location cannot be performed unless the tillage regime and soil nitrogen at $\mathbf{s}_0$ are known. This is a feature of all regression methods and should not be construed as a shortcoming. If the modeler determines that Y depends on X and the value of X is unknown the model cannot be used to produce a prediction of Y.

9.5 Spatial Regression and Classification Models

9.5.1 Random Field Linear Models

A spatial regression or classification model is a model for geostatistical data where interest lies primarily in statistical inference about the mean function $\mathrm{E}[Z(\mathbf{s})] = \mu(\mathbf{s})$. The most important application of these models in the crop and soil sciences is the analysis of field experiments where the experimental units exhibit spatial autocorrelation. The agricultural variety trial is probably the most important type of experiment to which the models in this section can be applied, but any situation in which $\mathrm{E}[Z(\mathbf{s})]$ is modeled as a function of other variables in addition to a spatially autocorrelated error process falls under this heading. Variety trials are particularly important here because of their respective size. Randomization of treatments to experimental units neutralizes the effects of spatial correlation among experimental units and provides the framework for statistical inference in which cause-and-effect relationships can be examined. These trials are often conducted as randomized block designs and, because of the large number of varieties involved, the blocks can be substantial in size. Combining adjacent experimental units into blocks in agricultural variety trials can be at variance with an assumption of homogeneity within blocks. Stroup, Baenziger and Mulitze (1994) notice that if more than eight to twelve experimental units are grouped, spatial trends will be removed only incompletely. Although randomization continues to neutralize these effects, it does not eliminate them as a source of experimental error.

Figure 9.25 shows the layout of a randomized complete block design conducted as a field experiment in Alliance, Nebraska. The experiment consisted of 56 wheat cultivars arranged in four blocks and is discussed in Stroup et al. (1994) and Littell et al. (1996).

Analysis of the plot yields in this RCBD reveals a p-value for the hypothesis of no varietal differences of $p = 0.7119$ along with a coefficient of variation of $CV = 27.58\%$. A p-value that large should give the experimenter pause. That there are no yield differences among 56 varieties is very unlikely. The large coefficient of variation conveys the considerable magnitude of the experimental error variance. Blocking as shown in Figure 9.25 did not eliminate the spatial dependencies among experimental units and left any spatial trends to randomization which increased the experimental error. The large p-value is not evidence of an absence of varietal differences, but of an experimental design lacking power to detect these differences.

Instead of the classical RCBD analysis one can adopt a modeling philosophy where the variability in the data from the experiment is decomposed into large-scale trends and smooth-scale spatial variation. Contributing to the large-scale trends are treatment effects, deterministic effects of spatial location, and other explanatory variables. The smooth-scale variation consists of a spatial random field that captures, for example, smooth fertility trends.

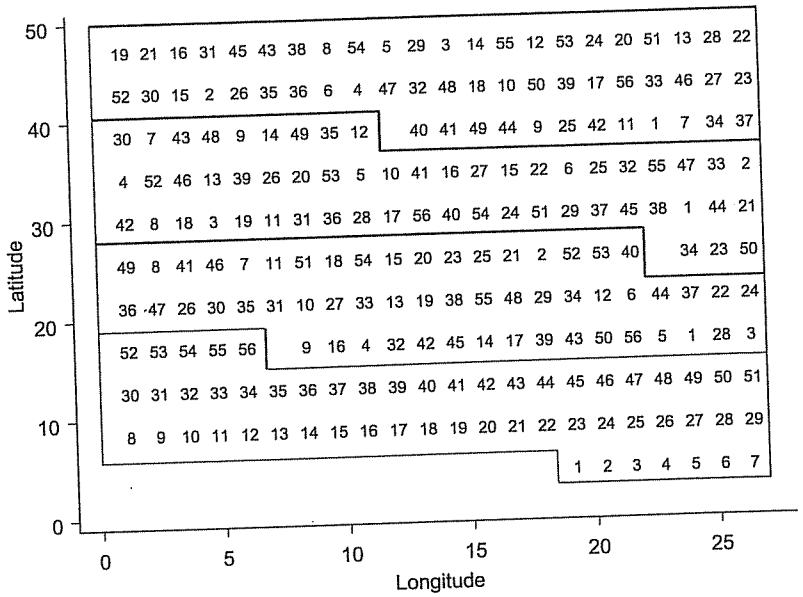

Figure 9.25. Layout of wheat variety trial at Alliance, Nebraska. Lines show block boundaries, numbers identify the placement of varieties within blocks. There are four blocks and 56 varieties. Drawn from data in Littell et al. (1996).

In the notation of §9.4 we are concerned with the spatial mean model

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}),$$

where $\delta(\mathbf{s})$ is assumed to be a second-order stationary spatial process with semivariogram $\gamma(\mathbf{h})$ and covariogram $C(\mathbf{h})$. The mean model $\mu(\mathbf{s})$ is assumed to be linear in the large-scale effects, so that we can write

$$Z(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \delta(\mathbf{s}). \qquad [9.54]$$

We maintain the dependency of the design/regressor matrix $\mathbf{X}(\mathbf{s})$ on the spatial location since $\mathbf{X}(\mathbf{s})$ may contain, apart from design (e.g., block) and treatment effects, other variables that depend on the spatial location of the experimental units, or the coordinates of observations themselves although that is not necessarily so. Zimmerman and Harville (1991) refer to [9.54] as a **random field linear model**. Since the spatial autocorrelation structure of $\delta(\mathbf{s})$ is modeled through a semivariogram or covariogram we take a direct approach to modeling spatial dependence rather than an autoregressive approach (in the vernacular of §9.3). This can be rectified with the earlier observation that data from field experiments are typically lattice data where autoregressive methods are more appropriate by considering each observation as concentrated at the centroid of the experimental unit (see Ripley 1981, p. 94, for a contrasting view that utilizes block averages instead of point observations).

