6-1996

Spatio-Temporal Statistical Modeling of Livestock Waste in Streams

Noel Cressie
Iowa State University

James J. Majure
Iowa State University

Spatio-Temporal Statistical Modeling of Livestock Waste in Streams

Abstract
Surface water runoff from large livestock operations finds its way into streams, rivers, and ultimately the larger watershed area. In this paper, the model measures the nitrate concentrations in the upper North Bosque (Texas) watershed, which is a region of concentrated dairy operations. Using 15 months of daily data collected at 17 stream monitoring sites allows the authors to obtain optimal predictions of unknown nitrate concentration at all stream locations at any given time, along with a measure of their variability.

Keywords
Agriculture, Livestock, Models and assessment tools, Watershed and ecoregion, Water quality quantity and management

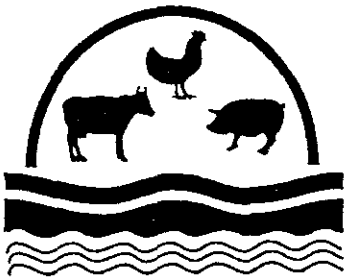
Disciplines
Agricultural and Resource Economics | Agriculture | Statistical Models | Water Resource Management

Spatio-Temporal Statistical Modeling of Livestock Waste in Streams

Livestock Series Report 5

Noel Cressie and James J. Majure

Staff Report 96-SR 81
June 1996



Spatio-Temporal Statistical Modeling of Livestock Waste in Streams

Livestock Series Report 5

Noel Cressie and James J. Majure

Staff Report 96-SR 81
June 1996

Center for Agricultural and Rural Development
Iowa State University
Ames, IA 50011-1070

Noel Cressie is Distinguished Professor of Statistics, Iowa State University; and James J. Majure was the manager of the GIS Support and Research Facility, Iowa State University.

This research was supported by CARD and the Texas Institute for Applied Environmental Research, the U.S. Environmental Protection Agency under Cooperative Agreement CR-822919-01-0, and by the National Science Foundation under Grant DMS-9204521. The National Pilot Project on Livestock and the Environment was funded by the U.S. Environmental Protection Agency under Cooperative Agreement #R820374010.

The results presented in this paper are based largely on James Majure's master's thesis, presented to the Department of Statistics, Iowa State University, in 1995.

ABSTRACT

Livestock agriculture (e.g., dairy, beef, pork, poultry) in the USA is tending rapidly toward operations where a large number of animals are concentrated in a relatively small area. The economies of scale are counterbalanced by the dangers of pollution from inadequate treatment of animal waste. Traditional methods of treatment involve lagoon retention and subsequent spreading on fields but the sheer volume of production seems to be outstripping these and other technologies. Surface-water runoff finds its way into streams and rivers, ultimately polluting all downstream segments of the watershed. The topic of this paper is spatio-temporal statistical modeling of (log) nitrate concentration in the upper North Bosque watershed, which is a region of concentrated dairy operations. A model is fitted from daily data collected over a period of 15 months, at 17 stream monitoring sites throughout the watershed. Optimal predictions of unknown nitrate concentration, at all stream locations at any given time, are obtained, along with a measure of their variability. The model allows for policy changes to be made, and assessed, based on the consequent spatio-temporal predictions.

SPATIO-TEMPORAL STATISTICAL MODELING OF LIVESTOCK WASTE IN STREAMS

Introduction

The possible harmful effects of livestock on the environment is beginning to be appreciated. Livestock waste may account for up to 20% of surface water pollution in the US (Long, 1992). An important source of pollution is livestock waste generated at concentrated animal feeding operations (CAFOs), including beef, dairy, swine, and poultry operations. Since 1972, the US Environmental Protection Agency (EPA) has attempted to control pollution from CAFOs by permitting all feedlots with more than 1,000 animal units as well as smaller operations that meet special water-related criteria. Compliance has been slow; by 1992, only about 10% of the eligible feedlots had been permitted (Long, 1992).

