linear unbiased parameter estimates. Thus, all of the general statistical theory associated with parameter estimation in ordinary linear regression models is immediately applicable (under either assumption). Additionally, the SU assumption leads to the best linear unbiased, ordinary regression model predictor $\hat{\mathbf{y}}_z=\mathbf{X}_z\hat{\beta}$. In theory, the EU assumption leads to a different set of best linear unbiased predictions. However, the simpler LUP $\hat{\mathbf{y}}_z=\mathbf{X}_z\hat{\beta}$ can again be used under this latter assumption, along with the ordinary variance formulas. In most cases these formulas will produce conservative variance estimates, provided that the ancillary sensor data exhibits a sufficiently smooth spatial structure. Additionally, these LU predictions will exhibit only trivial losses in efficiency (compared to their BLU counterparts) if/when $\sigma_s^2\approx 0$ or $\theta\approx 0$.

2.2.3. A regression based prediction formula for computing Prob($y_0>c$)

In many practical survey applications, determining the probability that a new prediction exceeds some specific threshold value is also of interest. In the geostatistical literature, indicator kriging and disjunctive kriging are two commonly used modeling techniques for estimating such probabilities (Matheron, 1976; Deutsch and Journel, 1992; Chilès and Delfiner, 1999; Webster and Oliver, 2001); note that cokriging equivalent forms of these techniques are sometimes used when ancillary sensor data is available.

Although not commonly discussed in most classical linear modeling textbooks, regression models can also be used to produce such probability estimates, at least from the Bayesian viewpoint. More specifically, given the SU residual assumption and upon adopting a Bayesian perspective, the probability that an unobserved y_0 lies within the interval (a, b) can be computed as

$$\pi_i[a,b] = \mathrm{Prob}(a\leq y_0\leq b) = \int_g^h t_{(n-p)}dt \quad (21)$$

where $t_{(n-p)}$ represent a central t-distribution having $n-p$ degrees of freedom (i.e., the regression model residual degrees of freedom),

$g = (a - \hat{y}_0) \Big/ \sqrt{\mathrm{Var}\{\hat{y}_0\}}$ and $h = (b - \hat{y}_0) \Big/ \sqrt{\mathrm{Var}\{\hat{y}_0\}}$ (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

$$\mathrm{RIE}[a,b] = \frac{100}{N-n} \sum_{i=1}^{N-n} \pi_i[a,b] \tag{22}$$

which represents a prediction of the percentage of non-sampled 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. Eqs. (21) and (22) 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. Technically, the above calculations are only strictly valid under the SU residual assumption, since a precise estimate for var$\{\hat{y}_0\}$ can not be computed under the weaker EU assumption (i.e., without *a priori* knowledge of the short-range spatial correlation structure).

3. Residual and prediction validation tests for LR models

3.1. The Moran test statistic

If an ordinary LR model is to be successfully used in place of the geostatistical MLM, then more restrictive modeling assumptions need to be met. In addition to the Gaussian error process, the critical assumption in the LR model is the EU residual assumption. In a spatial context, this assumption implies that the residual errors associated with the calibration sample site locations are (at least approximately) uncorrelated. Thus, some type of test for residual spatial correlation should always be performed before deciding to adopt the ordinary LR modeling approach.

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 MLM (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 LR 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 (Ripley, 1981; Schabenberger and Gotway, 2005). However, it can also be used to assess the EU residual assumption in the geostatistical modeling framework. The Moran residual test statistic (δ_M) is defined as

$$\delta_M = \frac{\mathbf{r}^T \mathbf{W} \mathbf{r}}{\mathbf{r}^T \mathbf{r}} \tag{23}$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\beta}$ (e.g., the vector of observed model residuals), $\mathbf{W}$ represents a suitably specified proximity matrix, and $\hat{\beta}$ is calculated using Eq. (9). While the specification of $\mathbf{W}$ 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

$$w_{ii} = 0 \text{ and } w_{ij} = d_{ij}^{-2} \Big/ \sum_{j=1}^{n} d_{ij}^{-2}, \tag{24}$$

respectively.

Brandsma and Ketellapper (1979) showed that the first two moments of δ_M are

$$E(\delta_M) = tr(\mathbf{MW})/(n-p) \tag{25}$$

and

$$\mathrm{Var}(\delta_M) = \frac{tr(\mathbf{MWMW}^T) + tr(\mathbf{MWMW}) + \{tr(\mathbf{MW})\}^2}{(n-p)(n-p+2)} - \{E(\delta_M)\}^2 \tag{26}$$

where $E(\ )$, $Var(\ )$, and $tr(\ )$ represent expectation, variance, and trace functions; n and p represent the number of sample sites (calibration sample size) and regression model parameters (including the intercept); and $\mathbf{M}$ is defined to be $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The corresponding Moran test score can then be computed as

$$z_M = (\delta_M - E(\delta_M)) / \sqrt{\mathrm{Var}(\delta_M)} \tag{27}$$

and 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 fixed effects in the regression model have been correctly specified, such a test score implies that the ordinary LR model residuals exhibit significant spatial correlation. In this situation, the LR 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 is available (or additional data can be collected), then a suitable spatial or geostatistical linear modeling approach should instead be employed.

3.2. Techniques for assessing the Gaussian (Normal) error assumption

In addition to the EU residual assumption, one must also verify that the LR model residuals satisfy the usual standard Gaussian 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 LR model. These include assessing the assumption of residual Normality using quantile–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, added variable plots, etc.).

The standard jack-knifing techniques commonly used to assess the predictive capability of an ordinary regression model are also directly applicable to the LR model in the spatial setting. Most standard statistical software packages can readily produce jack-knifed residual and/or prediction estimates in a computationally efficient manner. (Note that in the geostatistical literature, jack-knifing is typically referred to as cross-validation.) Cook and Weisberg (1999) and Myers (1986) offer a good review of many relevant regression model diagnostic and assessment techniques.

