Latin Hypercube Sampling of Gaussian Random Fields

Edzer J. PEBESMA and Gerard B. M. HEUVELINK

Faculty of Environmental Sciences
University of Amsterdam
1018 VZ Amsterdam
The Netherlands
(*e.pebesma@geog.uu.nl*)

Following the method of Stein, this article shows how a Latin hypercube sample can be drawn from a Gaussian random field. In a case study the efficiency of Latin hypercube sampling is compared experimentally to that of simple random sampling. The model outputs studied are the mean and the 5- and 95-percentile of the areal fraction where point concentration of zinc in the topsoil exceeds a given threshold. The Latin hypercube sampling procedure slightly distorts the short-distance correlation, and in an artificial example, it is shown that this distortion is modest for small samples and vanishes for large samples.

KEY WORDS: Environmental model; Geostatistical simulation; Monte Carlo; Risk analysis; Uncertainty analysis.

Many process models in the earth and environmental sciences need spatially distributed values for characteristics of an area (or a volume) of the earth as input. For instance, the hydraulic conductivity and transmissivity at all locations in an aquifer are relevant parameters for modeling groundwater flow (Kitanidis and Vomvoris 1983). Most often, the values of these spatially distributed input variables (or parameters) are not known everywhere: In some cases, measurements are available on a limited number of locations; in other cases, only a rough guess of the variables (e.g., a range of values) is available. It may then be relevant to study the effect of the uncertainty about input variables on the model output to get an idea of the error bounds of the model output due to this uncertainty, or of the risk of making wrong decisions, based on the model output.

Uncertainty in the output resulting from uncertainty in the input can be studied by a Monte Carlo analysis. This works as follows: A sample consisting of many realizations of the input variables is created and used as input to run the model. The collection of all model outputs is then taken as representative of the output probability distribution function (PDF). The sample size determines how well the outcomes represent the true PDF: The larger the sample, the more accurate the information. Other, for instance, analytical methods for calculating uncertainty of model output resulting from uncertainty in model inputs (e.g., Heuvelink 1998) may be preferred in special cases (e.g., for some simple models) but they will not be considered here.

For certain problems, it is sufficient to perform the Monte Carlo analysis for each spatial location separately. For instance, this will be the case when the model is a point location model that only uses input variable values at point locations and when the spatial correlation of the output is of no interest. For problems in which dependencies between input variables at nearby locations are relevant because the model includes spatial interactions (e.g., spatial averaging, transport of matter) or where the spatial correlation of model output is of interest, however, it is necessary to perform an integrated spatial Monte Carlo analysis. This means that the input is characterized by a full spatial PDF (the multivariate PDF for the variable at all spatial locations). In general, the spatial PDF should incorporate spatial correlation because it is present in most real-world variables.

A multivariate PDF that is often used to characterize continuous spatial variables is the Gaussian random field (Cressie 1993). One reason for this model's popularity is that normal (Gaussian) theory is thoroughly worked out and often simple in form. Recently, this model has received considerable criticism when used to model permeabilities (Gómez-Hernández 1997; Journel 1997). Still for many other purposes, including the case study that we will present in this article, the Gaussian random field remains a reasonable choice.

For the input to the Monte Carlo analysis, the usual way to create multiple realizations of a spatial random field is to draw a simple random sample: Subsequent realizations are drawn independently (Gómez-Hernández and Journel 1993; Gotway 1994). Taking a larger sample of realizations will allow more accurate estimation of the output PDF. If the run time of each Monte Carlo run is short, then taking a large sample suffices to make the sampling error negligibly small.

When each model run is relatively expensive (e.g., Smith and Freeze 1979; Beven and Binley 1992; Thiele, Rao, and Blunt 1996), the sample size has to be kept small for practical reasons. In that case, more accurate assessment of the output PDF can be obtained when more efficient sampling methods, such as stratified random sampling or

TECHNOMETRICS, NOVEMBER 1999, VOL. 41, NO. 4

its multivariate version, Latin hypercube sampling (McKay, Conover, and Beckman 1979; Ross 1990), are used.

In this article we first recall how a Latin hypercube sample is drawn from independent and dependent random vectors. We then show how a Latin hypercube sample is obtained from a Gaussian random field. We discuss the specific problems of the method, and we use a case study to demonstrate the efficiency gain compared to simple random sampling. In an artificial example, the bias in spatial correlation introduced by the Latin hypercube sampling is quantified. A discussion concludes this article.

1. LATIN HYPERCUBE SAMPLING

1.1 Independent Variables

Latin hypercube sampling (LHS) is a stratified random sampling technique in which a sample of size N from multiple (continuous) variables is drawn such that for each individual variable the sample is (marginally) maximally stratified (McKay et al. 1979). A sample is maximally stratified when the number of strata equals the sample size N and when the probability of falling in each of the strata equals N^{-1}. The classic figure with the Latin hypercube sample drawn from two independent uniform $U[0,1]$ variables is shown in Figure 1: The number of categories per variable equals the sample size (6), each row or column contains one element, and the width of rows and columns is 1/6.

To draw a Latin hypercube sample of size N for K independent variables, the ith sample element for variable j is obtained as

$$z_{ij} = F_j^{-1}((p_{ij} - \xi_{ij})/N), \tag{1}$$

where F_j is the cumulative distribution function of variable j, where, for each $j = 1,\ldots,K, p_{ij}$ $(i = 1,\ldots,N)$ is a random permutation of $1,\ldots,N; \xi_{ij}$ is $U[0, 1]$, a uniform distributed random number between 0 and 1; and the K permutations and the NK uniform variates ξ_{ij} are mutually independent. An example is given in Table 1.

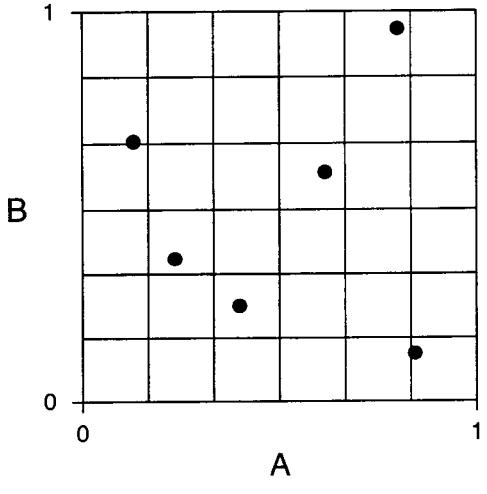

Figure 1. A Latin Hypercube Sample With N = 6 for Two Independent Uniform U[0, 1] Random Variables, A and B.

