Incorporation of Auxiliary Information in the Geostatistical Simulation of Soil Nitrate Nitrogen

S. Grunwald,* P. Goovaerts, C. M. Bliss, N. B. Comerford, and S. Lamsal

ABSTRACT

In north-central Florida the potential risk for movement of nitrate into the aquifer is high due to the large extent of well-drained marine-derived quartz sand overlying porous limestone material coupled with high precipitation rates. Our objective was to estimate spatio-seasonal distributions of soil NO$_3$–N across the Santa Fe River Watershed in north-central Florida. We conducted spatially distributed synoptic and seasonal sampling (September 2003—wet summer/fall season, January 2004—dry winter season, May 2004—dry spring season) of soil NO$_3$–N. Prior distributions of probability for NO$_3$–N were inferred at each location across the watershed using ordered logistic regression. Exploratory variables included environmental spatial datasets such as land use, drainage class, and the Floridian aquifer DRASTIC index. These prior probabilities were then updated using indicator kriging, and multiple realizations of the spatial distribution of soil NO$_3$–N were generated by sequential indicator simulation. Cross-validation indicated that smaller prediction errors are obtained when secondary information is incorporated in the analysis and when indicator kriging is used instead of ordinary kriging to analyze these datasets characterized by the presence of extreme high values and a nonnegligible number of data below the detection limit. The NO$_3$–N values were lowest in September 2003 as a result of excessive leaching caused by large, intense tropical storms. Overall the NO$_3$–N values in January 2004 were high and could be attributed to fertilization of crops and pastures, low plant uptake, and low microbial transformation during the winter period. Despite seasonal trends reflected by the values of observed and estimated NO$_3$–N, we found areas that showed consistently high soil NO$_3$–N throughout all seasons. Those areas are prime targets to implement best management practices.

WATER RESOURCES perform multiple functions supporting economic, ecological, and sociocultural activities. Multiple stresses are imposed on Florida’s water resources by humanity (e.g., ecotourism, urban, recreation and agricultural activities) and nature (e.g., high-intensity rainfall-runoff events, and hurricanes). According to the Florida Department of Environmental Protection (Scott, 2002), Florida’s aquifer system contains more than 8.33 $\times$ 10$^{12}$ m$^3$ of fresh water and 90% of Florida’s population relies on groundwater resources for its drinking water. The vadose zone provides an interface between land use activities and the Floridian (deep), Intermediate, and Surficial aquifer systems. In north-central Florida, the Floridian aquifer is confined (artesian) on the Highlands side but unconfined (nonartesian) on the Lowlands side, providing ample pathways for nutrients and contaminants to leach into the deep aquifer system (Fernald and Purdum, 1998). The potential risk for nutrient and contaminant movement in this region is high because of the large extent of well-drained marine-derived quartz sand overlying porous limestone material coupled with high annual precipitation rates (>1300 mm) (National Climatic Data Center, NOAA, 2005) that causes excessive leaching.

The Suwannee Basin has the largest area of elevated NO$_3$–N concentrations in the Floridian aquifer system in the state of Florida (Suwannee River Water Management District [SRWMD], 2002). According to the groundwater monitoring program operated by SRWMD, NO$_3$–N was elevated in the Middle Suwannee River Watershed and along a NW-SE axis extending through the Santa Fe River Watershed (SFRW) (SRWMD, 2002). Springs in the Middle Suwannee River Watershed have NO$_3$–N concentrations ranging from 1.2 to 17.0 mg L$^{-1}$ (SRWMD, 2002). The groundwater in the Middle Suwannee River watershed flows toward the Suwannee River and affects the surface water quality of the Suwannee River via springs and seeps in the riverbed. Although the SFRW (3585 km$^2$) covers only 13.8% of the Suwannee Basin (25 900 km$^2$) it contributed 22.2% (2676 t NO$_3$–N) of the total NO$_3$–N loads in 2000 and 19.7% (2971 t NO$_3$–N) of the total NO$_3$–N loads in 2003 into the Gulf of Mexico (SRWMD, 2000, 2003). Katz et al. (1999) used $^3$H-N-O isotope tracers to track sources of nitrate in the Suwannee Basin. Their results indicated that nitrate most likely originated from a mixture of inorganic fertilizers and organic animal waste sources.

In this paper we adopt a geostatistical simulation approach for modeling the spatial distribution of soil NO$_3$–N in the SFRW. This approach is hybrid in that the prior probabilities of exceeding a series of NO$_3$–N threshold values are first derived from secondary information using ordered logistic regression (McCullagh, 1980), and these probabilities are updated using nonparametric geostatistics (i.e., indicator kriging). Goovaerts (1999), Heuvelink and Webster (2001), and McBratney et al. (2000, 2003) presented state-of-the-art geostatistical and hybrid modeling techniques for landscape-based modeling. McBratney et al. (2000) compared 24 different univariate and multivariate statistical, geostatistical, and hybrid models at field, subcatchment, and regional scales and found that hybrid geospatial techniques outperformed all other methods at all scales. They emphasized that mixed models

Abbreviations: ccdf, conditional cumulative distribution function; DL, detection limit; ESRI, Environmental Systems Research Institute; IK, indicator kriging; PI, probability interval; SFRW, Santa Fe River Watershed; SIS, sequential indicator simulation; SKlm, simple kriging with local mean; SRWMD, Suwannee River Water Management District; SSURGO, Soil Survey Geographic Database.