The model for the semivariogram/covariogram is critically important for the quality of spatial predictions in kriging methods. In spatial random field models, where the mean function is of primary importance, it turns out that it is important to do a reasonable job at modeling the second order structure of $\delta(\mathbf{s})$, but as Zimmerman and Harville (1991) note, treatment comparisons are relatively insensitive to the choice of covariance functions (provided the set of functions considered is a reasonable one and that the mean function is properly

specified). Besag and Kempton (1986) found that inclusion of a nugget effect also appears to be unnecessary in many field-plot experiments.

Before proceeding further with random field linear models we need to remind the reader of the adage that *one modeler's random effect is another modeler's fixed effect.* Statistical models that incorporate spatial trends in the analysis of field experiments have a long history. In contrast to the random field models, previous attempts of incorporating the spatial structure focused on the mean function $\mu(\mathbf{s})$ rather than the stochastic component of the model. The term *trend analysis* has been used in the literature to describe methods that incorporate covariate terms that are functions of the spatial coordinates. In a standard RCBD analysis where Y_{ij} denotes the observation on treatment i in block j, the statistical model for the analysis of variance is

$$Y_{ij} = \mu + \rho_j + \tau_i + e_{ij},$$

where the experimental errors e_{ij} are uncorrelated random variables with mean 0 and variance σ^2. A trend analysis changes this model to

$$Y_{ij} = \mu + \tau_i + \theta_{kl} + e_{ij}, \qquad [9.55]$$

where θ_{kl} is a polynomial in the row and column indices of the experimental units (Brownie, Bowman, and Burton 1993). If r_k is the k^{th} row and c_l the l^{th} column of the field layout, then one may choose $\theta_{kl} = \beta_1 r_k + \beta_2 c_l + \beta_3 r_k^2 + \beta_4 c_k^2 + \beta_5 r_k c_l$, for example, a second-order response surface in the row and column indices. The difference to a random field linear model is that the deterministic term θ_{kl} is assumed to account for the spatial dependency between experimental units. It is a fixed effect. It does, however, appeal to the notion of a smooth-scale variation in the sense that the spatial trends move smoothly across block boundaries. The block effects have disappeared from model [9.55]. Applications of these trend analysis models can be found in Federer and Schlottfeldt (1954), Kirk, Haynes, and Monroe (1980), and Bowman (1990). Because it is assumed that the error terms e_{ij} remain uncorrelated they are not spatial random field models in our sense and will not be discussed further. For a comparison of trend and random field analyses see Brownie et al. (1993).

A second type of model that maintains independence of the errors are the nearest-neighbor models which are based on differencing observations with each other or by taking differences between plot yields and cultivar averages. The Papadakis nearest-neighbor analysis (Papadakis 1937), for example, calculates residuals between plot yields and arithmetic treatment averages in the East-West and North-South direction and uses these residuals as covariances in the mean model (the θ_{kl} part of the trend analysis model). The Schwarzbach analysis relies on adjusted cultivar means which are arithmetic means corrected for average responses in neighboring plots (Schwarzbach 1984).

In practical applications it may be difficult to choose between these various approaches to model spatial dependencies and to discriminate between different models. For example, changing the fixed effects trend by including or eliminating terms in a trend analysis will change the autocorrelation of the model residuals. Brownie and Gumpertz (1997) conclude that it is necessary to account for major spatial trends as fixed effects in the model but also that random field analyses are surprisingly robust to moderate misspecification of the fixed trend and retain a high degree of validity of tests and estimates of precision. The reason, in our opinion, is that a model which simultaneously models large- and small-scale stochastic trends is able – within limits – to capture omitted trends in the mean model through the spa-

tial dependency structure in the error process. Zimmerman and Harville (1991) refer to this effect as the covariance function "soaking up" spatial heterogeneity that would otherwise be fitted through fixed effects in the mean function. A trend analysis model or nearest-neighbor model that assumes that the mean function is correctly specified and the errors are uncorrelated will be invalid if the mean function is not modeled properly. There is nothing in the error structure that can "soak up" the ill-specification of the fixed effects.

9.5.2 Some Philosophical Considerations

Modeling data from a field experiment with random field methods seems like a win-win situation. The modeler can add or delete terms to the fixed effects part of the model that capture large-scale trends and let the covariance function of the error process $\delta(\mathbf{s})$ pick up any smooth-scale spatial variation of the omitted effects. As always, there is no free lunch and the analyst must be aware of the differences between modeling the data from a designed experiment vs. relying on randomization theory. The classical analysis of an experimental design stems from its underlying linear model which in turn is generated by the particular error-control, treatment, and observational designs. The ability to perform cause-and-effect inferences rests on these design components. Randomization ensures that the unaccounted effects — such as systematic spatial trends among the experimental units — are balanced out. This implies that expectations are reckoned over the randomization distribution of the design. In the Alliance, Nebraska wheat yield variety trial this distribution is formed by all possible arrangements of the 56 treatments to the $56 \times 4 = 224$ experimental units. The observed outcomes are considered fixed in the randomization approach. Assume, for a moment, that the three rightmost columns of experimental units in Figure 9.25 are systematically different from the other units. Should we take this into account in specifying the statistical model for the analysis or appeal to the fact that under randomization such effects are washed (balanced) out? There are three schools of thought:

1. Appeal to the randomization distribution because it allows causal inference. In effect, stick with the randomized complete block analysis. If it does not work out because blocking was carried out incorrectly, learn from the mistake and fix the problem the next time a variety trial with fifty-six treatments is conducted.