3.3. Additional prediction validation tests for the LR model

Suppose that a plausible LR model has been specified that describes some type of soil property / sensor data relationship. Suppose also that after acquiring a data set of n samples, a Moran or likelihood ratio test verifies that the EU residual assumption is reasonable and that the other usual residual assumptions hold. Hence, this spatially referenced LR model can be conveniently expressed in matrix notation as $\mathbf{y} = \mathbf{X}\beta + \varepsilon(\mathbf{s})$, where $\varepsilon(\mathbf{s}) \sim N(0, \tau^2\mathbf{I}_n)$ and $\mathbf{y}$ and $\mathbf{X}$ are defined as in Eq. (2). In most surveys, the ultimate goal will be to use the fitted equation for prediction purposes, but assume first that we wish to assess the "validity" of our proposed LR model. Assume also

that it is possible to naturally split the acquired data set into two distinct sub-sets, having n_1 and n_2 samples, respectively.

There are three types of statistical tests that can be readily employed to assess the validity of the spatially referenced LR model. These can all be expressed as *F*-tests (and/or *t*-tests), and are based on the idea of data partitioning. To facilitate this discussion, let the two distinct sample sub-sets represent a primary calibration set and a secondary validation set. Given this data partition of $n=n_1+n_2$ sample sites, assume further that we wish to fit the model using the calibration data and then test it's prediction adequacy using the validation data.

First, with respect to the pooled (calibration+validation) data set, note that the pooled **y** vector and **X** matrix can be partitioned as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{bmatrix} \tag{28}$$

where the subscripts index the calibration and validation data sub-sets and the dimension of the partitioned design matrix is $(n_1+n_2)\times 2p$. Given this partition, a "composite model" *F*-test can be performed by fitting the partitioned equation

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon(\mathbf{s}) \tag{29}$$

and then testing if $\beta_1=\beta_2$ (Cook and Weisberg, 1999). This is one of the more well known, standard model validation testing techniques suggested in the statistical literature; additional details can be found in most regression textbooks (Cook and Weisberg, 1999; Myers, 1986; Weisberg, 1985).

However, there are two other tests that can also be easily performed on partitioned data and are quite appropriate for directly assessing prediction reliability. These are not commonly discussed in introductory regression textbooks, although their theory is easily derived from standard linear modeling techniques. In the discussion that follows, we will refer to these as the "joint-prediction" *F*-test and "mean-prediction" *t*-test, respectively.

The joint-prediction *F*-test can be performed by first estimating the LR model using just the calibration data, next calculating the joint set of prediction errors across the validation sites as

$$\mathbf{r}_2 = \mathbf{y}_2 - \mathbf{X}_2\hat{\beta}_1 \tag{30}$$

and then by computing the statistic

$$F_1 = \mathbf{r}_2^T\mathbf{V}^{-1}\mathbf{r}_2/s_1^2 \text{ where } \mathbf{V} = \left(\mathbf{I} + \mathbf{X}_2\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_2^T\right). \tag{31}$$

This test statistic, originally suggested by Lieberman (1961), essentially defines the joint (simultaneous) prediction region for multiple predictions from a single regression model. Given the EU residual assumption and under the null hypothesis (i.e., that the fitted calibration model is correct), F_1 follows a central $F(n_2,\ n_1-p)$ distribution where n_2 and n_1-p represent the number of validation sites and the (calibration) model degrees of freedom, respectively, and s_1^2 represents the estimated calibration model mean square error (MSE) estimate (Lieberman, 1961; Rao and Toutenburg, 1999). In a similar manner, the mean-prediction *t*-test can be performed by first calculating the average prediction error as

$$\bar{r} = \mathbf{q}^T\mathbf{r}_2 \text{ where } \mathbf{q}^T = [\,1/n_2,\ \ \ldots\ \ 1/n_2\,] \tag{32}$$

and then computing the statistic

$$t_1 = \bar{r}/\left(s_1\sqrt{h}\right) \text{ where } h = \left[(1/n_2) + \left(\mathbf{q}^T\mathbf{X}_2\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_2^T\mathbf{q}\right)\right]. \tag{33}$$

Note that t_1 follows a central *t* distribution (with n_1-p degrees of freedom) under the null hypothesis, where s_1 represents the square root of the calibration model MSE estimate (Rao and Toutenburg, 1999).

Intuitively, the composite-model *F*-test represents a test for non-equivalent parameter estimates across the partitioned calibration and prediction (validation) sample sites. In contrast, the joint-prediction *F*-test assesses the ability of the regression model (fit using the calibration data only) to make unbiased predictions at all new validation sites, and simultaneously tests if these predictions are within the specified tolerance (precision) of the model. The mean-prediction *t*-test follows from the joint-prediction *F*-test, and hence assesses the ability of the regression model to make an unbiased prediction of the average value across the new n_2 validation sites.

Like the composite model *F*-test, these latter two tests turn out to be simple to implement in standard regression modeling software. First, redefine the full regression equation to be

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{A}\psi + \varepsilon(\mathbf{s}) \tag{34}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \text{ and } \mathbf{A} = \begin{bmatrix} 0 \\ \mathbf{I}_2 \end{bmatrix}. \tag{35}$$

Note that the second **A** design matrix essentially introduces n_2 additional 0/1 indicator variables into the regression model, thus Eq. (34) can equivalently be viewed as a special type of ANOCOVA model. In this model, the $n_2\psi_j$ parameter estimates can be shown to be exactly equal to the calculated prediction errors defined in Eq. (30) and the corresponding estimated parameter variance matrix is exactly $s_1^2\mathbf{V}$ (see proof A2 in the Appendix). Therefore, the two parameter tests of $H_o: \psi_j = 0 \forall j$ and $H_o: \sum\psi_j = 0$ produce the exact same *F* and *t*-test statistics as those shown in Eqs. (31) and (33), respectively. Hence, both the joint-prediction *F*-test and mean-prediction *t*-test can be easily preformed using any standard statistical software package that supports regression modeling.