S. Grunwald, C.M. Bliss, N.B. Comerford, and S. Lamsal, Soil and Water Science Dep., Institute of Food and Agric. Sciences, Univ. of Florida, 2169 McCarty Hall, P.O. Box 110290, Gainesville, FL 32611-0290, USA; P. Goovaerts, BioMedware, Inc., 516 N. State St., Ann Arbor, MI 48104, USA. Received 27 Feb. 2005. *Corresponding author (SGrunwald@ifas.ufl.edu).


Special Section: Landscape Processes

© Soil Science Society of America

677 S. Segoe Rd., Madison, WI 53711 USA

© Reproduced from Vadose Zone Journal. Published by Soil Science Society of America. All copyrights reserved.
that combine auxiliary environmental factors with sparsely sampled soil field data are well suited for regional predictive modeling. Other comparisons of numerous hybrid geostatistical applications were presented by Odeh et al. (2004). Our objective was to estimate spatioseasonal distributions of soil NO$_3$–N across the SFRW. Because of the expected spatiotemporal variation of NO$_3$–N, we conducted spatially distributed synoptic and seasonal sampling of soil NO$_3$–N. We used profile averaged soil NO$_3$–N field observations and auxiliary environmental spatial datasets such as land use, drainage class, DRASTIC Index, and other attributes to develop spatial probability trends for three different seasons (wet summer/fall, dry winter, dry spring seasons). The approach presented in this paper is the first application of ordered logistic regression to estimate the prior probabilities of occurrence in indicator kriging. Cross-validation was used to assess the prediction performances of the proposed approach and quantify the benefit over alternative approaches, such as ordinary kriging or univariate indicator kriging.

**MATERIALS AND METHODS**

**Study Area**

The SFRW, which is part of the Suwannee Basin, is located in north-central Florida and covers an area of approximately 3585 km$^2$. The soils are predominantly sandy in texture. Dominant soil orders include Ultisols (36.7%), Spodosols (25.8%), Entisols (14.7%), and there is a minor extent of Histosols (2.0%), Inceptisols (1.1%), and Alfisols (1.0%) (Fig. 1). Land use is mixed, consisting of established pine (*Pinus* spp.) plantation (32.2%), wetlands (16.2%), upland forest (14.7%), improved pasture (14.0%), urban (8.8%), newly planted forest plantations (6.0%), crops (5.0%), rangeland (3.7%), and a variety of high intensity land uses such as tree groves, nurseries, dairies, and feeding operations (SRWMD and SJWMD, 1995). Two main physiographic regions in the watershed are the Gulf Coastal Lowlands and the Northern Highlands, which are separated from one another by the Cody Scarp (Schmidt, 1997). The underlying geology in the western part of the watershed is limestone, providing potential leaching pathways for soil NO$_3$–N into the aquifer. In contrast, the Cody Scarp is clayey, phosphate-rich, and dissects the watershed along a northwest–southeast axis. The eastern part of the watershed is dominated by sand-rich sediments with varying loamy subsoils (Brown et al., 1990; Randazzo and Jones, 1997). The elevation ranges from around 3 to 91 m above mean sea level (Fig. 1). By and large the land is level (0–2% slopes) or gently sloping and undulating (0–5% slopes), the major exception to this pattern being the moderately and strongly sloping land (5–12% slopes) along the Cody Scarp. The stream network shows a much higher density in the eastern part of the watershed, compared with the western part that is dominated by karst terrain and sinkholes (Fig. 1). The Santa Fe River flows underground for about 3 km bypassing the Cody Scarp. The mean annual precipitation based on 30 yr of records from 1971 to 2000 across the watershed was 1334 mm, and the mean annual temperature was 20.4°C (National Climatic Data Center, NOAA).

**The Datasets**

**Field Observations**

Soil samples were collected during September 2003, January 2004, and May 2004 at four depth increments (0–30, 30–60, 60–120, and 120–180 cm) in composites proportional to the depth of sampling to ensure a constant sampling support for each increment. A stratified-random sampling design was used where strata were derived from land use–soil order factor combinations in ArcGIS (Environmental Systems Research Institute [ESRI], Redlands, CA). Sample sites were randomly selected within strata proportional to their aerial extent in addition to targeting high-intensity land uses such as tree groves and feedlots. Not all sampling sites could be sampled because of field conditions (e.g., high water table during rainy season) and/or private land ownership restrictions. Therefore, the number of samples collected varied slightly by season and depth (Table 1). The samples were extracted with 2 M KCl (Keeney and Nelson, 1982), and the extraction solution was analyzed for NO$_3$–N content and expressed in micrograms per gram of soil. The