Table 1. Numerical Example of a Latin Hypercube Sample of Size 5 From Two Independent Gaussian Distributions, $\mathcal{N}(0, 1)$ (j = 1) and $\mathcal{N}(5, 2)$ (j = 2)

i	j	p_{ij}	ξ_{ij}	$p_{ij} - \xi_{ij}$	$(p_{ij} - \xi_{ij})/N$	z_{ij}
1	1	2	.80	1.20	.24	−.71
2	1	1	.54	.46	.09	−1.32
3	1	5	.02	4.98	1.00	2.64
4	1	4	.56	3.44	.69	.49
5	1	3	.25	2.75	.55	.12
1	2	2	.51	1.49	.30	3.94
2	2	5	.39	4.61	.92	7.84
3	2	4	.41	3.59	.72	6.16
4	2	3	.29	2.71	.54	5.21
5	2	1	.86	.14	.03	1.16

1.2 Dependent Variables

Stein (1987) presented a method for drawing a Latin hypercube sample from dependent variables. His procedure is as follows:

1. Obtain a simple random sample from the multivariate PDF, $s_{ij}, i = 1,\ldots,N, j = 1,\ldots,K$.
2. Let $s_j = (s_{1j},\ldots,s_{Nj})$ be the sample corresponding to the jth variable, and let r_j be the vector with the ranks of s_j (a permutation of $1,\ldots,N$) for $j = 1,\ldots,K$.
3. Draw NK random variates ξ_{ij} independently from $U[0, 1]$.
4. Obtain the Latin hypercube sample z_{ij} $(i = 1,\ldots, N, j = 1,\ldots,K)$ by substituting r_{ij}, the ith element of r_j, for p_{ij} in Equation (1).

Given a simple random sample from the target multivariate distribution, a Latin hypercube sample from (approximately) the same distribution is obtained by slightly shifting the sample elements. During this shift, the ranks of the sample elements (and the rank correlation) are preserved: They remain identical to the ranks in the simple random sample. For a bivariate normal distribution an example is given in Table 2, and a graphical illustration is given in Figure 2.

Clearly, step 4 ensures that the marginal distributions in the Latin hypercube sample are preserved. Stein (1987) also showed that asymptotically (for large N) the procedure yields approximately the correct joint distribution. For small N, however, meaningful deviations may occur.

Table 2. Numerical Example of a Latin Hypercube Sample of Size 5 From Two Correlated $\mathcal{N}(0, 1)$ Gaussian Distributions (correlation is .9)

i	j	s_{ij}	r_{ij}	ξ_{ij}	$r_{ij} - \xi_{ij}$	$(r_{ij} - \xi_{ij})/N$	z_{ij}
1	1	.03	3	.48	2.52	.50	.01
2	1	1.18	5	.94	4.06	.81	.89
3	1	−.13	2	.34	1.66	.33	−.44
4	1	−1.18	1	.15	.85	.17	−.96
5	1	.14	4	.69	3.31	.66	.42
1	2	.56	4	.18	3.82	.76	.72
2	2	1.06	5	.14	4.86	.97	1.92
3	2	−.67	2	.59	1.41	.28	−.57
4	2	−1.57	1	.01	.99	.20	−.85
5	2	.51	3	.23	2.77	.55	.14

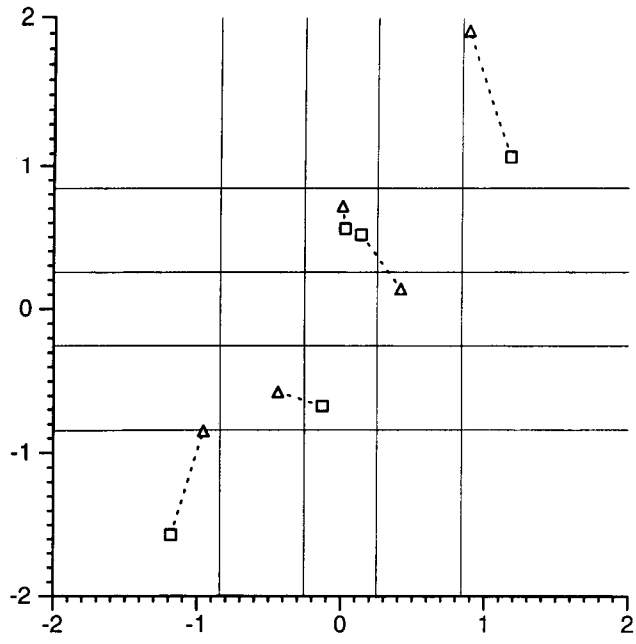

Figure 2. Resampling a Simple Random Sample (□) From a Bivariate Normal Distribution to a Latin Hypercube Random Sample (△), Preserving Ranks. Both variables (Table 2: z_{i1} on x axis, z_{i2} on y axis) are marginal $\mathcal{N}(0, 1)$; their correlation is .9. Dotted lines indicate the shift for individual sample elements, thin lines indicate stratum boundaries.

2. STOCHASTIC SIMULATION OF GAUSSIAN RANDOM FIELDS

2.1 Simple Random Sampling

In geostatistics (Journel and Huijbregts 1978; Cressie 1993), spatial data are considered as a realization of a random field $Z(\cdot) = \{Z(x)|x \in D\}$. Here, we assume that $Z(\cdot)$ is a Gaussian random field; that is we assume that for any finite set of locations $x_1, \ldots, x_m$, the $Z(x_1), \ldots, Z(x_m)$ are jointly Gaussian distributed.

In practice, stochastic simulation of $Z(\cdot)$ will be confined to simulating $Z(\cdot)$ at a finite number of K point locations in D, usually on a regular grid. Consequently, any method that is capable of sampling from a multivariate Gaussian distribution can also be used for the simulation of a Gaussian random field. One method that is especially useful is sequential simulation (Johnson 1987), which works as follows. Starting with a (possibly empty) set of observations $\{z(x_1), \ldots, z(x_m)\}$, we simulate the remaining values by repeating the following steps:

1. Go to an unvisited simulation location, and call this x_{m+1}.
2. Calculate the best linear predictor at $x_{m+1}, \hat{z}(x_{m+1})$, and the associated prediction error variance $\sigma^2(x_{m+1})$.
3. Draw a random value $\tilde{z}(x_{m+1})$ from $\mathcal{N}(\hat{z}(x_{m+1}), \sigma^2(x_{m+1}))$, the normal distribution with mean $\hat{z}(x_{m+1})$, and variance $\sigma^2(x_{m+1})$.
4. Add this value to the dataset (let $m = m + 1$) and return to step 1.

Repeat until there are no unvisited simulation locations left.