2. Do not appeal to the randomization distribution and model the variability and effects for *this* particular set of data. This is a modeling exercise determining which effects are modeled as part of the mean structure $\mathbf{X}(\mathbf{s})$ and which effects are "soaked up" by the error structure.

3. Appeal to randomization but also to the fact that stochastic elements beyond the randomization of treatments to units are at work. In developing the analysis appeal to a model where the errors are no longer independent and take expectation with respect to the joint distribution of randomization *and* the spatial process.

The three approaches differ in what is considered the correct model for analysis and how it is used. In (1) the correct model stems from the error-control, treatment, and observational design components. Treatment comparisons will always be unbiased under this approach, but can be inefficient if the design was not chosen carefully (as is the case in the Alliance-Nebraska case). In (2) the analyst is charged to develop a suitable model. Statistical inference

proceeds assuming that the selected model is correct. If a wrong model is used, treatment comparisons will be biased. Since there is never unshakable evidence that the final model is correct one can no longer make causal statements about the effect of treatments on the outcome. Statistical inference is *associative* rather than causal. The third approach is a mixture technique. It recognizes dependencies among the experimental units and the fact that treatments are randomly assigned to the units. Expectations of mean squares are calculated first over the randomization distribution conditional on the spatial process and then over the spatial process (see, for example, Grondona and Cressie 1991).

In spatial analyses the observed data are considered a realization of a random field and modeling the mean and dispersion structure proceeds in an observational manner. Whether a spatial model will provide a more efficient analysis will depend to what extent large-scale and small-scale trends are conducive to modeling. Besag and Kempton (1986) conclude that many agronomic experiments are not carried out in a sophisticated manner. The reasons may be convenience, unfamiliarity of the experimenter with more complex design choices, or tradition. We agree that it is hardly reasonable to conduct a field experiment with 56 treatments in a randomized complete block design. An incomplete block design or a resolvable, cyclic design may have been more appropriate. Nevertheless, many experiments are still conducted in this fashion. Bartlett (1938, 1978a) views analyses that emphasize the spatial context over the design context as ancillary devices to salvage efficiency in experiments that could have been designed more appropriately. Spatial random field models are more than salvage tools. They are statistical models that describe the variation in data, whether the data stem from a designed experiment or an observational study. By switching from a design-based analysis to one based on modeling, the ability to draw causal inferences is sacrificed, however.

9.5.3 Parameter Estimation

In matrix-vector notation model [9.54] can be written as

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \boldsymbol{\delta}(\mathbf{s}), \quad \boldsymbol{\delta}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \qquad [9.56]$$

and the parameters of the model to be estimated are $\boldsymbol{\phi} = [\boldsymbol{\beta}, \boldsymbol{\theta}]'$. $\boldsymbol{\theta}$ relates to the spatial dependency structure and $\boldsymbol{\beta}$ to the large-scale trend. As models for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ we usually consider covariograms that are derived from the isotropic semivariogram models in §9.2.2, keeping the number of parameters in $\boldsymbol{\theta}$ small. Because we work with covariances, it is assumed that the process is second-order stationary so that its covariogram is well-defined. Two general approaches to parameter estimation can be distinguished. Likelihood and likelihood-type methods which estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ simultaneously and least squares methods that estimate $\boldsymbol{\beta}$ given an externally obtained estimate of the spatial dependency.

Least Squares Methods

If $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ were known parameter estimates for $\boldsymbol{\beta}$ can be obtained by generalized least squares (GLS):

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}(\mathbf{s}).$$

Since $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is usually unknown we are faced with a similar quandary as in universal kriging.

Estimating $\boldsymbol{\theta}$ through semivariogram analysis requires detrending of the data, that is, an estimate of $\boldsymbol{\beta}$. Efficient estimation of $\boldsymbol{\beta}$ requires knowledge of $\boldsymbol{\theta}$. The usual approach is to

1. Assume $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\mathbf{I}$ and fit the model by ordinary least squares to obtain $\widehat{\boldsymbol{\beta}}_{OLS}$.
2. Obtain the OLS residuals $\widehat{\mathbf{e}}(\mathbf{s}) = \mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{OLS}$.
3. Fit a parametric, second-order stationary semivariogram based on the $\widehat{\mathbf{e}}(\mathbf{s})$ to obtain $\widehat{\boldsymbol{\theta}}$.
4. Use the estimates from the semivariogram fit to construct the $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ matrix.

These steps can (and should) be iterated, replacing OLS residuals in step 2. with GLS residuals after the first iteration. The final estimates of the mean parameters are estimated generalized least square estimates

$$\widehat{\boldsymbol{\beta}}_{EGLS} = \left(\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{Z}(\mathbf{s}). \qquad [9.57]$$

The same issues as in §9.4.4 must be raised here. The residuals lead to a biased estimate of the semivariogram of $\delta(\mathbf{s})$ and $\widehat{\boldsymbol{\beta}}_{OLS}$ is an inefficient estimator of the large-scale trend parameters. Since the emphasis in spatial random field linear models is often not on predicting but on estimation and hypothesis testing about $\boldsymbol{\beta}$ these issues are not quite as critical as in the case of universal kriging. If the results of a random field linear model analysis are used to predict $Z(\mathbf{s}_0)$ as a function of covariates and the spatial autocorrelation structure, the issues regain importance.