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.70</td>
<td>4.09</td>
<td>1.17</td>
</tr>
<tr>
<td>SEM</td>
<td>0.13</td>
<td>1.45</td>
<td>0.24</td>
</tr>
<tr>
<td>95% Confidence of mean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td>0.45</td>
<td>1.22</td>
<td>0.69</td>
</tr>
<tr>
<td>Upper</td>
<td>0.97</td>
<td>6.96</td>
<td>1.64</td>
</tr>
<tr>
<td>Median</td>
<td>0.23</td>
<td>0.11</td>
<td>0.29</td>
</tr>
<tr>
<td>Max.</td>
<td>6.54</td>
<td>103.71</td>
<td>19.92</td>
</tr>
<tr>
<td>SD</td>
<td>1.31</td>
<td>16.09</td>
<td>2.74</td>
</tr>
<tr>
<td>Skewness</td>
<td>2.77</td>
<td>4.09</td>
<td>4.37</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>7.59</td>
<td>25.36</td>
<td>22.30</td>
</tr>
<tr>
<td>Detection limit (DL)</td>
<td>0.01</td>
<td>0.06</td>
<td>0.105</td>
</tr>
<tr>
<td>% data below DL</td>
<td>29.7</td>
<td>41.5</td>
<td>11.7</td>
</tr>
</tbody>
</table>
analytical instrument has a detection limit (DL) of about 0.02 ppm of NO$_3$–N in the extraction solution. In Table 1 the statistical properties of profile averaged soil NO$_3$–N for all three sampling events are summarized.

In Florida, wet and dry seasons are pronounced with highest precipitation rates typically during the summer months (August–September) and lowest rates during winter and spring seasons. Wet seasons cause excessive leaching of soil NO$_3$–N and other nutrients through the vadose zone into the surface and groundwater. Precipitation in September 2003 was high (680.7 mm), whereas precipitation was much lower in January 2004 (274.3 mm) and May 2004 (231.4 mm).

Figure 2 shows the spatial distribution of observed soil NO$_3$–N values, which indicates persistently high values in the northeast for all sampling events. Soil NO$_3$–N was highest in January 2004, with a maximum value of 103.7 µg g$^{-1}$. Overall, the NO$_3$–N values in January 2004 could be attributed to fertilization of crops and pastures, low plant uptake, and low microbial immobilization and nitrification during the winter period. Observed NO$_3$–N values were lowest in September 2003 due to excessive leaching caused by large intense rainfall–runoff events. High NO$_3$–N values were measured in feedlots and tree groves, with a maximum value of 8.9 µg g$^{-1}$ (May 2004). Improved pasture showed low NO$_3$–N values in September 2003, with a mean of 1.2 and 2.9 µg g$^{-1}$ in May 2004, but a much higher mean in January 2004 of 11.4 µg g$^{-1}$. Throughout all sampling events, pine plantations showed very low NO$_3$–N measurements, with means throughout all seasons of 0.2 µg g$^{-1}$ and wetlands with 0.3 µg g$^{-1}$ (September 2003), 0.1 µg g$^{-1}$ (January 2004), and 0.5 µg g$^{-1}$ (May 2004).

**Auxiliary Environmental Spatial Data**

We assembled a comprehensive suite of auxiliary environmental spatial data to characterize soils, geology, topography, land cover, land use, and landscape properties for each soil sampling location. Table 2 provides a summary of the auxiliary environmental variables and their data sources. While some variables were point specific (e.g., elevation) others were lumped across a local neighborhood of varying size. For example, spectral reflectance values derived from Landsat ETM+ satellite imagery (pixel resolution 30 m) were averaged within 3 × 3 and 10 × 10 pixel neighborhoods corresponding to localized areas of 8100 and 90 000 m$^2$, respectively. Majority land use was derived by identifying the major (dominant) land use category within a 3 × 3 pixel neighborhood. The soil data were derived from the Soil Survey Geographic Database (SSURGO; map scale 1:24 000) that contain taxonomic and soil property data lumped across soil map units. The Floridian aquifer DRASTIC Index was derived as a weighting average using the following variables: (i) depth to water, (ii) net recharge, (iii) aquifer media, (iv) soil media, (v) topography, (vi) impact of the vadose zone, and (vii) hydraulic conductivity (data source: Florida Dep. of Environmental Protection; detailed description DRASTIC indices available at http://www.fgdl.org [verified 4 Nov. 2005]). High DRASTIC indices correspond to high vulnerability to groundwater pollution. All auxiliary environmental data were analyzed and stored within the ArcGIS Vers. 8.3 geographic information system (ESRI, 2005).

**Geostatistical Approach**

The following sections outline our geostatistical modeling approach combining environmental factors and observed soil NO$_3$–N point data. Geostatistics was used to model the spatial uncertainty of soil NO$_3$–N concentration at any location $u$ in the study area $A$ (i.e., the SFRW). This model of uncertainty takes the form of a set of $L$ equally probable realizations of the spatial distribution of NO$_3$–N values, $z^{(l)}(u); l = 1, \ldots, L, \forall u \in A$. Each realization represents a possible scenario for the distribution of NO$_3$–N values across the watershed. At any particular location $u$, the set of simulated values $\{z^{(l)}(u); l = 1, \ldots, L\}$ provides an empirical conditional cumulative distribution function (ccdf) $F(u,z;[\text{Info}])$, which gives the probability that the concentration is no greater than any given threshold $z$. The conditioning information, “Info,” consists of the set of $n$ soil NO$_3$–N concentrations $\{z(u); \alpha = 1, \ldots, n\}$ plus the $V$ secondary attributes available across the study area $\{y(u); j = 1, \ldots, V, \forall u \in A\}$.