4. Case studies

The following case studies describe two examples where an ordinary LR modeling approach was successfully used to quantify the relationship between soil salinity and non-invasive electromagnetic (EM) conductivity survey data. The model estimation, parameter tests, and prediction computations described in each study highlight the various statistical issues discussed in Sections 2 and 3. All of the statistical analyses discussed below have been performed using SAS/STAT (SAS Inc., 1999a), SAS/IML (SAS Inc., 1999b), and the ESAP software package (Lesch et al., 2000).

4.1. Case study I: a survey of a marginally saline lettuce field in Indio, CA

An EM survey was performed by the Coachella Valley Resource Conservation District in June 2003 within a 14-ha vegetable field located in Indio, CA. A total of 2040 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 model-based sample design (Lesch et al., 2000). Soil samples were collected from 0–0.6 m and 0.6–1.2 m

Table 1
Basic EM38 and soil salinity 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
		0–0.6 m	12	1.86	1.18	0.72	4.22
EC_e	dS/m	0.6–1.2 m	12	1.93	1.28	0.26	3.92

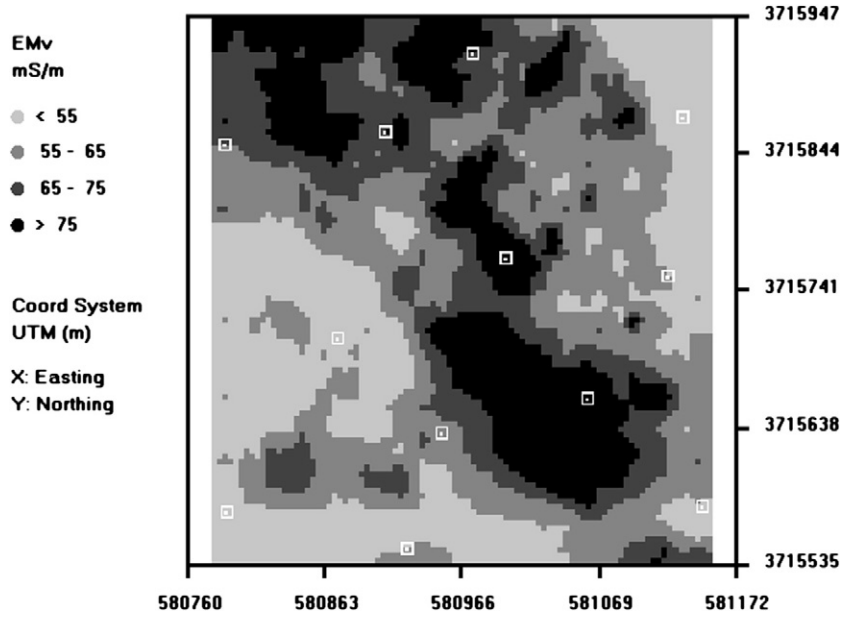

Fig. 1. Interpolated EMV signal map and corresponding sample site locations; Indio lettuce field.

depths and analyzed for soil salinity (EC_e, dS/m), soil saturation percentage (%), and percent gravimetric water content (%). Table 1 presents the univariate summary statistics (mean, standard deviation, minimum, and maximum) for the EM38 survey and soil salinity data, respectively. Fig. 1 shows the interpolated EM_V signal map for this field, along with the spatial positions of the 12 sampling locations. The primary goals of this survey were two-fold: (i) to construct an accurate soil salinity inventory for the field and (ii) to determine if this field should be leached before the fall cropping season.

The results from an exploratory regression modeling analysis performed in ESAP suggested that the following natural log(EC_e)/log (EM) regression equation best described the soil salinity / signal conductivity relationship in this field:

$$\ln(EC_{ij}) = \beta_{0j} + \beta_{1j}(x_{1i}) + \beta_{2j}(x_{2i}) + \beta_{3j}(x_{1i}^2) + \varepsilon_{ij} \quad (36)$$

where

$$x_{1i} = \ln(EM_{V,i}) + \ln(EM_{H,i}), \text{ and } x_{2i} = \ln(EM_{V,i}) - \ln(EM_{H,i}). \quad (37)$$

In Eq. (36), the subscript $j = 1, 2$ corresponds to the two sampling depths, $i = 1, 2, \ldots, 2040$ corresponds to the EM38 sampling locations, β_{0j} through β_{3j} 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. The upper portion of Table 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 an associated p-value, and the corresponding Moran test score and p-value. The Moran scores suggest that the EU residual assumption is valid. Likewise, residual QQ plots (not shown) confirm that the regression model errors follow a Normal distribution and hence the ordinary LR modeling approach can be adopted. Additionally, the R^2 values suggest that these LR models can be used to describe 92% and 80% of the 0–0.6 m and 0.6–1.2 m observed spatial log(EC_e) patterns in this field, respectively.

Table 2
Summary and prediction statistics for depth-specific ln(EC_e) LR models; Indio lettuce field

Model summary statistics						
Depth	R^2	Root MSE	F-score	$Pr>F$	Moran score (δ_M)	p-value
0–0.6 m	0.922	0.196	31.37	<0.001	0.652	0.257
0.6–1.2 m	0.798	0.490	10.54	0.004	−0.067	>0.5

Prediction Statistic	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)
% Area of field >3.0 dS/m	10.5	25.3

The spatial salinity pattern in the 0–0.6 m depth was of primary interest in this survey. More specifically, the field was scheduled to be leached if (i) the field average ln(EC_e) level exceeded ln(2)=0.693 or (ii)>25% of the field was predicted to exhibit 0–0.6 m depth salinity levels>3 dS/m. The predicted field average ln(EC_e) levels (and corresponding 95% confidence intervals) are shown in the lower portion of Table 2, along with the estimated area of the field that is exceeds 3 dS/m for both sampling depths. These predictions can be automatically calculated in the ESAP software package (using Eqs. (14) and (22), respectively). Fig. 2 shows the corresponding 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 2 and Fig. 2 suggest that this field does not need to be leached. The 0–0.6 m field average ln(EC_e) estimate is 0.494 and only 10.5% of the individual 0–0.6 m depth predictions are calculated to exceed 3 dS/m. Thus, neither of the specified thresholds for implementing a leaching process are met in this field.