Likelihood Methods

Likelihood methods circumvent these problems because the mean and covariance parameters are estimated simultaneously. On the other hand they require distributional assumptions about $Z(\mathbf{s})$ or $\delta(\mathbf{s})$. If $\delta(\mathbf{s})$ is a Gaussian random field, then twice the negative log-likelihood of $\mathbf{Z}(\mathbf{s})$ is

$$\text{ϖ}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}(\mathbf{s})) = n\ln\{2\pi\} + \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + (\mathbf{z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}).$$

and the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_M$, $\widehat{\boldsymbol{\theta}}_M$ minimize this expression. This process is generally iterative and can be simplified by profiling the likelihood. This numerically efficient method can be applied if some parameters have a closed-form solution given the others. First consider $\boldsymbol{\theta}$ fixed and known. Minimizing $\text{ϖ}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}(\mathbf{s}))$ is then equivalent to minimizing $(\mathbf{z} - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})'\ \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z} - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})$. Since this is a generalized residual sum of squares, the maximum likelihood estimate of $\boldsymbol{\beta}$ (given $\boldsymbol{\theta}$) is

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}(\mathbf{s}).$$

The profiled (negative) log likelihood is obtained by substituting this expression back into $\text{ϖ}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}(\mathbf{s}))$ which is then only a function of $\boldsymbol{\theta}$ and is minimized with respect to $\boldsymbol{\theta}$. The resulting estimate $\widehat{\boldsymbol{\theta}}_M$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and the MLE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_M = \left(\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_M)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_M)^{-1}\mathbf{Z}(\mathbf{s}). \qquad [9.58]$$

The maximum likelihood ([9.58]) and estimated generalized least squares estimates [9.57] are very similar. They differ only in the covariance parameter estimate that is being substituted. To reduce the bias in maximum likelihood estimates of the covariance parameters it is again recommended to perform restricted maximum likelihood estimation. The REML estimates of the large-scale trend parameters are obtained as

$$\widehat{\boldsymbol{\beta}}_R = \left(\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_R)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_R)^{-1}\mathbf{Z}(\mathbf{s}). \qquad [9.59]$$

Software Implementation

The three methods, GLS, ML, and REML, lead to very similar formulas for the $\boldsymbol{\beta}$ estimates. The `mixed` procedure in The SAS® System can be used to obtain any one of the three. The spatial covariance structure of $\delta(\mathbf{s})$ is specified through the `repeated` statement of the procedure. In contrast to clustered data models in §7, all data points are potentially autocorrelated which calls for the `subject=intercept` option of the repeated statement.

Assume that an analysis of OLS residuals leads to an exponential semivariogram with practical range 4.5, partial sill 10.5, and nugget 2.0. The spatial coordinates of the data points are stored in variables `xloc` and `yloc` of the SAS data set. The mean model consists of treatment effects and a linear response surface in the coordinates. The following statements obtain the EGLS estimates [9.57], preventing `proc mixed` from iteratively updating the covariance parameters (`noiter` option of `parms` statement). The `noprofile` option prevents the profiling of an extra scale parameter from $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The Table of `Covariance Parameter Estimates` will contain three rows entitled `Variance`, `SP(EXP)`, and `Residual`. These correspond to the partial sill, the range, and the nugget effect, respectively. Notice that the parameterization of the exponential covariogram in `proc mixed` considers the range parameter to be one third of the practical range.

```
/* ------------------------------------------------------ */
/* Fit the model by EGLS for fixed covariogram estimates */
/* ------------------------------------------------------ */
proc mixed data=RFLMExample noprofile ;
  class treatment;
  model Z = treatment xloc yloc xloc*yloc / s;
  parms /* sill   */ ( 10.5 )
        /* range  */ ( 1.5  )
        /* nugget */ ( 2.0  ) / noiter;
  /* The local option of the repeated statement adds the */
  /* nugget effect                                       */
  repeated /subject=intercept local type=sp(exp)(xloc yloc);
run; quit;
```

Restricted maximum likelihood estimates are obtained in `proc mixed` with the statements

```
proc mixed data=RFLMExample noprofile ;
  class treatment;
  model Z = treatment xloc yloc xloc*yloc / s;
  parms /* sill   */ ( 6   to 12 by 2   )
        /* range  */ ( 0.5 to  3 by 1.5 )
        /* nugget */ ( 1   to  4 by 1.0 );
  repeated /subject=intercept local type=sp(exp)(xloc yloc);
run; quit;
```

The `noiter` option was removed from the `parms` statement which prompts the procedure to iteratively update the covariance parameter estimate $\boldsymbol{\theta}$. For each element of $\boldsymbol{\theta}$ a range of starting values is given. This can considerably speed up estimation, which can require formidable resources for large data sets. If the grid of starting values is too fine this is somewhat counterproductive as the procedure then has to evaluate many combinations of possible starting values before settling on the best set. The default estimation procedure for covariance parameter estimation is restricted maximum likelihood and the code example above yields $\widehat{\boldsymbol{\beta}}_R$ as in [9.59]. To obtain maximum likelihood estimates add the `method=ml` option to the `proc mixed` statement.