**Soft Indicator Kriging**

The ccdf $F(u,z;[\text{Info}])$ can be modeled using two main categories of geostatistical algorithms, which are referred to as parametric or nonparametric depending on whether a parametric model (i.e., multiGaussian) is adopted for the random function $Z(u)$ (for a review see Goovaerts, 2001). In this paper, a nonparametric approach, known as indicator kriging (IK), is adopted because the large proportion of data below the detection limit makes hazardous any transform to achieve
a Gaussian distribution. This method requires a preliminary coding of each piece of information into a vector of indicators, defined for a set of \( K \) thresholds, with \( z_k \) discretizing the range of variation of the soil attribute. Two sets of indicator data can be created: hard indicators resulting from the coding of NO\textsubscript{2}–N data and soft indicators derived from secondary information.

Hard indicators, which take a value of 0 or 1, are computed at the sampled location \( u_s \) as:

\[
I(u_s; z_k) = \begin{cases} 
1 & \text{if } z(u_s) \leq z_k \\
0 & \text{otherwise} 
\end{cases} \quad k = 1, 2, \ldots, K 
\tag{1}
\]

Soft indicators \( j(u_s; z_k) \) are valued between 0 and 1, and they were calculated from the set of secondary variables (continuous and categorical) by performing ordered logistic regression using the proportional odds model (Procedure Logistic, SAS Institute, 1989). This model assumes that the impact of the different secondary variables is the same across thresholds (i.e., same regression coefficients) and only the intercept of the regression model is threshold-specific. In theory, a more complex cumulative logit model with nonproportional odds would have better predictive ability. Yet, the larger number of regression parameters associated with this model could not be estimated accurately (i.e., lack of convergence) given the small size of the dataset and the fact that most secondary variables are categorical and require the estimation of \((S - 1)\) coefficients where \( S \) is the number of categories.

Soft indicators can be viewed as the local means of the hard indicator variable; in other words, the GIS data are used to inform on the regional trend of the NO\textsubscript{2}–N data. At any location \( u \in A \), the prior probability \( j(u; z_k) \) can be updated into a posterior probability \( P(z_k | u) \) using the following simple kriging with local mean (SKlm) estimator (Goovaerts and Journel, 1995):

\[
I_{SKlm}(u; z_k) = F_{IK}(u; z_k | \text{[Info]}) \quad k = 1, 2, \ldots, K \\
= j(u; z_k) + \sum_{n=1}^{N(u)} \lambda_n(u; z_k) [I(u_n; z_k) - j(u_n; z_k)] 
\tag{2}
\]

The weights \( \lambda_n(u; z_k) \) are the solution of the following system:

\[
\sum_{n=1}^{N(u)} \lambda_n(u; z_k) C_{R_R}(u_n - u_p; z_k) = C_R(u_n - u_p; z_k) \\
\alpha = 1, \ldots, N(u) 
\tag{3}
\]

where \( C_R(u_n - u_p; z_k) \) is the covariance function of the residual indicator variable \( R(u; z_k) = I(u; z_k) - j(u; z_k) \) for the vector joining the locations \( u_n \) and \( u_p \).

Although the kriging system (Eq. [3]) is written in terms of covariances, common practice consists of inferring and modeling the semivariogram rather than the covariance function. The experimental semivariogram for a given lag vector \( h \) is estimated as:

\[
\hat{\gamma}_b(h; z_k) = \frac{1}{2N(h)} \sum_{a=1}^{N(h)} [r(u_a; z_k) - r(u_a + h; z_k)]^2 
\tag{4}
\]

where \( N(h) \) is the number of data pairs within the class of distance and direction used for the lag vector \( h \). A continuous function must be fitted to \( \hat{\gamma}_b(h; z_k) \) to deduce semivariogram values for any possible lag \( h \) required by prediction algorithms, and also to smooth out sample fluctuations. In this study, the semivariograms were modeled using least-square regression (Pardo-Iguzquiza, 1999). All semivariogram models were bounded (i.e., reached a sill), and the covariance models were derived by subtracting the semivariance values from the sill value.

**Modeling of the Cumulative Conditional Distribution Function**

Because the \( K \) probabilities are estimated individually (i.e., \( K \) indicator kriging systems are solved at each location) the following constraints, which are implicit to any probability distribution, might not be satisfied by all sets of \( K \) estimates:

\[
0 \leq F_{IK}(u; z_k | \text{[Info]}) \leq 1 \quad \forall k 
\tag{5}
\]

\[
F_{IK}(u; z_k | \text{[Info]}) \leq F_{IK}(u; z_{k'} | \text{[Info]}) \quad \text{if } z_k \leq z_{k'} 
\tag{6}
\]
All probabilities that are not between 0 and 1 are first reset to the closest bound, 0 or 1. Then, condition (Eq. [6], shown above) is ensured by averaging the results of an upward and downward correction of ccdf values (Deutsch and Journel, 1998).