In surface waters (lakes, reservoirs, and streams), typical pollutants from CAFOs are nitrates, phosphorous, ammonia, and coliform bacteria, caused by natural or artificial flushing/ drainage. Further, nitrates, coliform bacteria, and metals and salts in manure contaminate groundwater; this is caused by natural seepage in regions of sensitive hydrogeology, leaking storage lagoons, and misapplication of waste onto agricultural land.

Livestock waste causes fish kills (high ammonia; excess nutrients cause excess growth of algae which reduce dissolved oxygen), eutrophication of lakes (excess nutrients), changes in water habitat ecosystems (excess algae reduce sunlight to submerged aquatic vegetation), unhealthy wildlife populations (coliform bacteria cause avian botulism and cholera that kill thousands of migratory waterfowl annually; zinc, copper, etc. adversely affect bottom-feeding aquatic birds), soil pollution (metals and salts), and acid deposition (ammonia). The cost of livestock-waste pollution can be measured in the short term by the cost of closure of sources of drinking water, nearby fisheries, agricultural industries, and recreational areas. In the long term, the loss of large sources of potable water, caused by polluted groundwater seeping into aquifers, would have disastrous consequences, particularly on the burgeoning population in the south and west of the USA.

Since the middle of 1992, the United States Environmental Protection Agency has funded a large project, *Livestock and the Environment: A National Pilot Project*, to study the effects on the environment of dairy CAFOs in and around Erath County, Texas. Principal partners in the

project are Iowa State University's Center for Agricultural and Rural Development (CARD) and Tarleton State University's Texas Institute for Applied Environmental Research (TIAER). This project is mainly concerned with surface waters and odor, although some consideration is also being given to groundwater.

In this article, emphasis is on CAFOs in the dairy industry and their impact on surface waters. We shall build a spatio-temporal statistical model of nitrate concentration in streams of the upper North Bosque watershed (in and around Erath County, Texas). As such, it will be a statistical *description* of exogenous, spatial, and temporal dependencies supported by data sampled from upper North Bosque streams over a period of more than a year.

Dairy CAFOs are in direct contrast to the popular image of a small dairy farm. Indeed, it would be more accurate to call them "milk production facilities" rather than "dairy farms." There are enormous economies of scale in milk production by CAFOs. In spite of punitive fines to any CAFO that violates the Texas Water Commission's (a Texas state agency) no-discharge policy, it is clear that the current economic system pushes the long-term cost of a polluted environment on to county, state, and federal tax payers. A more equitable system, whereby the (potential) polluter pays, seems beyond the reaches of a private market institution (Libby and Boggess, 1990), implying the need for executive and legislative action. Thus, economic and statistical studies are meant to support policy recommendations that would both minimize the impact of animal waste on the environment and distribute its cost among those responsible.

A few basic facts about the study area are enlightening. Erath County, with a 1990 population of 28,300 and an area of 697,500 acres, is located 65 miles south of Fort Worth, Texas. In 1980, the county had 181 dairies and 20,000 dairy cows. In 1990, the county had 197 dairies and 70,000 dairy cows; and over one billion pounds of milk were produced (for Houston and Dallas markets) bringing in $144 million in 1990. Erath County is currently the top milk-producing county in Texas. However, in any one day, approximately six million pounds of manure is produced in this region of only approximately 1,000 square miles.

Pollution of surface waters is the most common complaint in and around Erath County, originating either from downstream neighbors or a local association called the Cross Timbers Concerned Citizens (TIAER, 1992, p. 8). In 1987, in part as a response to growing complaints, the Texas Water Commission (TWC) revised its technical guidelines for CAFOs and embarked on a permitting campaign. Permits are required for dairy farms with more than 250 head and involve establishing stringent waste-management facilities for cumulative rainfall events and regular monitoring controls. Enforcing the permit requirements on a regular basis is another

matter; in 1991, the TWC conducted only about 10% of its required inspections (TIAER, 1992, p. 59). At the local level, there is considerable disagreement between permitted dairy operators, unpermitted (small) dairy operators, and concerned citizens. In the middle is the TWC trying to make and enforce scientifically sound and reasonable regulations. It is in front of this backdrop where one finds this spatio-temporal statistical analysis of non-point-source pollution of streams.