The prediction statistics generated by the ESAP software package are always calculated under an SU residual assumption. As discussed previously, Eqs. (18) and (19) can be used to determine the correct variance estimates under the more realistic EU residual assumption, for any *a priori* specified spatial covariance structure. For example, we can compare the SU variance estimates to variance estimates derived under an assumed spherical covariance structure with a spatial range value ≤ sample minimum separation distance, etc.

In this specific EM survey, the minimum separation distance between the two nearest soil sampling locations was 88.4 m. Table 3 presents some relevant variance statistics for this field, using an alternative isotropic spherical covariance function that exhibits a range value of either 44 or 88 meters and a partial sill/total variance ratio of either 0.95 (5% nugget) or 0.5 (50% nugget). These statistics include (i) the relative variance estimate for the average prediction across the 2028 non-sampled survey locations,

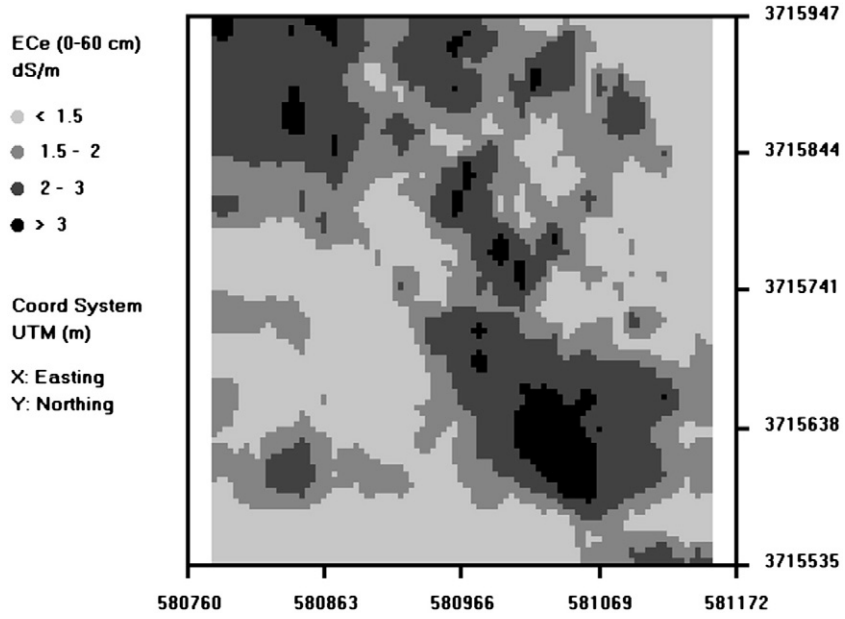

Fig. 2. Back-transformed and interpolated LR model salinity predictions, 0–0.6 m sample depth; Indio lettuce field.

(ii) the average value of the variance ratios associated with the individual, non-sampled survey locations, (iii) the maximum value of the variance ratios associated with the individual, non-sampled survey locations, (iv) the number of individual variance ratios that were found to be <1, and (v) the estimated percentage of non-sampled survey locations >3 dS/m. These statistics can be used to determine the sensitivity of the LR model prediction variance estimates under alternative EU residual scenarios.

The results shown in Table 3 confirm that the ordinary LR model variance formulas generally produce conservative variance estimates if the residual errors instead exhibit a short range spatial correlation structure. The calculated variance for the average prediction is clearly conservative, when compared to each of the four correlated error scenarios. More than 91% (1848 of 2028) of the individual prediction variance estimates are also conservative when the spherical range coefficient (θ) is set to 88 m, and over 98% of these prediction variance estimates are conservative for θ=44. Not surprisingly, the calculated average values for the individual variance ratios are always <1; more importantly, the maximum values imply that the ordinary LR model variance estimates are never under-estimated by more than 8.3% (under the alternative assumption of an isotropic spherical correlation structure). Finally, the recalculated percentage of survey sites >3 dS/m changes very little, suggesting that these probability calculations are relatively insensitive to short range residual correlation (at least in this particular example).

Table 3
A comparison of the average relative variance estimates, variance ratios and the probability of accidence (RIE) statistic under the SU and various EU residual assumptions; Indio lettuce field

Variance ratio or prediction statistic	SU residuals	EU residuals (isotropic spherical covariance)			
		θ=range, $\alpha=(\sigma_s^2/\tau^2)$			
	IID	α=0.95		α=0.50	
		θ=88 m	θ=44 m	θ=88 m	θ=44 m
$Var(\bar{y}_z)/\tau^2$	0.1006	0.0802	0.0953	0.0899	0.0978
Ave var($\hat{y}_0$|SU)/var($\hat{y}_0$|EU)	n/a	0.920	0.975	0.958	0.987
max Var($\hat{y}_0$|SU)/Var($\hat{y}_0$|EU)	n/a	1.083	1.026	1.044	1.014
# of Variance ratios<1	n/a	1848	1983	1848	1983
% Of survey sites>3 dS/m	10.52%	10.35%	10.48%	10.43%	10.50%

4.2. Case Study II: Model validation tests for a salinity survey in the San Joaquin Valley, CA

Corwin et al. (2006) describes a monitoring project undertaken to determine spatial-temporal changes in various soil properties resulting from drainage water reuse, in a 32.4-ha saline–sodic forage field in the San Joaquin Valley, CA. Electromagnetic induction surveys and multiple soil sampling projects have been performed in this field