The best linear predictor at a location x is

$$\hat{z}(x) = \mu(x) + c^T C^{-1}(z - \mu), \tag{2}$$

where $\mu(x) = E(Z(x)), c^T = (\text{cov}(Z(x_1), Z(x)), \ldots, \text{cov}(Z(x_m), Z(x)))$, element i, j of matrix C is $\text{cov}(Z(x_i), Z(x_j)), z = (z(x_1), \ldots, z(x_m))^T$, and $\mu = (\mu(x_1), \ldots, \mu(x_m))^T$. The prediction error variance is given as

$$\sigma^2(x_{m+1}) = \text{var}(Z(x_{m+1})) - c^T C^{-1} c. \tag{3}$$

Note that the procedure assumes that the mean and covariance of $Z(\cdot)$ are known.

When simulating large grids, at some stage the dataset becomes too large for calculating $c^T C^{-1}$. To avoid this, in practice the best linear predictor is approximated by using only a subset of the nearest (most correlated) data in (2) and (3) (Deutsch and Journel 1992). This local approximation may introduce spurious structures in the simulated fields when the simulation locations are visited in a regular sequence (e.g., rowwise in a grid). To avoid such structures the sequence is usually randomized.

One realization is a sample with size $N = 1$ from the multivariate distribution of the random field. Larger samples contain multiple realizations and can be generated by repeating the preceding procedure over and over. Computationwise, it is more efficient, however, to generate all required simulations during a single pass over the simulation locations. Then, the "expensive" results ($c^T C^{-1}$ and the data selection in a local neighborhood) need to be calculated only once for each simulation location (Journel 1989).

When the simulation starts with no data, the realizations are *unconditional*: They are not conditioned on available data and therefore, except for having the same mean and spatial covariance, they are completely independent (e.g., top row, Fig. 3). When the simulation starts with a set of observations and spatial correlation was present, all realizations will more or less follow the pattern already present in the observation data (Fig. 4b and second row Fig. 3). With respect to the ensemble of possible unconditional realizations, these realizations are more alike than unconditionally simulated realizations, but *conditional on the data*, they are independent (this follows from step 3 of the preceding list).

At any location x the ensemble of possible realizations is (marginally) distributed as

$$\mathcal{N}(\hat{z}(x), \sigma^2(x)), \tag{4}$$

where $\hat{z}(x)$ and $\sigma^2(x)$ are [the simple kriging mean and variance (Journel and Huijbregts 1978)] obtained by applying (2) and (3) to the observed data only. Clearly, in absence of observed data (unconditional simulation) this will be $\mathcal{N}(\mu(x), \text{var}(Z(x)))$.

2.2 Latin Hypercube Sampling

We now describe how the procedure for drawing a Latin hypercube sample from dependent variables can be applied to the simulation of random fields. We describe the procedure for a single random field, but the generalization to multiple interdependent random fields is straightforward. Basically, the procedure is identical to that in Section 1:

1. Create a simple random sample of size N of the random field [i.e., generate N independent realizations of $Z(\cdot)$].

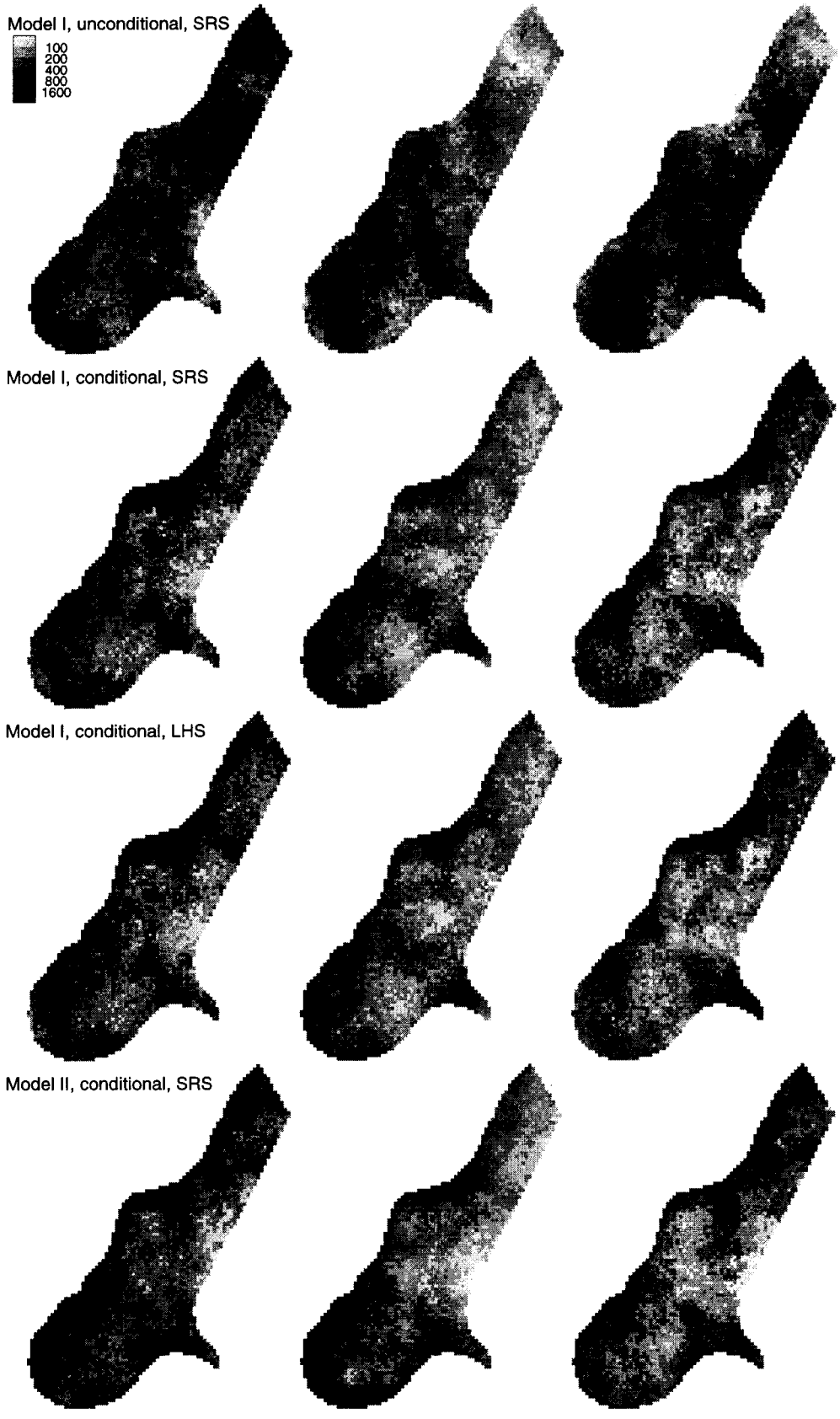

Figure 3. Independent Unconditional Simulations of a Gaussian Random Field (top row); Independent Conditional Simulations of a Gaussian Random Field Using Simple Random Sampling (second row); Latin Hypercube Sample Elements Corresponding to the Conditional Simulations in the Second Row (third row), and Independent Conditional Simulations from Model II (bottom row). Conditioning data are shown in Figure 4b. Model I assumes a constant and known mean; Model II assumes an unknown mean that is a function of distance to the River Meuse.