Once conditional probabilities were estimated and corrected for potential order relation deviations, the set of \( K \) probabilities must be interpolated within each class \((z_a, z_a + 1)\) and extrapolated beyond the smallest and the largest thresholds to build a continuous model for the ccdf. In this study, the resolution of the discrete ccdf was increased by performing a linear interpolation between tabulated bounds provided by the sample histogram of \( \text{NO}_2-N \) concentration (Deutsch and Journel, 1998).

Validation of the Prediction Models

Prediction errors, as well as the quality of the uncertainty models, were quantified using cross-validation. One \( \text{NO}_2-N \) observation is removed at a time, and the ccdf at this location is modeled using neighboring soil data and secondary information. The RMSE is computed as:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( z(U_i) - z_j(U_i) \right)^2}
\]

where the \( \text{NO}_2-N \) concentration at \( U_i \) is predicted using the mean of the ccdf there, \( z_j(U_i) \), which is referred to as the E-type estimate (Goovaerts, 1997).

At any location \( U \) knowledge of the ccdf \( F(U; z(U_\text{Info})) \) allows the computation of a series of symmetric \( p \) probability intervals (PI) bounded by the \((1 - p)/2\) and \((1 + p)/2\) quantiles of that ccdf. For example, the 0.5 PI is bounded by the lower and upper quartiles \([F^{-1}(U; 0.25|\{\text{Info}\}), F^{-1}(U; 0.75|\{\text{Info}\})]\). A correct modeling of local uncertainty would entail that there is a 0.5 probability that the actual \( z \) value at \( U \) falls into that interval or, equivalently, that over the study area 50% of the 0.5 PI include the true value. Cross-validation yields a set of \( z \) measurements and independently derived ccdfs at the \( n \) locations \( U_i \), allowing the fraction of true values falling into the symmetric \( p \) PI to be computed as:

\[
\bar{z}(p) = \frac{1}{n} \sum_{i=1}^{n} \left[ z(U_i) \right]_{(1 - p)/2}^{(1 + p)/2} | \{\text{Info}\}
\]

where \( w(p_i) \) equals 1 if \( z(U_i) \) lies between the \((1 - p)/2\) and \((1 + p)/2\) quantiles of the ccdf and equals zero otherwise. The scattergram of the estimated, \( z(U_i) \), vs. expected, \( p \), fractions is called the “accuracy plot.” Deutsch (1997) proposed assessing the closeness of the estimated and theoretical fractions with the following “goodness” statistic:

\[
G = 1 - \frac{1}{S} \sum_{s=1}^{S} \left[ w(p_s) \right]_p \left| \bar{z}(p_s) - p_s \right|
\]

where \( w(p_s) \) equals 1 if \( \bar{z}(p_s) > p_s \), and equals 2 otherwise. The statistics are computed over \( S = 99 \) thresholds corresponding to the percentiles of the histogram of \( \text{NO}_2-N \) data. Twice the importance is given to deviations when \( \bar{z}(p_s) < p_s \) (inaccurate case), that is, the case where the fraction of true values falling into the \( p \) PI is smaller than expected.

Not only should the true value fall into the PI according to the expected probability \( p \), but this interval should be as narrow as possible to reduce the uncertainty about that value. In other words, among two probabilistic models with similar goodness statistics one would prefer the one with the smallest spread (less uncertain). Different measures of ccdf spread can be used: variance, interquartile range, and entropy. Following Goovaerts (2001) the average width of the PI’s that include the true value are plotted for a series of probabilities \( p \). For a probability \( p \) the average width is computed as:

\[
\bar{W}(p) = \frac{1}{n} \sum_{a=1}^{n} \left[ z(U_a, p) - z(U_a, (1 + p)/2) | \{\text{Info}\} \right]
\]

Sequential Indicator Simulation

Indicator kriging allows one to model the uncertainty prevailing at a specific location \( U \) (i.e., local uncertainty) and to estimate the value of the soil attribute \( z(U) \) using the mean of the probability distribution (E-type estimate) at that location. The map of E-type estimates is, however, very smooth and does not provide a realistic picture of the spatial variability present in the field. Also, there is no straightforward way to combine the measures of local uncertainty to quantify the uncertainty attached to the value of the variable over blocks of irregular size and much larger than the measurement support. For example, we were interested in estimating the average \( \text{NO}_2-N \) concentration within each subbasin of the watershed and quantifying the attached uncertainty.

Unlike kriging, the aim of stochastic simulation is not to minimize a local error variance but to reproduce statistics such as the sample histogram or the semivariogram model in addition to honoring the data values. Thus, stochastic simulation is increasingly preferred to kriging for all applications where the spatial variability of the measured field must be preserved, such as the delineation of contaminated areas (Desbarats, 1996) or the modeling of solute transport in the vadose zone (Vanderborght et al., 1997). In particular, ccdf for supports larger than the measurement support (e.g., remediation units or flow simulator cells) can be numerically approximated by the cumulative distribution of block simulated values that are obtained by averaging values simulated within the block (Saito and Goovaerts, 2003).