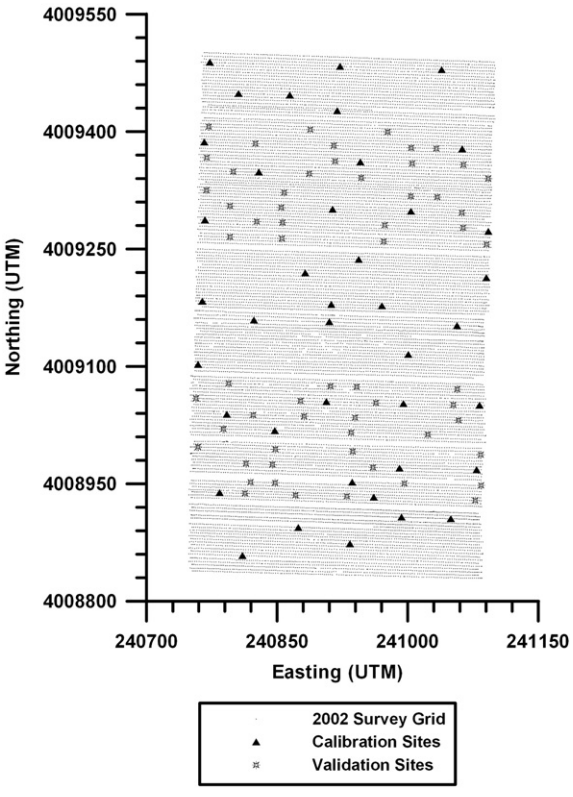

Fig. 3. Year 2002 EM survey grid and soil sampling plans; San Joaquin forage field.

Table 4
Basic EM38 and soil salinity summary statistics: San Joaquin Valley forage field

Variable	Units	N	Mean	Std. Dev.	Min	Max
EM_V	mS/m	22177	380.60	81.12	114.40	638.40
EM_H	mS/m	22177	297.53	61.89	132.80	604.80

Variable	Units	Data set	N	Mean	Std. Dev.	Min	Max
EC_e		Calibration	40	19.07	5.92	6.99	34.60
0–1.2 m	dS/m	Validation	60	19.74	6.31	8.36	32.48

every three years since 1999; details concerning the various survey protocols and sampling strategies can be found in the above mentioned reference. The initial EM survey grid acquired in 1999 was collected manually across the eight paddocks within this field and consisted of 384 EM38 horizontal and vertical survey locations. These readings were used to generate a prediction-based sampling strategy that selected 40 primary calibration sampling sites from these 384 grid locations (Lesch, 2005; Corwin et al., 2006).

In 2002, a mobilized EM survey was performed that resulted in the collection of 22,177 EM survey locations. The 40 EM survey locations within 1.5 m of the sample sites collected in 1999 were sampled again, along with 60 additional survey sites across four of the eight paddocks using a restricted, simple random sampling design (15 sites per paddock, selected from Paddocks 2, 3, 6, and 7). The one restriction incorporated into the random sampling design was that these 60 additional locations should correspond (approximately) to points on the original 1999 survey grid, thus insuring that the physical separation between any two sample sites was at least 25 m. Fig. 3 shows the 2002 EM survey grid, along with both soil sampling designs. Table 4 shows the corresponding EM survey and salinity sample data summary statistics (2002 survey/sample data). In the analysis that follows, the 0–1.2 m bulk average salinity levels associated with these 100 sampling locations are used to demonstrate the three regression model validation tests discussed previously in Section 3.

The initial regression function fit to the 2002 survey data was a natural log(EC_e) / log(EM) signal only regression equation; i.e.,

$$\ln(EC_i) = \beta_0 + \beta_1(x_{1i}) + \beta_2(x_{2i}) + \varepsilon_i \quad (38)$$

where the x_1 and x_2 signal terms are defined as in Eq. (37). This model produced an R^2 of 0.775, a root MSE estimate of 0.159, a non-significant β_2 parameter estimate (t-score p-value=0.349), and a statistically significant Moran residual test statistic (δ_M=2.64, p-value=0.004). A graphical analysis of the LR model residuals (trend and variogram plots, data not shown) suggested that the residuals exhibited long range, low-order (quadratic) spatial trends, rather than any short range spatial correlation structure. Thus, a modified equation using the x_1 signal data along with quadratic trend surface parameters was re-fit to the survey data; i.e.,

$$\ln(EC_i) = \beta_0 + \beta_1(x_{1i}) + \beta_2(s_{x,i}) + \beta_3(s_{y,i}) + \beta_4(s_{xy,i}) + \beta_5(s_{x^2,i}) + \beta_6(s_{y^2,i}) + \varepsilon_i \quad (39)$$

where the latter s_x and s_y terms represent the relative spatial coordinates of the 100 sampling locations. This revised model produced an R^2 of 0.811, a root MSE estimate of 0.149, a non-significant Moran residual test statistic (δ_M=0.63, p-value=0.263), and a clearly non-significant Shapiro–Wilk normality test statistic (SW=0.9945, p-value>0.5). The left-hand side of Table 5 shows the corresponding parameter estimates (and standard errors) for this estimated LR model.

Since the EU and normality residual assumptions appear to be reasonable, the three model validation tests can be used to further test the validity of a LR equation based solely on the non-random ESAP sampling locations. More specifically, these tests can be used to determine (i) if the parameter estimates appear to change across the two sampling designs (via the composite model F-test) and/or (ii) if the 40-site "calibration" equation can be used to predict the values of the 60 "validation" sites in an unbiased manner, within the precision of the estimated calibration model (using the joint-prediction F-test and mean-prediction t-test, respectively). As discussed previously, all three of these tests can be easily carried out in any standard LR modeling software package.

The right-hand side of Table 5 lists the parameter estimates (and standard errors) for the LR model estimated using just the 40 non-randomly selected calibration locations. After suitably partitioning the data set, the composite-model parameter test was performed in SAS using the TEST statement within the REG procedure; the corresponding composite-model F-score was 1.86 (p=0.086). Likewise, the joint-prediction F-test was conveniently carried out in SAS by first defining a new blocking variable that contained a unique site identification code for each new validation site (and a common, higher valued code for all calibration sites), and then fitting a standard ANOCOVA model using the GLM procedure. The corresponding F-score associated with this block effect was 0.72 (p=0.864). Additionally, a single ESTIMATE statement was used to produce the correct, mean-prediction t-score. In this example, the corresponding t-score was found to be −1.37 (p=0.181).