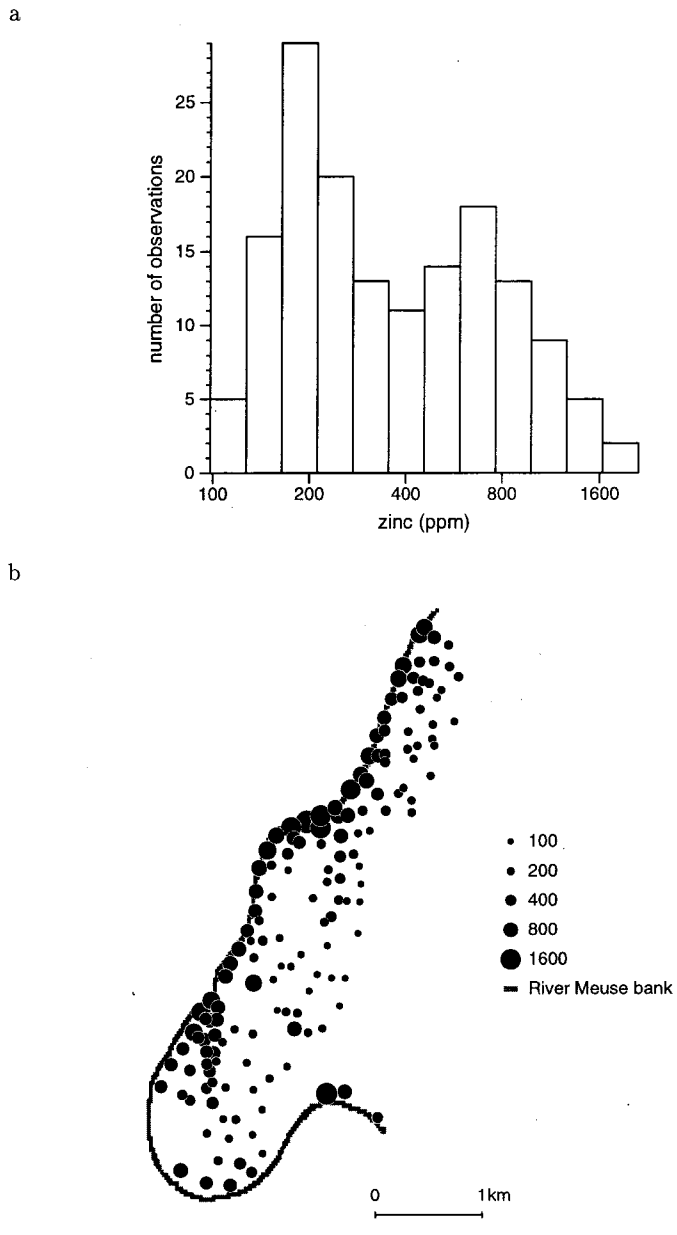

Figure 4. Zinc Concentrations (ppm) Measured Along the River Meuse Near Meers, The Netherlands: Histogram (a) and Dot Map (b).

2. Let $v(x_j) = (v_1(x_j), \ldots, v_N(x_j))$ be the sampled values at location x_j, and let $r(x_j)$ be the vector with the ranks of $v(x_j)$ (a permutation of $1, \ldots, N$) for $j = 1, \ldots, K$.

3. Draw NK random variates ξ_{ij} independently from $U[0, 1]$.

4. Obtain the Latin hypercube sample $z_i(x_j)$ at location x_j as $z_i(x_j) = F^{-1}_{Z(x_j)}((r_i(x_j) - \xi_{ij})/N), i = 1, \ldots, N, j = 1, \ldots, K$, with $r_i(x_j)$ the ith element in $r(x_j)$.

Thus, all that is required to obtain a Latin hypercube sample is (a) a simple random sample of the random field and (b) the (inverse) marginal distributions $F^{-1}_{Z(x_j)}$. For a Gaussian random field, a simple random sample and the marginal distributions are easily obtained. As an example, three elements from a sample of size $N = 20$ are given in Figure 3: The second row shows three elements of the simple random sample; the third row shows the corresponding Latin hypercube sample elements.

3. CASE STUDY: ZINC CONCENTRATIONS ALONG THE RIVER MEUSE

Consider topsoil zinc concentrations measured in an area along the River Meuse (Fig. 4; Burrough and McDonnell 1998). Clearly, the zinc concentrations tend to decrease when moving from the river. A plot of measured zinc concentration versus the square root of distance to the river (Fig. 5) suggests a linear relation at the log scale. Because topsoil zinc concentration is related to flooding of the river banks, we would preferably try to explain variation in measured zinc concentration by variables such as flood frequency or topographic level. This information is not readily available, however, and distance to the river is used as a substitute.

Let the variable $z(x)$ be the zinc concentration at location $x \in D$, with D the study area. We will consider two different models for the spatial variation of $z(\cdot)$. Model I is the simpler model: It assumes that the log-concentrations are a realization of a Gaussian random field and that the mean zinc concentration μ is spatially invariant and known (taken as the unweighted sample mean):

$$\log(z(x)) = \mu + e(x),$$

with $e(x)$ a zero-mean second-order stationary residual. Model II treats the spatial variation as the sum of a trend, taken as a first-order linear model in the square root of distance to the River Meuse $M(x)$, and a zero-mean second-order stationary residual,

$$\log(z(x)) = \beta_0 + \beta_1\sqrt{M(x)} + e(x),$$

with β_0 and β_1 unknown constants.