The objective of stochastic simulation is to generate a set of simulated values \( \{z^{(i)}(U); i = 1, \ldots, L; \forall U \subset A\} \) that reproduce some target statistics (i.e., histogram, semivariogram) conditionally to the finite set of \( z \) data \( \{z(U_i); a = 1, \ldots, j\} \) and the secondary information \( \{y(U); j = 1, \ldots, J; \forall U \subset A\} \). Like estimation, simulation can be accomplished using a growing variety of techniques that differ in the underlying random function model (multiGaussian or nonparametric), the amount and type of information that can be accounted for, and the computer requirements (Goovaerts, 1997). In this study, realizations were generated using sequential indicator simulation (SIS), which relies on the aforementioned indicator semivariogram modeling and soft indicator kriging. The SIS algorithm proceeds as follows:

1. Define a random path visiting each node \( U \) of the simulation grid only once.
2. At each location \( U \), determine the ccdf using soft indicator kriging. The key difference with the implementation of indicator kriging is that not only the \( n \) original \( \text{NO}_2-N \) data, but also the previously simulated values, will be used as conditioning information for kriging.
3. Draw a simulated value from that distribution and add it to the data set.
4. Proceed to the next location along the random path, and repeat the two previous steps.
5. Loop until all \( N \) nodes are simulated.

The procedure is repeated using a different random path and set of random numbers to generate another realization.
RESULTS AND DISCUSSION

Ordered Logistic Regression

For each of the three time periods, a proportional odds model was developed using 10 indicator classes as dependent variables. The construction of these cumulative logits required the choice of $K = 9$ thresholds. For each time period, the DL was selected as the first threshold. The proportion of observations below the DL ranged from 11.7% (May 2004) to 41.5% (January 2004) (see Table 1). The other eight thresholds were selected so that the data above the DL were distributed equally among the different threshold classes. The explanatory variables included four attributes (Fig. 3): drainage, land use, majority land use, DRASTIC index Floridian aquifer; and one cross product to capture the interaction between drainage and land use.

Table 3 lists for each factor the $p$ value, which represents the probability of not rejecting the hypothesis that the factor is nonsignificant. A factor is said to be significant if the $p$ value is lower than 0.05, and its effect is highly significant if the $p$ values does not exceed 0.01. Land use was the only factor that is highly significant for all three time periods, and the associated $p$ values were the smallest. Drainage and its interaction with land use were only significant for September 2003, while the effect of DRASTIC index Floridian aquifer was significant in January and May 2004. The regression models were applied to the GIS layers. For example, Fig. 4 (top graph) shows the map of the “prior” probability of exceeding the seventh threshold of 0.582 $\text{m g}^{-1}$ in September 2003. The mean of these prior probability distributions (i.e., E-type estimates) was computed and mapped at the bottom of Fig. 4.

Higher NO$_3$–N estimates were found in the (western) central parts of the SFRW that is dominated by agriculture (crops and improved pasture) and rangeland. DRASTIC values were high in this area, indicating a high potential risk for leaching of nutrients that coincides with drainage that was classified as excessive due to karst terrain. Very low values were estimated for sites in the southwest that showed excessive drainage coinciding with Entisols. These soils are characterized by sandy soil profiles with low organic matter content, suggesting that they store little soil NO$_3$–N within the profile. Low to intermediate values were estimated along the Cody Scarp that is dominated by Ultisols with confining layers of clay. Wetlands, pine plantations, and upland forest dominate the eastern part of the watershed on Spodosols and Ultisols. In contrast, Spodosols are characterized by layers of spodic material that consists of illuvial organic materials and sesquioxides and is fragile and less confining. DRASTIC indices were relatively low in the eastern part of the watershed, suggesting low risk for NO$_3$–N leaching.

Soft Indicator Kriging

Hard indicators were created according to Eq. [1] using the same nine thresholds as in the ordered logistic regression. Residuals were obtained by subtracting the soft indicators (i.e., probabilities estimated by the proportional odds model) from the colocated hard data, and the $K$ corresponding indicator semivariograms were computed following Eq. [4]. Figure 5 shows the experimental indicator semivariograms with the model fitted for the data collected in January 2004. Any anisotropic

![GIS layers: drainage classes, Floridian aquifer DRASTIC index, and land use. (Drainage class code: VP = very poorly; SP = somewhat poorly; E = excessive; SE = somewhat excessive; MW = moderately well; P = poorly; W = well drained).](image-url)
(i.e., direction-dependent) variability was captured by the secondary information; hence, the omnidirectional semivariograms of the residuals were computed. For the larger thresholds in particular, the range was only a few kilometers, which means that most of the regional variability was explained by the secondary information and that the residual variability essentially reflects short-scale fluctuations.