Table 5
LR model parameter estimates and associated standard errors for Eq. (39), using both the full (n=100) and calibration only (n=40) sample sites; San Joaquin Valley forage field

	n=100 (calibration+validation samples)		n=40 (calibration samples only)	
Parameter	Estimate	Std.Dev.	Estimate	Std. Dev.
β_0	2.975	0.031	3.034	0.050
β_1	0.244	0.018	0.246	0.034
β_2	−0.027	0.015	−0.064	0.026
β_3	−0.023	0.009	−0.011	0.015
β_4	0.002	0.010	0.007	0.016
β_5	−0.041	0.017	−0.069	0.027
β_6	−0.012	0.006	−0.012	0.008

Notes: relative spatial coordinates defined as s_x=(Easting−240923.5)/100 and s_y=(Northing−4009166.1)/100 in Eq. (39).

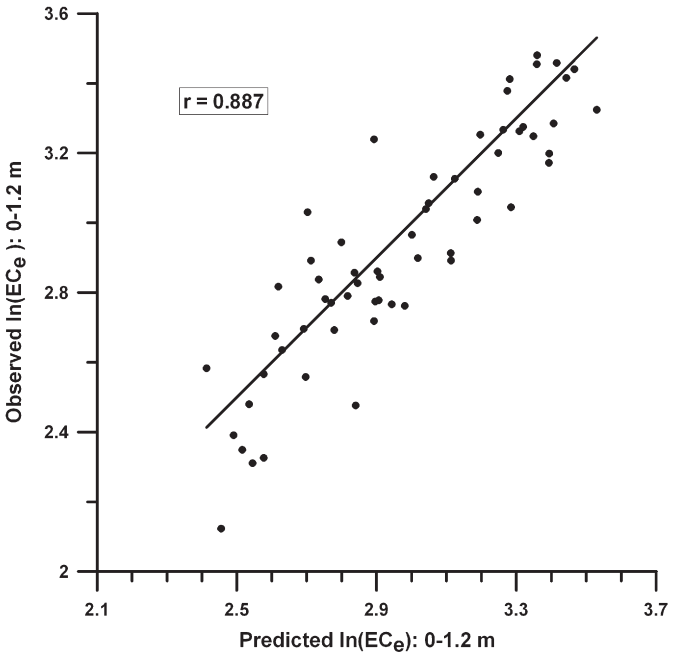

Fig. 4. Observed versus predicted 0–1.2 m ln(ECe) across the 60 validation sample locations; San Joaquin forage field.

All three of the above model validation tests produced non-significant test results at the 0.05 significance level. These non-significant results suggest that the salinity sample data associated with the model-based sampling design can be used to estimate a reliable and unbiased LR calibration equation. Additionally, these test results confirm that the ln (EC_e) values at the secondary set of validation sample locations can be predicted within the reported accuracy and precision of the estimated LR model. A plot of these observed versus predicted ln(EC_e) values at the 60 randomly selected validation locations is shown in Fig. 4.

5. Discussion and conclusion

The preceding two case studies demonstrate two situations where an ordinary LR model can be used in place of the more elaborate geostatistical MLM for spatial prediction purposes. Case study 1 represents a typical example of a field scale, soil salinity study based on a detailed EM survey (N=2040 survey sites) and a very limited set of soil calibration samples (n=12 sites). Given the high degree of correlation between the two data sets, an ordinary LR model can be effectively used to produce a spatially detailed salinity map. This map (and associated field summary statistics) can in turn be used to address the leaching question in a cost-effective manner. Furthermore, when the variances of the prediction estimates were computed under the SU residual assumption, we found that these variance estimates were generally conservative under the examined alternative (and perhaps more reasonable) EU assumptions.

Unlike the first study, in the second case study there are a sufficient number of sampling locations to estimate some type of suitable geostatistical MLM. In fact, our intention had been to use a more advanced spatial linear model to assess the two sampling strategies, had the data warranted such an approach. However, the data analysis results suggest that the linear model residuals exhibited no detectable spatial structure and thus a full geostatistical modeling approach was unnecessary. Additionally, since an ordinary LR modeling approach could be used here, the simpler classical model validation tests could also be employed to assess the adequacy of the non-random spatial sampling strategy.

Clearly, there are many applications where a geostatistical MLM (or similar geostatistical modeling technique) is obviously needed. However, we have also found many instances where the acquired survey / sample data suggest that a full geostatistical model is unnecessary. The correlation of bulk soil electrical conductivity sensor data (to various soil properties) and/or remotely sensed imagery data (to various crop biomass or yield indices) represent two current areas of research where ordinary LR modeling techniques often prove to be accurate, reliable and statistically valid (Barns et al., 2003; Corwin and Lesch, 2005; Eigenberg et al., in press). In these situations, we believe that an ordinary LR modeling approach can be used to great advantage, particularly when the project economics precludes the collection of a large number of calibration sample sites.

To summarize, in this article we have reviewed the connection between the simpler ordinary LR model and the full geostatistical MLM. The formulas for the ordinary LR model parameter estimates and best linear unbiased predictions have been derived under two different (SU and EU) residual error assumptions, along with the computable linear unbiased predictions and variance estimates under the EU error assumption. The Moran test for detecting spatial correlation in LR model residuals has also been discussed, in addition to three LR model validation tests that can be derived from classical linear modeling theory.

The two case studies highlight and demonstrate the various parameter estimation, response variable prediction, and model validation techniques reviewed in this article. The LR modeling and prediction method arises naturally (as a special case of the geostatistical MLM) whenever the residual errors satisfy either the SU or EU error assumptions. Indeed, providing the underlying modeling assumptions are satisfied, this simpler modeling approach can be effectively used to describe many different soil property/sensor data relationships, compute cost-effective spatial predictions, and produce classical model parameter and/or validation tests that are directly applicable to a broad array of spatial surveying projects.