There is no claim that our statistical analysis shows causation. Water-chemistry sampling sites were located with classification factors such as geography, stream permanence, and site accessibility in mind. It is difficult, if not impossible, in such studies to produce a sampling design that allows properly for "treatment" factors such as upstream locations of dairies, dairy size, dairy-management practices, soil types, and topography. In other words, the statistical analysis given in this paper is of an *observational study* rather than of a carefully controlled, well designed experiment.

The goal of this statistical modeling effort is spatio-temporal prediction of surface-water nitrate contamination with known confidence. Each of the sections that follows plays a role in achieving that goal. The next section gives a brief description of the computing environment (Geographic Information System and statistical software) in which the spatio-temporal analysis was carried out, and of the data base used for the analysis. The third section discusses the modeling and fitting of the large-scale variation, while the fourth section concentrates on the small-scale (in space and time) variation. Spatio-temporal prediction results are given in the fifth section and a discussion is given in the final section.

The Computing Environment and the Data Base

The Geographic Information System (GIS) Arc/Info and Arcview was used to derive important geographic variables, to manage both the original and the derived data, to provide spatial displays (maps) of the upper North Bosque watershed, and to display kriging predictors (and kriging standard errors) at various locations on the stream network. The statistical software Splus was used for all statistical analysis, such as weighted regression, variogram estimation, variogram-model fitting, and kriging.

The data base is made up of two parts, the original data (e.g., contaminant concentrations, locations of sampling sites on streams) and the derived data base (e.g., topography, stream distance, landscape characteristics, area-of-influence management practices).

Original Data Base

The region for which the statistical analysis was conducted is the upper North Bosque

river basin, which is contained mainly within Erath County, Texas. The study area contains 87 dairies and has 24 surface-water sampling sites, made up of 17 stream sites and 7 reservoir sites. Figure 1 shows the study area with the stream network, locations of dairies, locations of precipitation gauges, and locations of surface-water sampling sites.