The semivariograms (Journel and Huijbregts 1978) of the log-measurements (Model I) and the predicted regression

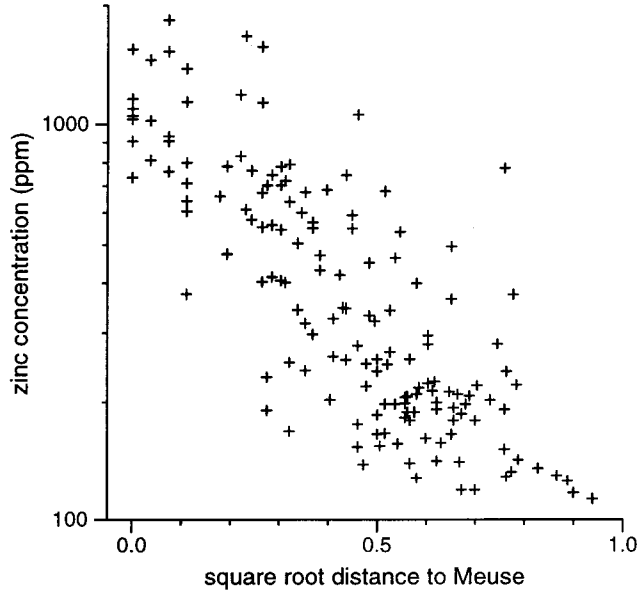

Figure 5. Scatterplot of Zinc Concentrations and Square Root of Distance to the River Meuse (normalized).

residuals (Model II) are shown in Figure 6. The semivariogram models fitted to the log-measurements is the spherical model

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ .054 + .58\left(\frac{3}{2}\,\frac{h}{893} - \frac{1}{2}\left(\frac{h}{893}\right)^3\right) & \text{if } 0 < h \le 893 \\ .054 + .58 & \text{if } h > 893, \end{cases}$$

and that of the residuals from Model II is the exponential model

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ .036 + .19\left(1 - \exp\left(\frac{-h}{254}\right)\right) & \text{if } h > 0. \end{cases}$$

The semivariogram models were fitted to the sample semivariograms using weighted least squares (Cressie 1993; Pebesma and Wesseling 1998). As a check, the semivariogram model from the ordinary least squares (OLS) predicted residuals was used to obtain generalized least squares (GLS) predicted residuals (Cressie 1993). The semivariogram fitted to the latter residuals was identical to the one obtained from OLS predicted residuals. Covariances between data pairs are derived from semivariogram models when the observations (Model I) or the residuals (Model II) are modeled as a second-order stationary random field.

3.1 Simulation of $Z(x)$

For the conditional simulation of a simple random sample of a Gaussian random field with known mean [$\log(z(\cdot))$ under Model I and $e(\cdot)$ under Model II], the sequential simulation algorithm (Sec. 2.1) was used. In our case, for each best linear (simple kriging) prediction, the nearest 40 (measured or simulated) values were used. The visiting sequence of simulation locations was randomized, starting on a coarse grid that was increasingly refined (Gómez-Hernández and Journel 1993; Pebesma and Wesseling 1998). One sample of a given size was simulated during a single pass through the simulation locations.

For the simulation of $z(\cdot)$ under Model II, the uncertainty in the regression coefficient vector $\beta = (\beta_0, \beta_1)^T$ was taken into account as well. Realizations from β were drawn from $\mathcal{N}(\hat{\beta}_{\text{gls}}, \Sigma_{\text{gls}})$, with $\hat{\beta}_{\text{gls}}$ the GLS estimate of β and Σ_{gls} the corresponding estimation (co)variance matrix (Cressie 1993). For the GLS estimation of β, a global neighborhood (all data) was used. The ith realization for $\log(z(\cdot))$ at x_0 was obtained by combining the ith element from a sample of β, β^i, with the ith realization of $e(\cdot), e_i(x)$ using Model II:

$$\beta_0^i + \beta_1^i \sqrt{M(x_0)} + e_i(x_0).$$

Note that β is spatially invariant, whereas $e(x)$ is not. LHS was applied to the sampling of β (taking the estimation error covariance into account) and to the sampling of $e(x)$, before combining the two. In the estimation, estimation error of β and estimation error in $e(x)$ were assumed to be independent, and the ensemble of possible realizations at any location x_0 under this model is (marginally) distributed as

$$\mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 \sqrt{M(x_0)} + \hat{e}(x_0), p_0^T \Sigma_{\text{gls}} p_0 + \sigma_e^2(x_0)), \quad (5)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the GLS estimates, $p_0 = (1, \sqrt{M(x_0)})^T$, and $\hat{e}(x_0)$ and $\sigma_e^2(x_0)$ are obtained when, in (2) and (3), residuals at observed data locations are substituted for data. The bottom row of Figure 3 shows three elements of a simple random sample of conditional simulations under Model II. Clearly, compared to Model I they show a stronger tendency of increasing zinc concentrations toward the river, especially in sparsely sampled areas.

For both models, the offset of the semivariogram [the value of $\gamma(h)$ when h approaches 0] was taken as a nugget effect—that is, small-distance variation. Although the true measurement error will be larger than 0, we had no information on it and treated it as 0. This results in more variable realizations of the random fields than those that would have been obtained when measurement error was taken larger than 0 (Cressie 1993).

3.2 Areal Fraction Above a Threshold

Suppose we want to estimate the areal fraction at which the zinc concentration exceeds a certain threshold c. The quantity of interest then is

$$a(c) = \frac{1}{|D|} \int_{x \in D} I(z(x), c)\, dx, \quad (6)$$

with $|D|$ the area of D and

$$I(z(x), c) = \begin{cases} 1 & \text{if } z(x) > c \\ 0 & \text{otherwise.} \end{cases}$$

For practical reasons the integral in (6) is approximated by a fixed sum,

$$a(c) \approx \frac{1}{K} \sum_{j=1}^{K} I(z(x_j), c), \quad (7)$$

with K the number of locations x_j discretizing D.

Replacing $z(x)$ with the random variable $Z(x)$ in (6) yields the random variable $A(c)$. Given a Gaussian model for $Z(\cdot)$, we can easily calculate the expected value of $A(c)$

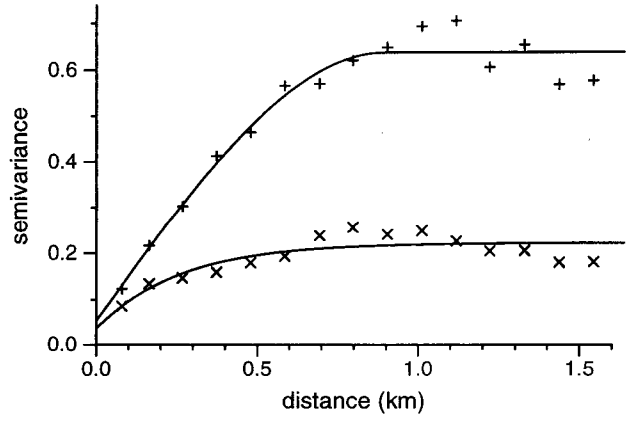

Figure 6. Sample Semivariograms of the Log(zinc, ppm) Measurements (+), of Regression Residuals (×) and Fitted Models (——).

using the sum approximation

$$E(A(c)) \approx \frac{1}{K}\sum_{j=1}^{K} E(I(Z(x),c)) = \frac{1}{K}\sum_{j=1}^{K}(1 - F_{Z(x_j)}(c)) \quad (8)$$

and the standard normal cumulative distribution function and (4) for Model I or (5) for Model II. Other properties of the distribution of $A(c)$ (such as the variance or percentiles), however, are not as easily derived analytically. Therefore, we use Monte Carlo simulation for this. We are interested in the entire probability distribution of $A(c)$, conditional to the data, because this tells us how much we know about the true value $a(c)$.