Appendix

Definitions:		
Vector or matrix	Description	Dimension
$\mathbf{Y}$	Calibration data vector	$n\times1$
$\mathbf{y}_z$	Prediction data vector	$m\times1$
$\mathbf{X}$	Design matrix for calibration data	$n\times p$
$\mathbf{X}_z$	Design matrix for prediction data	$m\times p$
$\mathbf{\Sigma}_s(\mathbf{C})$	Spatial covariance (correlation) matrix, calibration data	$n\times n$
$\mathbf{\Sigma}_{zs}(\mathbf{C}_z)$	Spatial covariance (correlation) matrix, prediction data	$m\times m$
$\mathbf{\Sigma}_{yzs}(\mathbf{C}_{yz})$	Spatial covariance (correlation) matrix between $\mathbf{y}$ and $\mathbf{y}_z$	$m\times n$
β	Regression model parameter vector	$p\times1$

Proof A1. Variance of the linear unbiased predictor

Let $\Sigma_z=\tau^2\Omega_z$ for $\tau^2=\sigma_n^2+\sigma_s^2$, $\mathbf{\Omega}_z=\alpha\cdot\mathbf{C}_z(\theta)+(1-\alpha)\mathbf{I}_z$ and $\alpha=\sigma_s^2/\tau^2$ (as discussed in the text). The variance of the linear unbiased predictor (LUP) $\hat{\mathbf{y}}_z=\mathbf{X}_z\hat{\beta}$ under the effectively uncorrelated (EU) residual assumption can be most easily established by first noting that

$$\mathrm{Var}(\mathbf{y}_z-\hat{\mathbf{y}}_z)=\mathrm{Var}\left(\mathbf{y}_z-\mathbf{X}_z\hat{\beta}\right)=\mathrm{Var}(\mathbf{y}_z-\mathbf{A}\mathbf{y})\ \text{for}\ \mathbf{A}=\mathbf{X}_z\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T.$$

Thus,

$$\begin{aligned}\mathrm{Var}(\mathbf{y}_z-\hat{\mathbf{y}}_z)&=\mathrm{Var}(\mathbf{y}_z)-2\cdot\mathrm{cov}\left(\mathbf{y}_z,[\mathbf{A}\mathbf{y}]^T\right)+A\cdot\mathrm{Var}(y)\cdot A^T\\&=\Sigma_z-2\cdot\Sigma_{yz}A^T+\tau^2AA^T\end{aligned}$$

since under the EU assumption $\mathrm{var}(\mathbf{y})=\tau^2\mathbf{I}$. Direct substitution of the above matrix equalities yields Eq. (16).

Proof A2. An alternative way to calculate the joint-prediction F-test

Recall that the joint-prediction F-test, as originally derived by Lieberman (1961) is defined as

$$F_1=\mathbf{r}_2^T\mathbf{V}^{-1}\mathbf{r}_2/s_1^2$$

where

$$\mathbf{r}_2=\mathbf{y}_2-\mathbf{X}_2\hat{\beta}_1\ \text{and}\ \mathbf{V}=\mathbf{I}_2+\mathbf{X}_2\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_2^T.$$

Likewise, for

$$\mathbf{y}=\begin{bmatrix}\mathbf{y}_1\\\mathbf{y}_2\end{bmatrix},\quad\mathbf{X}=\begin{bmatrix}\mathbf{X}_1\\\mathbf{X}_2\end{bmatrix},\quad\mathbf{A}=\begin{bmatrix}0\\\mathbf{I}_2\end{bmatrix},$$

the ANOCOVA model discussed in the text can be expressed as

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{A}\psi+\varepsilon,\quad\mathrm{Var}(\varepsilon)=\sigma^2,$$

where the model errors are assumed to satisfy the EU residual assumption. To prove that the joint-prediction F-test is identically equivalent to the ANOCOVA joint parameter test of $\psi=0$, it is sufficient to show that $\hat{\psi}=\mathrm{r}_2$, $Var(\hat{\psi})=\sigma^2\mathbf{V}$, and $\hat{\sigma}^2=s_1^2$.

Define $\mathbf{Z}=\begin{bmatrix}\mathbf{X}_1&0\\\mathbf{X}_2&\mathbf{I}_2\end{bmatrix}$ and $\Delta=\begin{bmatrix}\beta\\\psi\end{bmatrix}$, so that the ANOCOVA equation can be re-expressed as $\mathbf{y}=\mathbf{Z}\Delta+\varepsilon$. For this model, we find that

$$\mathbf{Z}^T\mathbf{Z}=\begin{bmatrix}\mathbf{X}_1^T\mathbf{X}_1+\mathbf{X}_2^T\mathbf{X}_2&\mathbf{X}_2^T\\\mathbf{X}_2&\mathbf{I}_2\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}&\mathbf{A}_{12}\\\mathbf{A}_{21}&\mathbf{A}_{22}\end{bmatrix}=\mathbf{Q}$$

$$\Rightarrow$$

$$\mathbf{Q}^{-1}=\begin{bmatrix}\mathbf{B}_{11}&\mathbf{B}_{12}\\\mathbf{B}_{21}&\mathbf{B}_{22}\end{bmatrix}$$

for $\mathbf{B}_{11}=(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$, $\mathbf{B}_{12}=-\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$, $\mathbf{B}_{21}=\mathbf{B}_{12}^{T}$ and $\mathbf{B}_{22}=\mathbf{A}_{22}^{-1}+\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$. Direct imputation of the above design matrices yields
$\mathbf{B}_{11}=(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$, $\mathbf{B}_{12}=(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_2^T$, $\mathbf{B}_{21}=-\mathbf{X}_2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$ and
$\mathbf{B}_{22}=\mathbf{I}_2+\mathbf{X}_2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_2^T$.

Hence $\hat{\boldsymbol{\Delta}}=(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$ is equal to

$$\begin{bmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & -(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_2^T \\ -\mathbf{X}_2(\mathbf{X}_1^T\mathbf{X}_1)^{-1} & \mathbf{I}_2+\mathbf{X}_2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_2^T \end{bmatrix} \times \begin{bmatrix} \mathbf{X}_1^T & \mathbf{X}_2^T \\ 0 & \mathbf{I}_2 \end{bmatrix} \times \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} =$$

$$\begin{bmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}_1 \\ \mathbf{y}_2-\mathbf{X}_2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \mathbf{y}_2-\mathbf{X}_2\hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta} \\ \hat{\psi} \end{bmatrix}.$$

Thus, $\hat{\psi}=\mathbf{r}_2$ and $\mathrm{var}(\hat{\psi})=\sigma^2\mathbf{B}_{22}=\sigma^2\mathbf{V}$. Additionally, since $\hat{\beta}=\hat{\beta}_1$ and the ANOCOVA model contains the same corrected degrees of freedom as the simpler regression model fit just to the first set of sample data, $\hat{\sigma}^2=s_1^2$. Therefore, the two test statistics must be exactly equivalent.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Boca Raton, FL.