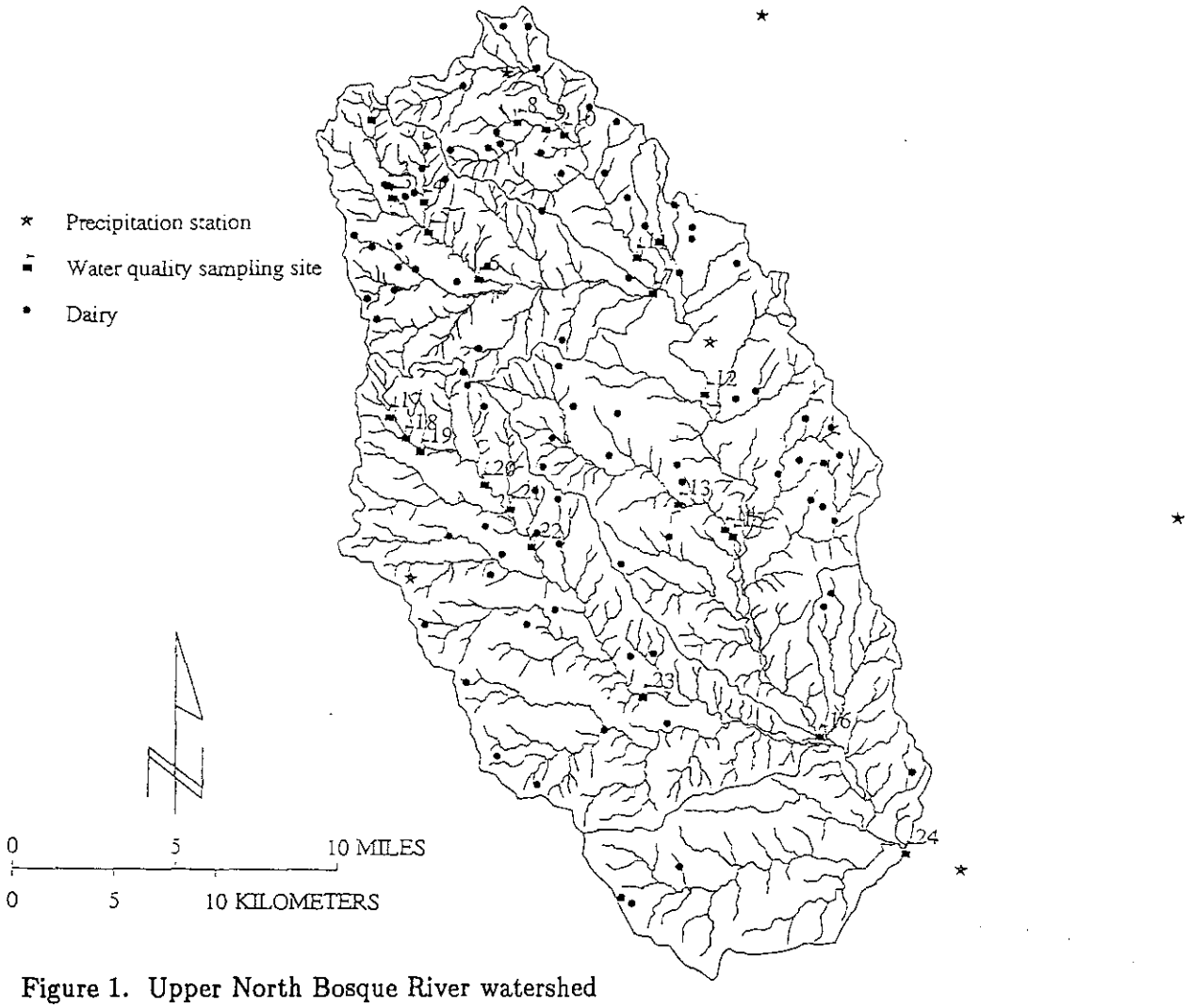


Figure 1. Upper North Bosque River watershed

The original data base consists of the following information:

- locations of 24 surface-water sampling sites along with the results of the analyses of 467 samples taken over the period from March 6, 1991 through May 27, 1992. The water-quality information consisted of 18 variables including field measurements (e.g., depth, air temperature, and so forth), and laboratory measurements (e.g., ammonia concentrations, nitrate concentrations, ortho-phosphate concentrations, and so forth);

- locations of 87 dairies along with information about the management characteristics used at each dairy (e.g., number of head, number of waste lagoons, waste application methods, and so forth);
- the DEM (digital elevation model) with a resolution of 30m, acquired from the US Geological Survey; and
- locations of six precipitation gauges with daily values for the period of interest.

Based on stratification decisions, the 24 sites were reduced to 17 stream (not reservoir) sites, with which there were 176 samples associated during the period March 6, 1991 through May 27, 1992. Of those, there were 157 values of nitrate concentration above the detection limit. These log-nitrate concentrations were standardized by their sample mean and sample variance; samples below the detection limit were not included in the analysis.

Derived Data Base

An underlying assumption in the modeling described in this article is that conditions near a sampling site affect the occurrence of pollutants at that site more than they would if the sampling site were further away. Furthermore, we assume that conditions that influence the occurrence of pollutants at a site are those that occur within the drainage basin of a site. Therefore, the variables used to explain large-scale variation in observed (log) nitrate concentrations are calculated via a *three-day area of influence* for a site, which is defined as that area within the site's drainage basin that drains to the site within three days. Three days was used because literature suggests that pollutant concentrations fluctuate considerably after a rain event and return to baseline conditions within three days (Johengen and Beeton, 1992). Because data on flow rate were unavailable, we cannot confirm a suspected flushing effect of initially high concentrations, followed by lower concentrations (as discharge increases due to the rain event), followed by increasing concentrations again, back to baseline, during the three-day period.

The three-day areas of influence are obtained as described below using hydrological modeling tools available in the GRID subsystem of ARC/INFO. These tools allow the extraction of hydrologic features from a digital representation of topography (i.e., a DEM). The features that can be extracted include flow lines (streams) and drainage basins. For a complete description of the algorithms used, see Jensen and Domingue (1988).