The critical level $c = \log(200\ \text{ppm})$ was studied, and the study area was represented by $K = 3{,}103$ point locations on a square grid, having a grid spacing of 40 meters (Fig. 3). We focus on the mean $E(A(c))$ and the 5- and 95-percentile of $A(c)$. The latter two can be used as an approximate 90% prediction interval for $A(c)$.

To investigate the efficiency gain of LHS, both a simple random sample and a Latin hypercube sample of sizes 20, 50, and 100 realizations of $Z(\cdot)$ are used to estimate the mean, 5- and 95-percentile of $A(c)$. For a given model, this setup results in 2 (sampling methods) $\times$ 3 (sample sizes) $\times$ 3 (sample statistics) $=$ 18 parameters from a single sampling round.

To assess sampling error, this sampling round was repeated 100 times, resulting in a set of 100 independent replicates. For each replicate, mean values and sample 5- and 95-percentiles are reported for all 18 parameters. The results are shown in Figure 7.

Stein (1987) showed that the efficiency gain of LHS compared to simple random sampling (SRS) depends on how the target quantity (the function outcome) depends on the variables sampled: The more the function is additive in the variables, the more LHS improves on SRS. Although the simple function (7) used in the case study is not linear with respect to the $Z(x_j)$, both models saw a tremendous gain in sampling efficiency with LHS for the mean value of $A(c)$.

For Model I, to reach the sampling error of a Latin hypercube sample of size 20, approximately, a simple random sample of size 5,400 would be needed. The efficiency gain is much less for the 5- and 95-percentile of $A(c)$, which is no surprise because these quantities are not nearly as additive in the variables as the mean value. Choosing percentiles closer to .5 or a c value closer to the median of the data would result in a larger efficiency gain because then the additive part of the model becomes more important. Both simple random sampling and LHS result in unbiased estimates of $E(A(c))$ (dashed line in Fig. 7).

For Model II (Fig. 7, bottom), all uncertainties are larger, and the percentiles for $A(c)$ are much farther apart. Moreover, the efficiency gain of Latin hypercube is slightly smaller. For the larger sample sizes, LHS results in a small bias for the mean of $A(c)$. This may be caused by the fact that, for the estimation of the residual component of $\hat{z}(x_j)$ in (8), residuals were assumed to be known, whereas GLS-predicted residuals were actually used. Although ignoring correlation between the trend and residual component in Model II yields somewhat larger variances than the universal kriging prediction variance (Cressie 1993, 3.4.64), the results of Model II are judged to be more realistic than those obtained under Model I. This is because Model I unrealistically assumes that the mean of $Z(\cdot)$ is known and because it disregards the strong relationship between zinc concentration and distance to the river (Figs. 4b and 5).

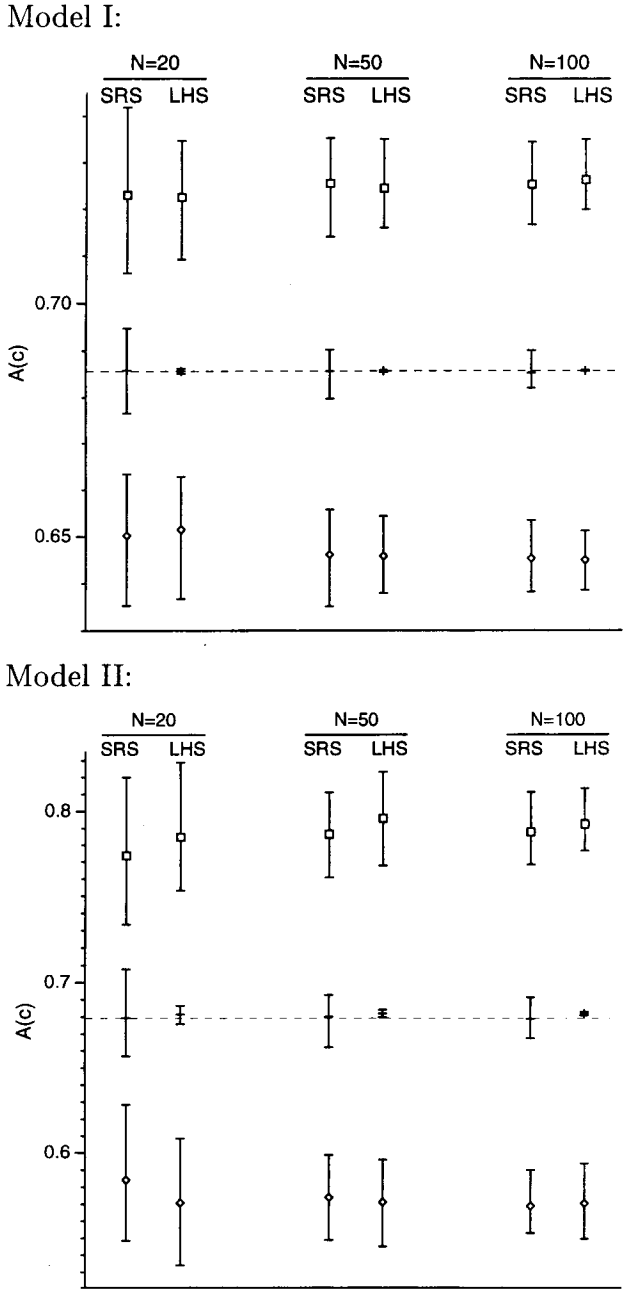

Figure 7. 90% Confidence Intervals for the 5-Percentile (◇), the Mean Value (+), and the 95-Percentile (□) of A(c), Depending on Sample Size (N = 20, 50, 100) and Sampling Strategy (SRS—simple random sampling, LHS—-Latin hypercube sampling) for Model I (top) and Model II (bottom). The dashed line shows the analytical result according to (8).

Note also that the uncertainty about the estimated semivariograms was not taken into account when the prediction intervals for $A(c)$ were obtained. This could lead to overly optimistic prediction intervals for $A(c)$.

Because $A(c)$ was defined in terms of "point" values $z(x)$, the support (physical size, Journel and Huijbregts 1978) of the sampled material has a strong influence on $A(c)$, especially for more extreme values of c. In this case, 10 topsoil subsamples of approximately 200 cm^3 topsoil were selected randomly within a radius of 10 m from the sampling site center, and zinc concentration was measured after mechanical mixing of the subsamples for a site.