Barns, E.M., Sudduth, K.A., Hummel, J.W., Lesch, S.M., Corwin, D.L., Yang, G., Doughtry, C.S.T., Bausch, W.C., 2003. Remote- and ground-based sensor techniques to map soil properties. Photogramm. Eng. Remote Sensing 69, 619–630.

Brandsma, A.S., Ketellapper, R.H., 1979. Further evidence on alternative procedures for testing of spatial autocorrelation amonst regression disturbances. In: Bartels, C.P.A., Ketellapper, R.H. (Eds.), Exploratory and Explanatory Statistical Analysis of Spatial Data. Martinus Nijhoff, Boston, MA, pp. 113–136.

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

Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. John Wiley, New York, NY.

Cook, R.D., Weisberg, S., 1999. Applied Regression including Computing and Graphics. John Wiley, New York, N.Y.

Corwin, D.L., Lesch, S.M., 2005. Characterizing soil spatial variability with apparent soil electrical conductivity. I. Survey protocols. Comp. Electron. Ag. 46, 103–134.

Corwin, D.L., Lesch, S.M., Oster, J.D., Kaffka, S.R., 2006. Monitoring management-induced spatio-temporal changes in soil quality through soil sampling directed by apparent electrical conductivity. Geoderma 131, 369–387.

Cressie, N.A.C., 1991. Statistics for Spatial Data. John Wiley, New York, N.Y.

Deutsch, C.V., Journel, A.G., 1992. GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press, New York, NY.

Eigenberg, R.A., Lesch, S.M., Woodbury, B., Nienaber, J.A. in press. Geospatial methods for monitoring a vegetative treatment area. J. Environ. Qual.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression. John Wiley, New York, NY.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear model. J. Am. Stat. Assoc. 57, 369–375.

Gotway, C.A., Stroup, W.W., 1997. A generalized linear model approach to spatial data analysis and prediction. J. Ag. Bio. Environ. Stats. 2, 157–178.

Graybill, F.A., 1976. Theory and Application of the Linear Model. Wadsworth Publishing Co., Inc., Belmont, CA.

Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge, UK.

Harville, D.A., 1990. BLUP (Best Linear Unbiased Prediction), and beyond. Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, Berlin, Germany, pp. 239–276.

Harville, D.A., Jeske, D.R., 1992. Mean square error of estimation or prediction under a general linear model. J. Am. Stat. Assoc. 87, 724–731.

Haskard, K.A., Cullis, B.R., 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., Kachanoski, R.G., 2002. Indirect measurement of solute concentration. In: Dane, J.H., Topp, G.C. (Eds.), Methods of Soil Analysis, Part 4 — Physical Methods. . Soil Sci. Soc. Am. Book Series, vol. 5. Soil Science Society of America, Madison, WI, pp. 1274–1306.

Irvine, K.M., Gitelman, A.I., Hoeting, J.A., 2007. Spatial designs and properties of spatial correlation: effects on covariance estimation. J. Ag. Bio. Environ. Stats. 12, 450–469.

Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY.

Lesch, S.M., 2005. Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. Comp. Electron. Ag. 46, 153–180.

Lesch, S.M., Strauss, D.J., 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.

Lesch, S.M., Rhoades, J.D., Corwin, D.L., 2000. ESAP-95 Version 2.10R: User Manual and Tutorial Guide. Research Rpt. 146. USDA-ARS, George E. Brown, Jr. Salinity Laboratory, Riverside, CA, USA.

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

Lieberman, G.J., 1961. Prediction regions for several predictions from a single regression line. Technometrics 3, 21–27.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., 1996. SAS System for Mixed Models. SAS Institute Inc., Cary, NC.

Matheron, G., 1976. A simple substitute for conditional expectation: the disjunctive kriging. In: Guarascio, M., David, M., Huijbregts, C. (Eds.), Advanced Geostatistics in the Mining Industry. Reidel, Dordrecht, pp. 221–236.

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., 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, MA.

Press, S.J., 1989. Bayesian Statistics: Principles, Models, and Applications. John Wiley, New York, NY.

Rao, C.R., Toutenburg, H., 1999. Linear Models: Least Squares and Alternatives. Springer-Verlag, New York, NY.

Rhoades, J.D., Chanduvi, F., Lesch, S.M., 1999. Soil salinity assessment: methods and interpretation of electrical conductivity measurements. FAO Irrigation and Drainage Paper #57. Food and Agriculture Organization of the United Nations, Rome, Italy.

Ripley, B.D., 1981. Spatial Statistics. John Wiley, New York, NY.

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

SAS Institute Inc, 1999a. SAS/STAT User's Guide, Version 8. SAS Institute Inc., Cary, NC.

SAS Institute Inc, 1999b. SAS/IML User's Guide, Version 8. SAS Institute Inc., Cary, NC.

Schabenberger, O., Gotway, C.A., 2005. Statistical Methods for Spatial Data Analysis. CRC Press, Boca Raton, FL.

Schabenberger, O., Pierce, F.J., 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for Normality (complete samples). Biometrika 52, 591–611.

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, NY.

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

Upton, G., Fingleton, B., 1985. Spatial Data Analysis by Example. John Wiley, New York, NY.

Wackernagel, H., 1998. Multivariate Geostatistics, 2nd Ed. Springer-Verlag, Berlin, Germany.

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

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. Johy Wiley, New York, NY.

Weisberg, S., 1985. Applied Linear Regression, 2nd Ed. John Wiley, New York, NY.

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