9.6 Autoregressive Models for Lattice Data

Box 9.10 Lattice Models

- **Models for spatial lattice data are close relatives of time series models.**
- **A lattice model commences with the user's definition of spatial connectivity among sites. This choice is then combined with an appropriate model for the marginal or conditional distribution of the data that is consistent with the neighborhood structure.**
- **Depending on whether the joint or conditional distribution of $Z(\mathbf{s})$ is being modeled, SSAR and CSAR models for lattice data are distinguished.**

9.6.1 The Neighborhood Structure

Lattice data are spatial data where the index set D is a fixed, discrete subset of $\mathbb{R}^2$ of countable points and $Z(\mathbf{s})$ is a random variable at location $\mathbf{s} \in D$. Examples of lattice data are observations made by census tract, county, or city blocks, data from field trials and remotely sensed images. Keeping with the literature on lattice data we call the locations $\mathbf{s} \in D$ the **sites** of the lattice. It is common to enumerate the countable set of sites in a lattice, for example, counties or census tracts can be numbered from 1 to n. Since the numbering in itself does not convey any spatial information it is necessary to define a location feature of each site such as the county center or the seat of the county government. On rectangular lattices (field experiments) the center of the unit is often used or experimental units can be identified by row and column number.

Modeling the spatial dependence among observations via the semivariogram or covariogram requires a smooth-scale spatial structure and a continuous spatial process. With lattice data other means of capturing the spatial dependence are needed. The notion of stationarity is of somewhat questionable value for processes operating on irregularly shaped area units or partitions (census tracts, counties, landscapes, regions, states, etc.). Even if there exists an underlying stationary continuous-space process, variances and covariances will not be the same for all areas if the observations arise from different area integrations. Stationary co-

variogram with range 88.754 m^2 fits the empirical semivariogram well. With a properly crafted mean model the block total $Z(A)$ can then be obtained by universal block-kriging. The estimate of the total amount of lead so obtained is 13.958 tons with a prediction standard error of 1.68 tons. A 95% prediction interval for the total is thus [10.66 tons, 17.26 tons]. The surface of the universal kriging predictions (Figure 9.44) differs little from back-transformed ordinary kriging predictions on the logarithmic scale (Figure 9.43).

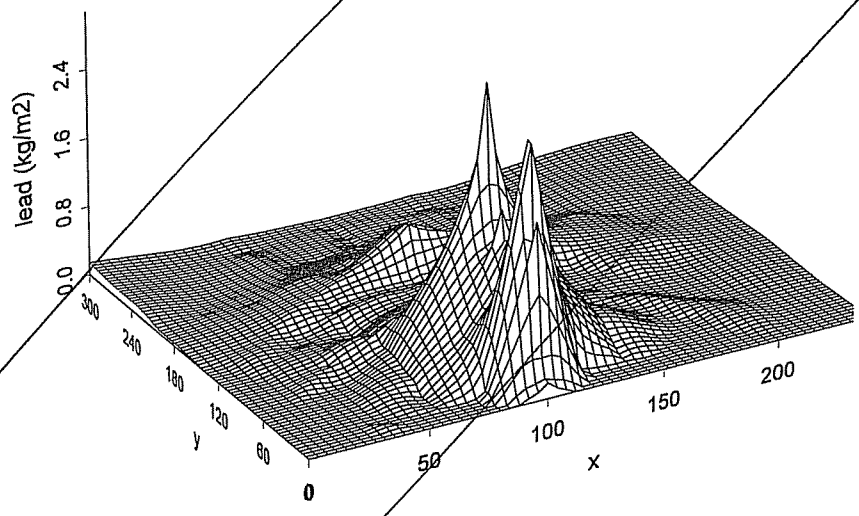

Figure 9.44. Predicted surface of lead in kg/m^2 obtained by universal kriging on the original scale.

9.8.4 Spatial Random Field Models — Comparing C/N Ratios among Tillage Treatments

When data are collected under different conditions, such as treatments, an obvious question is to determine whether the conditions are different from each other, and if so, how the differences manifest themselves. In a classical field experiment contrasts among the treatment means are estimated and tested to formulate statements about the differences among and effects of the experimental conditions. If the data collected under various conditions are autocorrelated, then one needs to rethink what precisely we mean by *differences* in the conditions. We now return to the soil carbon data first introduced in §9.8.2. After ten years of a corn-soybean rotation without tillage, intermediate strips of the field were chisel-plowed. Two months after the soils were first chisel-plowed in the spring samples from 0 to 2 inch depths were collected and total N percentage (TN) and total carbon percentage (CN) were determined. The sampling locations and the strips are shown in Figure 9.45.

Since sampling occurred very soon after tillage we do not anticipate fundamental changes in the TC and TN values or the C/N ratio between the two treatments. Because of the spatial sampling context and the presence of two conditions on the field, however, the data are perfectly suited to demonstrate the basic manipulations and computations involved in a random field analysis that involves treatment structure. We furthermore note from Figure 9.45 that the strips were not randomized. An analysis as a randomized experiment with subsampling of six replications of two treatments is therefore tenuous. Instead, we analyze the data as

a spatial random field with a mean structure given by the two treatment conditions and possible spatial autocorrelation among the sampling sites.