4. DISTURBANCE OF SPATIAL CORRELATION BY LHS

To obtain a Latin hypercube sample from a simple random sample, the sample elements are shifted (Fig. 2). Because the multivariate density is not preserved exactly during this shift, the result does not necessarily honor the original multivariate distribution. Obviously, the larger the sample size, the smaller the shift, and the closer the approximation will be. But the rationale for using LHS instead of simple random sampling is to get more out of small samples. Therefore, we will now study the reproduction of the main multivariate characteristic of Gaussian fields—that is, the spatial correlation. The shift basically consists of two parts (Fig. 2)—(1) displacement to the category (stratum), determined by the rank in the original sample, and (2) simple random sampling within this category (addition of ξ_{ij}). Step 1 does not necessarily change the spatial correlation much because correlated values at nearby spatial locations will likely have similar ranks in the sample and therefore undergo a similar displacement. Step 2, however, introduces "white noise"—shifts within categories that are spatially uncorrelated (compare rows 2 and 3 in Fig. 3 for an illustration of this effect). This will result in Latin hypercube samples with smaller short-distance correlations than those obtained from simple random sampling.

To illustrate this, the reproduction of the semivariance of locations x_i and x_j,

$$\gamma(x_i, x_j) = \frac{1}{2} E(Z(x_i) - Z(x_j))^2,$$

was studied in an artificial example. For a Gaussian random field having an exponential semivariogram, $\gamma(x_i, x_j) = 1 - \exp(-|x_i - x_j|)$, three location pairs (x_i, x_j) were chosen such that $\gamma(x_i, x_j)$ was equal to 5%, 50%, and 95% of $\text{var}(Z(x))$. Next, Latin hypercube samples of sizes $N = 20, 50$, and 100 were drawn unconditionally for these pairs, and this was replicated 1,000 times. Sample semivariances for the three location pairs were calculated for each realization (sample element), and mean values with error bounds are shown in Figure 8. As a reference check, the same procedure was done for a simple random sample of size 100. The magnitude of the bias is listed in Table 3.

Figure 8 and Table 3 show that for short distances LHS yields sample semivariances that have a positive bias of approximately $\text{var}(Z(x))/N$. This corresponds to a negative bias in short-distance spatial correlation of approximately $1/N$.

5. DISCUSSION AND CONCLUSIONS

For models that are at least partially additive in their input variables, given a sample size, the model output distribution obtained by LHS of the model input variables is more accurate than that obtained by simple random sampling. The simple model of the case study showed that for the prediction of the mean areal fraction having a zinc concentration above a critical level, a Latin hypercube random sample of size 20 performed equally as well as a simple random sample of approximately size 5,000.

The method used to obtain the Latin hypercube sample adopted from Stein (1987) is simple and easy to implement. Simple random sampling and subsequent LHS of Gaussian random fields have been implemented in the gstat computer program, which is freely available in source code (Pebesma and Wesseling 1998). It creates multiple realizations of the simple random sample following a single random path (Sec. 2). When all realizations can be kept in memory during simulation, conversion to a Latin hypercube sample is done after simulation and before saving to disk (or running a model on them). Relative to the CPU time required to create the simple random sample, the time spent on the LHS is small.

Stein (1987) warned of distortion of the multivariate distribution under LHS when the sample size N is not much larger than K, the number of variables. In the practice of spatial simulation this will often be the case: The number of grid cells used to represent a two- or three-dimensional space will easily exceed the maximum feasible number of Monte Carlo runs. We think that this is not necessarily an objection to the method.

Using replicated LHS (Iman and Conover 1980), an artificial example in Section 4 indicated that for Gaussian random fields the negative bias in spatial correlation is approximately inversely related to the sample size. In applications in which reproduction of short-distance spatial correlation is of critical importance (e.g., flow simulation where short-distance behavior relates to preferential flow and where the Monte Carlo sample size should be kept very small), the introduction of bias resulting from using LHS should be seriously considered, and it may even prohibit use of the method. In many applications, however, this bias may well be ignored, either because the Monte Carlo sample size is large enough or because the bias is negligible compared to the accuracy with which spatial correlation was assessed from observed data. In applications in which these correlations are well known and in which white noise is part of the spatial variation (semivariograms with a nugget effect), the semivariogram of the random field could be preadjusted to compensate for the bias. In any case, replicated LHS can be used to assess the bias.

Iman and Conover (1980) presented another method to generate a Latin hypercube sample for dependent variables, and a slightly modified version of their method was given