The ability of indicator kriging to predict NO$_3$–N concentrations at unsampled locations and to model the associated uncertainty was assessed using cross-validation. The mean of each ccdf was computed (E-type estimate) and plotted vs. the actual concentration (Fig. 6). For the first two time periods, data below the detection limits were hard to predict. However, there is generally a reasonable correlation between E-type estimates and observed concentrations. Table 4 indicates that the largest prediction error was observed for the January 2004 sampling campaign, where more than 40% of the data were below the detection limit. To assess the benefit of secondary information on prediction performances, univariate IK was conducted. Table 4 shows that the RMSE value is systematically higher than the one obtained using soft IK for all time periods. Similar prediction errors were obtained using lognormal kriging, which indicates that although a lognormal transform does not properly address the issue of data below the detection limit, the prediction is as good as the one obtained using a nonparametric approach. Differences between the three alternate interpolation techniques are more pronounced in terms of correlation coefficient between observed and estimated soil nitrate concentrations, see Table 4. This drop in correlation reflects the smoothness of estimates obtained when the secondary information is ignored in the analysis.

The accuracy plots (Fig. 7, left column) show that the uncertainty models are generally accurate, with a slightly
higher than expected proportion of data within the probability intervals. The good agreement between expected and observed proportions was illustrated by the high value for the goodness statistic. Not only should the true soil NO$_3$–N concentration fall into the PI according to the expected probability $p$, but this interval should be as narrow as possible to reduce the uncertainty about that value. Figure 7 (right column) shows the width of the PI computed from the local (ccdf) and global probability distributions (cdf) for increasing probability levels. In particular for January 2004, the cdf-based probability intervals are narrower than the ones calculated from the ccdf, which is caused by the very large proportion of data below the detection limit. However, for larger probability values, the intervals computed from the global distribution are systematically much wider (less precise) than the intervals obtained using soft indicator kriging. Thus, commonly used 90% probability intervals would be more precise (narrower) when computed from IK-based ccdfs while including the expected proportion of true values.

**Sequential Indicator Simulation**

Following the cross-validation of the soft indicator kriging approach, the next step was to generate realizations of the spatial distribution of NO$_3$–N concentrations using sequential indicator simulation. One hundred realizations were generated for each season. Figure 8 shows the first two realizations for the three sampling periods. The comparison of the top maps of Fig. 8 with the bottom map of Fig. 4 clearly illustrates the smoothness of the map of prior estimates, which mainly depicts regional patterns inferred from the secondary information. The spatial variability displayed by simulated maps is consistent with the sample variance and the structures displayed by the semivariograms. These maps honor both the observed soil nitrate concentrations at sampling sites and their histogram, leading to average nitrate concentrations that are the highest in January and the smallest in September.

Each grid of simulated values was aggregated to the subbasin level, and Fig. 9 shows the mean and standard deviation of the distribution of 100 simulated NO$_3$–N concentrations within each subbasin. The maps of average NO$_3$–N concentrations provide a clear description of regional patterns and their seasonal changes. The maps of standard deviations highlight the larger uncertainty of estimates computed for small subbasins, which is in agreement with the sampling theory. The variability of NO$_3$–N concentrations within each subbasin can also be quantified using the variance of values simulated within...
each basin. These variances were averaged across all 100 realizations. Figure 10 indicates that the largest within-basin variability was observed in January 2004 in the northeastern part of the watershed.

For each season different spatial patterns of soil NO$_3$–N were generated. The NO$_3$–N patterns indicated that land use management, precipitation (dry/wet season), and other environmental factors such as plant uptake and microbial immobilization, which change throughout the year, are important to consider in modeling N patterns in space and through time. Both the observed soil NO$_3$–N and realizations generated across the watershed in January 2004, representing the dry winter season, were highest among all seasons. Areas in the southwestern part of the watershed consistently showed low NO$_3$–N values throughout all seasons. In contrast, in the south-central and the northeastern part of the watershed, high values were found in September, January, and May, indicating that land use and vadose zone characteristics were less favorable for leaching of NO$_3$–N. These areas are typically in row crops, sprayfields, and pasture land uses that are fertilized, located on Ultisols and Spodosols, and show moderate to somewhat poor drainage.

Florida dairies need year-around forage systems that prevent loss of N to groundwater from waste effluent sprayfields (Woodard et al., 2002). Dairies in northern Florida are under governmental mandate to use manure nutrients while preventing excessive leaching. Most are involved in the year-round production of forage crops in manure effluent sprayfields. According to Woodard et al. (2002), forage crops remove large quantities of nutrients from the soil because, unlike grain production systems, nearly all aboveground biomass is removed during harvest. They found that forage systems in bermudagrass [Cynodon dactylon (L.) Pers.] with N concentrations in the plant tissue of 18.1 to 24.2 g kg$^{-1}$ were more favorable than corn (Zea mays L.)–forage sorghum [Sorghum bicolor (L.) Moench]–rye (Secale cereale L.) systems with N concentrations of 10.3 to 14.7 g kg$^{-1}$ in removing soil N. Therefore, the bermudagrass system was more effective at preventing NO$_3$–N leaching.

Fertilization rates and timing are critical to avoid excessive N loads in water bodies. Bakhsh et al. (2001) emphasized that residual soil NO$_3$–N from fertilization depends on the time of application as well as the tillage management practice. Katz et al. (1999) documented that long-term trends of elevated NO$_3$–N in spring waters of the Suwannee Basin followed the increase in fertilizer use. Therefore, the documentation of land use, land use shifts, and management practices is critical to make inferences about soil NO$_3$–N and N observed in surface waters and wells.