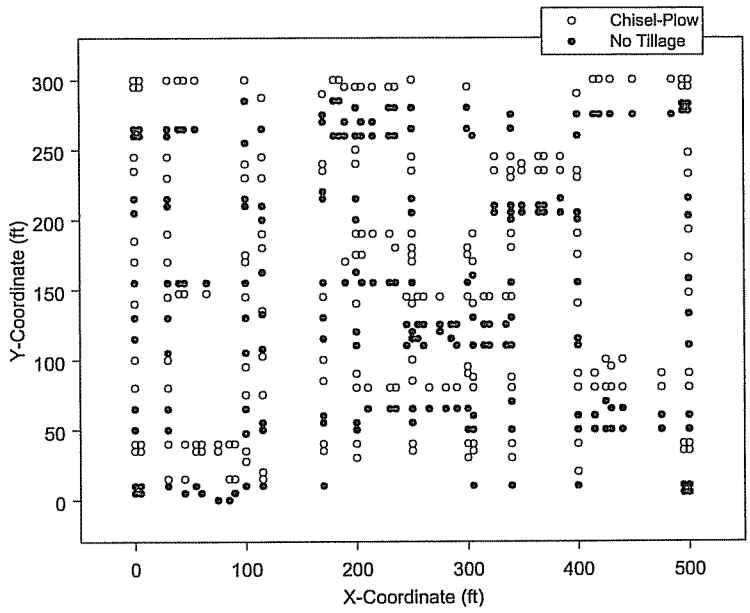

Figure 9.45. Sampling locations at which total soil N (%) and total soil C (%) were observed for two tillage treatments. Treatment strips are oriented in East-West direction.

The target attribute for this application is the C/N ratio and a simplistic pooled t-test comparing the two tillage treatments leads to a p-value of 0.809 from which one would conclude that there are no differences in the average C/N ratios. This test does not account for spatial autocorrelation treating the 195 samples on chisel-plow strips and 200 samples on no-till strips as independent. Furthermore, it does not convey whether there are differences in the spatial structure of the treatments. Even if the means are the same the spatial dependency might develop differently. This, too, would be a difference in the treatments that should be recognized by the analyst. Omnidirectional semivariograms were calculated with the `variogram` procedure in The SAS® System and spherical semivariogram models were fit to the empirical semivariograms (Figure 9.46) with `proc nlin` by weighted least squares:

```
proc sort data=CNRatio; by tillage; run;
proc variogram data=CNRatio outvar=svar;
   compute lagdistance=13.6 maxlag=19 robust;
   coordinates xcoord=x ycoord=y;
   var cn;
   by tillage;
run;
proc nlin data=fitthis nohalve method=newton noitprint;
  parameters sillC=0.093 sillN=0.1414 rangeC=116.6 rangeN=197.2
             nugget=0.1982;
  if tillage='ChiselPlow' then
    sphermodel = nugget + (distance <= rangeC)*sillC*(1.5*(distance/rangeC) -
                 0.5*((distance/rangeC)**3)) + (distance >  rangeC)*sillC;
  else
    sphermodel = nugget + (distance <= rangeN)*sillN*(1.5*(distance/rangeN) -
                  0.5*((distance/rangeN)**3)) + (distance >  rangeN)*sillN;
  model rvario = sphermodel;
  _weight_ = 0.5*count/(sphermodel**2);
run;
```

In anticipation of obtaining generalized least squares and restricted maximum likelihood inferences in `proc mixed` a common nugget effect was fit for both tillage treatments but the sills and ranges of the semivariogram were varied. The sill and range estimates for the chisel-plow treatment were 0.092 and 127.0, respectively. The corresponding estimates for the no-till treatment were 0.1397 and 199.2 (Output 9.13). Notice that to a considerable degree variability in C/N ratios is due to the nugget effect. The relative structured variability is 31% for the chisel-plow and 41% for the no-till treatment. The C/N ratio of the undisturbed no-till sites is more spatially structured, however, as can be seen from the larger range.

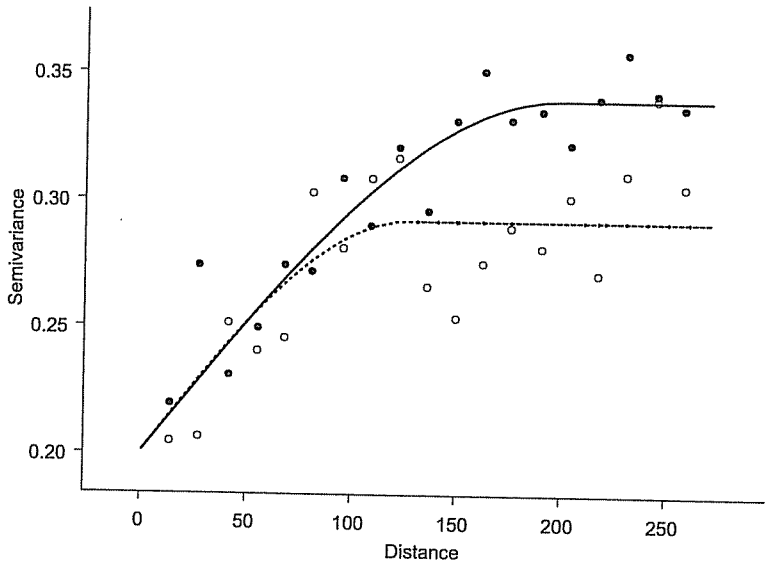

Figure 9.46. Omnidirectional empirical semivariograms for C/N ratio under chisel-plow (open circles) and no-till (full circles) treatments. Weighted least squares fit of spherical semivariograms are shown.