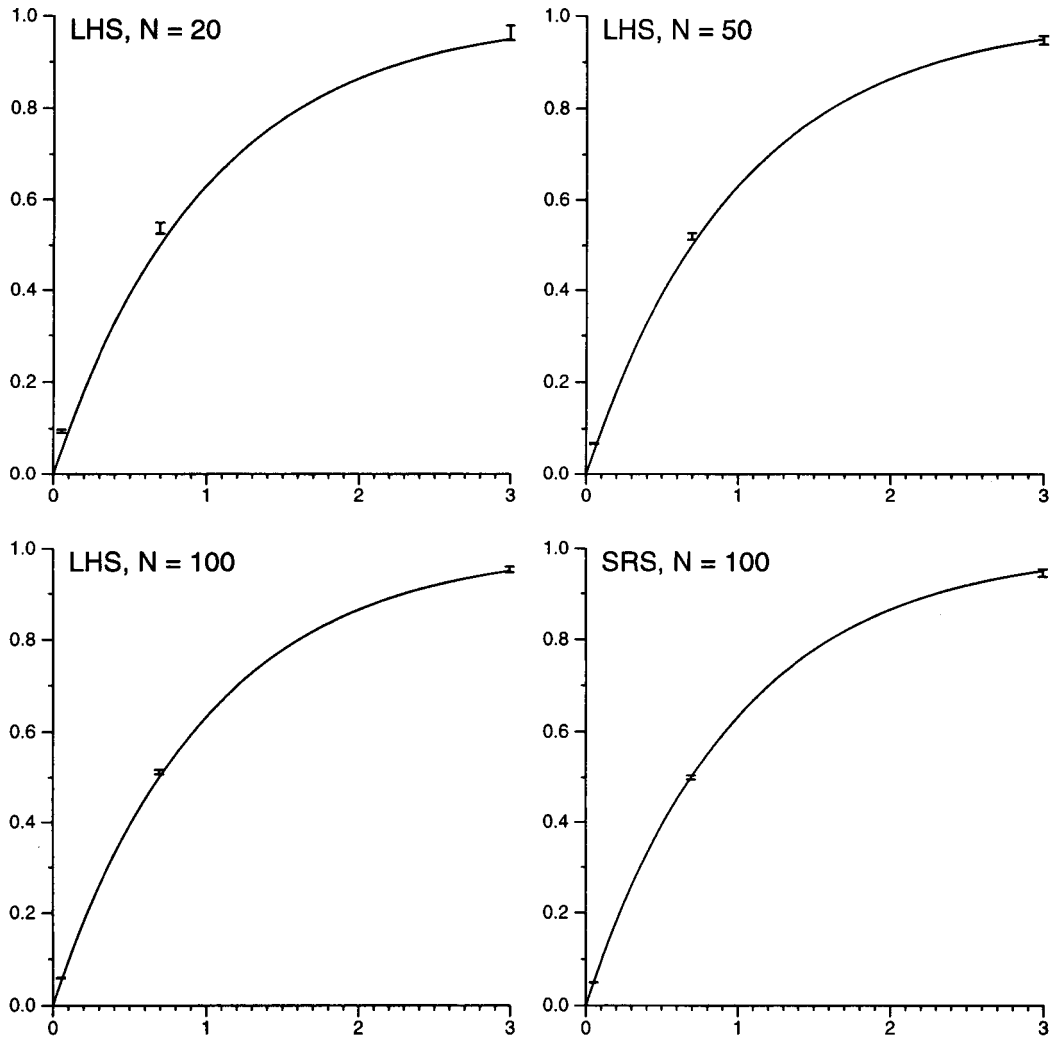

Figure 8. Semivariance Reproduction at Distances Where $\gamma(x_i, x_j)$ (curved line) Is .05, .5, and .95. Error bounds indicate the mean ±2 s.e., with s.e. calculated from (the averages of) 1,000 independently drawn samples. The x axis reads distance, the y axis semivariance.

by Owen (1994). Their method calls for repeated Choleski decomposition of a $K \times K$ matrix, however, and the number of locations K will be large in typical applications of Gaussian random fields.

In the case study, a very simple function of a random field was evaluated using both a simple model and a more realistic model for the data. An application with a more complex function, non-Gaussian random fields, and multiple cross-correlated spatial attributes was given by Kros, Pebesma, Reinds, and Finke (1999).

Table 3. Bias in Semivariogram Reproduction due to Latin Hypercube Sampling (reproduced—true semivariance) as a Function of Sample Size and (true) Semivariance Value

True $\gamma(h)$	*.05*	*.5*	*.95*
Bias LHS, N = 20	.0445*	.0373*	.0134
Bias LHS, N = 50	.0183*	.0202*	−.0052
Bias LHS, N = 100	.0097*	.0119*	.0033

NOTE: A * denotes values for which the interval defined through bias ±2 s.e. does not contain 0.

ACKNOWLEDGMENTS

We thank Ruud van Rijn, Mathieu Rikken, and Peter Burrough for permission to use the Meuse dataset. We thank Marc Bierkens, Tommy Norberg, Jaime Gómez-Hernández, and anonymous referees for helpful comments. This research was funded by the EU Environment and Climate Research and Technological Development Programme (contract number ENV4-CT95-0070).

[Received April 1998. Revised April 1999.]

REFERENCES

Beven, K., and Binley, A. (1992), "The Future of Distributed Models: Model Calibration and Uncertainty Prediction," *Hydrological Processes*, 6, 279–298.

Burrough, P. A., and McDonnell, R. A. (1998), *Principles of Geographical Information Systems*, Oxford, U.K.: Oxford University Press.

Cressie, N. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley.

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

Gómez-Hernández, J. J. (1997), "Issues on Environmental Risk Assessment," in *Geostatistics Wollongong '96*, eds. E. Y. Baafi and N. A.

Schofield, Dordrecht, The Netherlands: Kluwer, pp. 15–26.

Gómez-Hernández, J. J., and Journel, A. G. (1993), "Joint Sequential Simulation of MultiGaussian Fields," in *Geostatistics Tróia '92*, ed. A. Soares, Dordrecht, The Netherlands: Kluwer, pp. 85–94.

Gotway, C. A. (1994), "The Use of Conditional Simulation in Nuclear-Waste-Site Performance Assessment" (with discussion), *Technometrics*, 36, 129–161.

Heuvelink, G. B. M. (1998), *Error Propagation in Environmental Modelling With GIS*, London: Taylor & Francis.

Iman, R. L., and Conover, W. J. (1980), "A Distribution Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics, Part B—Simulation and Computation*, 11, 311–333.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, New York: Wiley.

Journel, A. G. (1989), "Fundamentals of Geostatistics in Five Lessons," in *Short Course in Geology* (vol. 8), eds. M. L. Crawford and E. Padovani, Washington, DC: American Geophysical Union, pp. iii–40.

——— (1997), "The Abuse of Principles in Model Building and the Quest for Objectivity," in *Geostatistics Wollongong '96*, eds. E. Y. Baafi and N. A. Schofield, Dordrecht, The Netherlands: Kluwer, pp. 3–14.

Journel, A. G., and Huijbregts, C. J. (1978), *Mining Geostatistics*, London: Academic Press.

Kitanidis, P. K., and Vomvoris, E. G. (1983), "A Geostatistical Approach to the Inverse Problem In Groundwater Modeling (Steady State) and One-Dimensional Simulations," *Water Resources Research*, 19, 677–690.

Kros, J., Pebesma, E. J., Reinds, G. J., and Finke, P. F. (1999), "Uncertainty in Modelling Soil Acidification at the European Scale: A Case Study," *Journal of Environmental Quality*, 28, 366–377.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979), "A Comparison of Three Methods for Selection Values of Input Variables in the Analysis of Output From a Computer Code," *Technometrics*, 2, 239–245.

Owen, A. B. (1994), "Controlling Correlations in Latin Hypercube Sampling," *Journal of the American Statistical Association*, 89, 1517–1522.

Pebesma, E. J., and Wesseling, C. G. (1998), "Gstat, a Program for Geostatistical Modelling, Prediction, and Simulation," *Computers & Geosciences*, 24, 17–31.

Ross, S. M. (1990), *A Course in Simulation*, New York: Macmillan.

Smith, L., and Freeze, R. A. (1979), "Stochastic Analysis of Steady State Groundwater Flow in a Bounded Domain: 2. Two-dimensional Simulations," *Water Resources Research*, 15, 1543–1559.

Stein, M. L. (1987), "Large Sample Properties of Simulations Using Latin Hypercube Sampling," *Technometrics*, 29, 143–151.

Thiele, M. R., Rao, S. E., and Blunt, M. J. (1996), "Quantifying Uncertainty in Reservoir Performance Using Stream Tubes," *Mathematical Geology*, 28, 843–856.