Wetlands (>16% of watershed) and sites with shallow water tables (flats and flatwoods) comprise large areas of the watershed. Although wetlands are endpoints of flowpaths accumulating material from upland areas, they did not show elevated soil NO$_3$–N values. This is likely due to the fact that denitrification rates in these

<table>
<thead>
<tr>
<th>Estimator</th>
<th>September 2003</th>
<th>January 2004</th>
<th>May 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td></td>
<td>Correlation</td>
</tr>
<tr>
<td>Soft indicator kriging</td>
<td>1.101</td>
<td>15.04</td>
<td>2.127</td>
</tr>
<tr>
<td>Univariate indicator kriging</td>
<td>1.272</td>
<td>15.87</td>
<td>2.670</td>
</tr>
<tr>
<td>Lognormal kriging</td>
<td>1.274</td>
<td>15.90</td>
<td>2.687</td>
</tr>
<tr>
<td></td>
<td>0.546</td>
<td>0.369</td>
<td>0.640</td>
</tr>
<tr>
<td>Univariate indicator kriging</td>
<td>0.235</td>
<td>0.139</td>
<td>0.217</td>
</tr>
<tr>
<td>Lognormal kriging</td>
<td>0.355</td>
<td>0.263</td>
<td>0.244</td>
</tr>
</tbody>
</table>

Table 4. Root mean square error (RMSE) of prediction (units = $\mu$g g$^{-1}$) and correlation coefficient between observed nitrate concentrations and estimates obtained using alternate methods for the different time periods.
Florida wetlands were high (White and Reddy, 2003) and remove N from the watershed by emitting nitrous oxides into the atmosphere.

Soil NO₃–N in pine plantations and upland forest was low for several reasons. First, N fertilizer use is low compared with other crop production systems. An intensive N fertilization practice for forest plantations might apply 50 kg N ha⁻¹ once in the first 5 yr and as much as 200 kg N ha⁻¹ once during the following 15 yr. While fertilization is common, it is not necessarily as intensive as indicated above, nor is it universally applied. Second, forest ecosystems impose a closed nutrient cycle on a landscape. Nitrogen deficiencies are commonly most pronounced by age 10 in these pine plantations because much of the bioavailable N is tied up in the forest floor as N is immobilized during decomposition. Third, forest plantations continue to absorb nutrients from the soil during the winter (Comerford et al., 1987) leaving less available for leaching. Fourth, leaching of N from the crowns of pine plantations is minimized as the needles in the crown senesce because much of the N is retranslocated from these needles back into the tree crown and

Fig. 7. Plots of the proportion of observed NO₃–N data falling within probability intervals (accuracy plot) and the width of these intervals vs. the probability $p$. The goodness statistics measures the similarity between the expected and observed proportions in the accuracy plots (best if one). Open circles in the graphs of right column represent the width of probability intervals computed from the sample histogram (cdf).
used to support the next season’s growth. It has been estimated that approximately 50% of the total N content of senescing needles is removed from the needle before it falls and begins to decompose (Dallatea and Jokela, 1994; Aerts, 1996). All these processes minimize N use, maintain N on the site, and leave less for leaching.

CONCLUSIONS

We presented a geostatistical approach that integrated sparse field observations of soil NO₃–N with auxiliary spatial environmental datasets. The approach was innovative in its use of ordered logistic regression to estimate prior probabilities from secondary information, which...
were later incorporated in soft indicator kriging. The characterization of seasonal and spatial patterns of NO$_3$–N in the vadose zone within the SFRW will enable managers to minimize adverse impacts on the environment. Three different typical seasons were considered to acknowledge seasonal trends in soil NO$_3$–N due to climate and land use management activities. The most pronounced spatial patterns in soil NO$_3$–N were found for the January sampling event, with maximum values >100 µg g$^{-1}$. Thus, in Florida soil sampling during winter months may be considered the most representative period to highlight differences among land uses, soils, and geographic locations within this watershed.

Despite seasonal trends reflected by the values of observed and estimated NO$_3$–N, we found areas that showed consistently high soil NO$_3$–N in January, May, and September. Those areas should be targeted by best management practices to reduce N loads. Since this watershed is characterized in some areas having karst terrain combined with high annual precipitation, attention...
has to be given to land use activities performed within these “high risk areas” that potentially serve as short-circuits for NO$_3$–N entering the aquifer.

**ACKNOWLEDGMENTS**

This project was supported by the USDA grant no. 2002-00501 funded through the “Nutrient Science for Improved Watershed Management” program. This research was supported by the Florida Agricultural Experiment Station and approved for publication as Journal Series no. R-10791. We thank Dr. Mark W. Clark for his extension efforts, Dr. Donald A. Graetz and Dr. Randy B. Brown for their advice on planning this project, and Wade Hurt for characterizing soils at each field location. We also thank the public and private land owners that provided us access to their property to collect soil samples.

**REFERENCES**


SAS Inst., Cary, NC.
SRWMD (Suwannee River Water Management District). 2003. Surface water quality and biological monitoring. WR02/03–03. SRWMD, Tallahassee, FL.