Output 9.13. (abridged)

The NLIN Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	5	13117.5	2623.5	23.92	<.0001
Residual	33	53.2243	1.6129		
Uncorrected Total	38	13170.7			
Corrected Total	37	207.6			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
sillC	0.0920	0.0151	0.0612	0.1228
sillN	0.1397	0.0152	0.1089	0.1706
rangeC	127.0	19.4203	87.4950	166.5
rangeN	199.2	29.7131	138.8	259.7
nugget	0.2000	0.0139	0.1717	0.2284

Next we obtain generalized least squares estimates of the treatment effect as well as predictions of the C/N ratio over the entire field with `proc mixed` of The SAS® System. A data set containing the prediction locations for both treatments (data set `filler`) is created

and appended to the data set containing the observations. The response variable of the filler data set is set to missing values. This will prevent `proc mixed` from using the information in the prediction data set for estimation. In calculating predicted values these observations can be used, however, since they contain all information apart from the response.

```
data filler;
  do tillage='ChiselPlow','NoTillage';
    do x = 0 to 500 by 10;  do y = 0 to 300 by 10; cn=.; output;  end; end;
  end;
run;
data fitthis; set filler cnratio; run;

proc mixed data=fitthis noprofile;
  class tillage;
  model CN = tillage /ddfm=contain outp=p;
  repeated / subject=intercept type=sp(sph)(x y) local group=tillage;
  parms /* sill  ChiselPlow */    0.0920
        /* range ChiselPlow */  127.0
        /* sill  NoTillage  */    0.1397
        /* range NoTillage  */  199.2
        /* nugget (common)  */    0.2000 /  noiter;
run;
```

The call to `proc mixed` has several important features. The `model` statement describes the mean structure of the model. C/N ratios are assumed to depend on the tillage treatments. The `outp=p` option of the `model` statement produces a data set (named `p`) containing the predicted values. The `repeated` statement identifies the spatial covariance structure to be spherical (`type=sp(sph)(x y)`). The `subject=intercept` option indicates that the data set comprises a single subject, all observations are assumed to be correlated. The `group=tillage` option requests that the spatial covariance parameters are varied by the values of the `tillage` variable. This allows modeling separate covariance structures for the chisel-plow and no-till treatments to reflect the differences in spatial structure evident in Figure 9.46. Finally, the `local` option adds a nugget effect. Since `proc mixed` adds only a single nugget effect, it was important in fitting the semivariograms to ensure that the nugget effect was held the same for the two treatments. The `parms` statement provides starting values for the covariance parameters. The order in which the values are listed equals the order in which the values appear in the `Covariance Parameter Estimates` table of the `proc mixed` output. A trial run is sometimes necessary to determine the correct order. The starting values are set at the converged iterates from the weighted least squares fit of the theoretical semivariogram (Output 9.13). The `noiter` option of the `parms` statement prevents iterations of the covariance parameters and holds them fixed at the starting values provided. To produce restricted maximum likelihood estimates of the covariance parameters, simply remove the `noiter` option. The `noprofile` option of the `proc mixed` statement prevents profiling of the nugget variance. Without this option `proc mixed` would make slight adjustments to the sill and nugget even if the `/noiter` option is specified.

The `Dimensions` table indicates that 395 observations were used in model fitting and 3162 observations were not used (Output 9.14). The latter comprise the `filler` data set of prediction locations for which the `CN` variable was assigned a missing value. The `-2 Res Log Likelihood` of 570.3 in the table of `Fit Statistics` equals minus twice the residual log likelihood in the `Parameter Search` table. The latter table gives the likelihood for all sets of starting values. Here only one set of starting values was used and the equality of the `-2 Res Log Likelihood` values shows that no iterative updates of the covariance parameters took place. The estimates shown in the `Covariance Parameter Estimates` table are identical to the

starting values provided in the `parms` statement. Finally, the `Type 3 Tests of Fixed Effects` table shows that there is no significant difference between the mean C/N ratios of the two tillage treatments ($p = 0.799$).

Output 9.14.

The Mixed Procedure

Model Information

Data Set	WORK.FITTHIS
Dependent Variable	cn
Covariance Structures	Spatial Spherical, Local Exponential
Subject Effect	Intercept
Group Effect	tillage
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
tillage	2	ChiselPlow NoTillage

Dimensions

Covariance Parameters	5
Columns in X	3
Columns in Z	0
Subjects	1
Max Obs Per Subject	395
Observations Used	395
Observations Not Used	3162
Total Observations	3557

Parameter Search

CovP1	CovP2	CovP3	CovP4	CovP5	-2 Res Log Like
0.09200	127.00	0.1397	199.20	0.2000	570.2618

Covariance Parameter Estimates

Cov Parm	Subject	Group		Estimate
Variance	Intercept	tillage	ChiselPlow	0.09200
SP(SPH)	Intercept	tillage	ChiselPlow	127.00
Variance	Intercept	tillage	NoTillage	0.1397
SP(SPH)	Intercept	tillage	NoTillage	199.20
Residual				0.2000

Fit Statistics

-2 Res Log Likelihood	570.3
AIC (smaller is better)	570.3
AICC (smaller is better)	570.3
BIC (smaller is better)	570.3

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tillage	1	393	0.06	0.7990

The predicted C/N surfaces for the two tillage treatments are shown in Figure 9.47. Both surfaces vary about the same mean but the greater spatial continuity (larger range) of the no-till sites is evident in a smoother, less variable surface. Positive autocorrelations are stronger over the same distance under this treatment as compared to the chisel-plow treatment. At this point it is worthwhile to revisit the question raised early in this application. What do we mean by *differences* in experimental conditions if the observations collected from each site have a spatial context? There is no difference in the average C/N values in this study as can be expected when sampling only two months after installment of the treatments. There appear to be differences in the spatial structure of the treatments, however. Fitting a single spherical semivariogram to the empirical semivariograms shown in Figure 9.46 a residual sum of square of 93.09 on 35 degrees of freedom is obtained. A sum of square reduction test leads to

$$F_{obs} = \frac{(93.09 - 53.2)/2}{53.2/33} = 12.37$$

with a p-value of 0.00009. If the semivariogram is estimated by ordinary (instead of weighted) least squares the statistics are $F_{obs} = 11.85$ and $p = 0.0001$. There are significant differences among the treatments in the autocorrelation structure, albeit not in the average C/N ratio. One can argue that after ten years of continuous no-till management there is greater continuity in the C/N ratios compared to what can be observed shortly after a disturbance through plowing.