The distance from all pixels within a drainage basin to the drainage outlet results in a *flow-length* grid, which is used to calculate the stream distance between stream sites. These stream distances are the appropriate metric in Section 4 when carrying out a geostatistical

analysis. Also, the flow length grid can be easily converted from distance to time using assumptions about the speed of flow over the landscape. To create the three-day areas of influence, flow-length grids were created for each of the sampling sites. The distances in the flow-length grids were converted to times using two assumptions: 1) water flows at a uniform rate of 0.5 m/s throughout the drainage basin, and 2) flow through reservoirs adds 24 hours to flow time (Hauck, 1993). A modification to allow for variable flow rates would be easy to implement (Maidment, 1993). The resulting *time-of-flow* grids were then truncated if flow time exceeded 72 hours.

This process creates two new data sets that are important in the calculation of explanatory variables: the three-day area of influence for each site and the time-of-flow grid for each site. The time-of-flow grid is used to create a *time-weight grid* defined as follows:

$$w_{ij} = \begin{cases} (TOF_{ij})^{-1} & , \text{ if } TOF_{ij} > 24 \text{ hours} \\ (1/24) & , \text{ if } TOF_{ij} \leq 24 \text{ hours} , \end{cases} \tag{1}$$

where j ranges over all pixels in the three-day area of influence of site i, and TOF_{ij} is the time of flow, in hours, from pixel j to site i. This time-weight grid is used in the calculation of several variables in order to give more weight to factors closer to the sampling site than to those further away. The largest weight assigned was (1/24) in order to keep factors that occur very close to a site from completely dominating all others; 24 hours is the finest temporal resolution available from the original data base. The weighting in (1) amounts to a truncated inverse-distance weighting and, although it is *ad hoc*, it serves us well in spatial modeling where we wish to downweight data further away. Further, if the weighting does not reflect reality for the physical processes, there is a self-correction mechanism in the statistical modeling that allows heteroskedastic variation to be fit; see the subsection below on heteroskedasticity.

The formula for the calculation of many of the explanatory variables includes the area of the three-day area of influence for a site. The area used is actually a *weighted area* with the weights being determined from the time-weight grid defined by (1). The weighted area for site i is defined as,

$$A_i = \sum_j a_{ij} w_{ij} , \tag{2}$$

where a_{ij} is the area of the j-th pixel in the three-day area of influence of site i (in this case a_{ij} is 900 m^2 for all i and j) and w_{ij} is given by (2.1). This quantity is calculated for each site and is the area used in the calculation of those explanatory variables requiring a per-unit-area standardization.

Explanatory Variables

The first step in the analysis is to find variables that explain the large-scale variation in nitrate concentrations in surface waters. Factors that are likely to affect the occurrence of nitrates in surface waters include the locations of possible sources of nitrate in the landscape and physical characteristics that aid or hinder the transport of nitrates from the source into the streams. Possible sources include dairies and crop land. Factors that affect transport might include land slope in the basin and precipitation conditions. The characterization of these factors consists of determining the three-day area of influence for each sampling site and summarizing the occurrence of a factor within that area. The use of a GIS to automate this task made the quantification of these explanatory variables possible.

Seventeen explanatory variables are considered. The definition of each of these variables is given below; more details on their calculation in the GIS can be found in Cressie and Majure (1996). The variables considered include information about *management practices* on dairies and the *physical characteristics* of a site's three-day area of influence. Also considered are variables describing seasonal variation and the distance from a site to the basin outlet.

Number of Dairies per Acre:

$$DPA_i \equiv \frac{\sum_k \sum_j w_{ij} I_{jk}}{A_i}, \tag{3}$$

where k ranges over the dairies that fall within the three-day area of influence, w_{ij} is the time-weight grid defined by (1), A_i is the weighted area of the three-day area of influence defined by (2), and I_{jk} is 1 if the j-th pixel contains the k-th dairy and 0 otherwise.