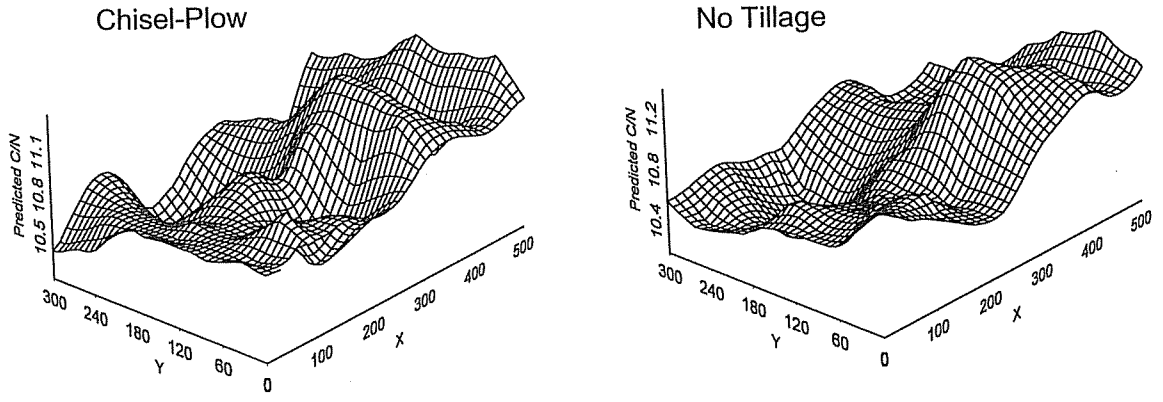

Figure 9.47. Predicted C/N surface under chisel-plow and no-till treatments.

The predicted surfaces in Figure 9.47 were obtained from the generalized least squares fit which assumed that the supplied starting values of the covariance parameters are the true values. This is akin to the assumption in kriging methods that the semivariogram values used in solving the kriging equations are known. Removing the `noiter` option of the `parms` statement in `proc mixed` the spatial covariance parameters are updated iteratively by the method of restricted maximum likelihood. Twice the negative residual log likelihood at convergence can be compared to the same statistic calculated from the starting values. This likelihood ratio test indicates whether the REML estimates are a significant improvement over the starting values. The `mixed` procedure displays the result of this test in the `PARMS Model Likelihood Ratio`

`Test` table (Output 9.15). In this application convergence was achieved after twelve time-consuming iterations with no significant improvement over the starting values ($p = 0.1981$).

Output 9.15. (abridged)

```
                The Mixed Procedure

                  Fit Statistics
-2 Res Log Likelihood                562.9
AIC (smaller is better)              572.9
AICC (smaller is better)             573.1
BIC (smaller is better)              592.8


 PARMS Model Likelihood Ratio Test
   DF     Chi-Square       Pr > ChiSq
    5           7.32           0.1981
```

9.8.5 Spatial Random Field Models — Spatial Regression of Soil Carbon on Soil N

In the previous application C/N ratio was modeled directly and compared between the two tillage treatments. In many applications one attribute emerges as the primary variable of interest and other variables are secondary attributes which are to be linked to the primary attribute. This approach is particularly meaningful if the secondary attributes are easy to measure or available in dense coverage (e.g., sensed images, GIS) and the primary attribute is more difficult to determine. If the relationship between primary and secondary attributes can be modeled for a particular data set where both variables have been measured, the model can then be applied to situations where only the secondary attributes are available. Consider that we are interested in predicting soil carbon as a function of soil nitrogen. From Figure 9.48 it is clearly seen that the relationship between TC and TN is very strong ($R^2 = 0.916$), close to linear, and differs not between the two tillage treatments.

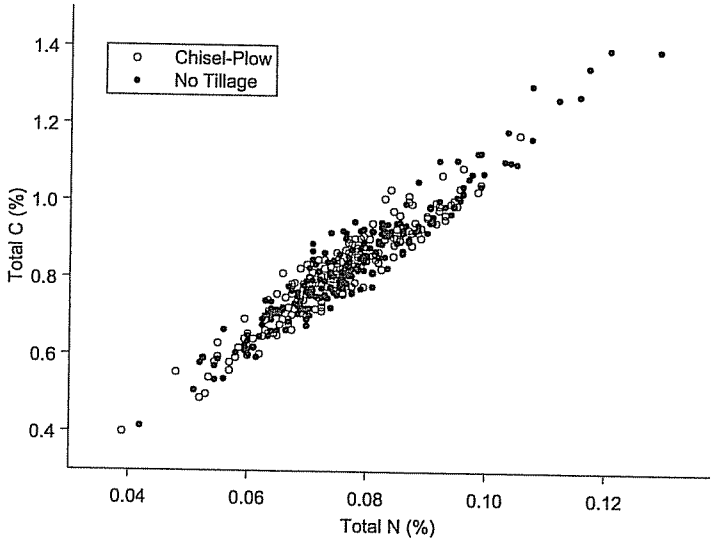

Figure 9.48. Relationship between total C (%) and total N (%) of chisel-plow and no-till areas.

CONTEMPORARY STATISTICAL MODELS

for the Plant and Soil Sciences

Oliver Schabenberger
Francis J. Pierce

2002

CRC PRESS

Boca Raton London New York Washington, D.C.