Number of Head per Acre:

$$HPA_i \equiv \frac{\sum_k \sum_j w_{ij} I_{jk} HEAD_k}{A_i}, \tag{4}$$

where $HEAD_k$ is the number of head maintained at the k-th dairy.

Waste Lagoons per Acre:

$$LPA_i \equiv \frac{\sum_k \sum_j w_{ij} I_{jk} LAGOONS_k}{A_i}, \tag{5}$$

where $LAGOONS_k$ is the number of waste lagoons at the k-th dairy.

Waste-Application Methods: There are six variables that describe liquid-waste application methods and three variables that describe solid-waste application methods.

$$WAM_{il} \equiv \frac{\sum_k \sum_j w_{ij} I_{jk} HEAD_k I(\text{method } l \text{ is used on the } k\text{-th dairy})}{\sum_k \sum_j w_{ij} I_{jk} HEAD_k}, \tag{6}$$

where l ranges from $1, \ldots, 9$ application methods (liquid and solid), and $I(B)$ is 1 if the statement B is true and 0 otherwise. Liquid-waste applications methods are: big gun ($l = 1$), irrigated ($l = 2$), center pivot ($l = 3$), sprinkler ($l = 4$), traveling big gun ($l = 5$), and wheel move ($l = 6$). Solid-waste application methods are: spread ($l = 7$), harrow spread ($l = 8$), and spread disc ($l = 9$).

Soil Hydrologic Code: The soil hydrologic code can take values of "A", "B", "C", or "D", where "A" indicates a soil with a high infiltration rate and "D" indicates a soil with a low infiltration rate. Because these codes are ordinal, they are converted to integers from one to four and then used to define the variable,

$$SHC_i \equiv \frac{\sum_j \sum_k w_{ij} I_{jk} HEAD_k HYDRO_k}{A_i}, \tag{7}$$

where $HYDRO_k$ is the soil hydrologic code of the soil on which wastes are applied on the k-th dairy.

Average Slope:

$$AS_i \equiv \frac{\sum_j SLOPE_j}{J_i}, \tag{8}$$

where J_i is the number of pixels in the three-day area of influence for the i-th site and $SLOPE_j$ is the slope at pixel j ($j = 1, \ldots, J_i$) derived from the DEM using the GIS.

Distance to Basin Outlet: This variable, referred to as DBO, is the stream distance from the sampling site to the basin outlet (at site 24). It is included as a surrogate for a spatial trend over the river basin.

Precipitation: This variable represents the amount of rain that fell within the three-day area of influence in the two days prior to and the day of collection of the sample.

$$PREC_{it} \equiv P_{it} + P_{i,t-1} + P_{i,t-2} \ , \tag{9}$$

where $P_{i,t-m+1}$ is the total precipitation (in inches) that fell $(m-1)$ days prior to the date t that the sample was taken, within the area that drains to the i-th site in m days; $m = 1, 2, 3$.

Seasonal Variation: This variable represents the seasonal variation observed in the data and is defined as,

$$SV_t = \begin{cases} 1 & \text{July } 1 \leq t \leq \text{September } 30 \\ 2 & \text{April } 1 \leq t \leq \text{June } 30 \\ 3 & \text{October } 1 \leq t \leq \text{December } 31 \\ 4 & \text{January } 1 \leq t \leq \text{March } 31 \ . \end{cases} \tag{10}$$

Figure 2 shows (log) nitrate concentrations plotted against time with a Lowess filter (Cleveland, 1979) superimposed. The four seasonal periods are delineated using vertical lines.

Modeling the Large-Scale Variation

Let $Z(\mathbf{s}, t)$ denote the log-nitrate concentration (standardized by its sample mean and sample standard deviation) at stream location $\mathbf{s}$ and time t. The statistical modeling strategy taken here is to decompose Z into *deterministic* large-scale variation (mean) plus *stochastic* small-scale variation (error). The model can be written as:

$$Z(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \delta(\mathbf{s}, t) \ . \tag{11}$$

The large-scale variation, represented by $\mu(\mathbf{s}, t)$, is expressed as a linear function of k regressors $\mathbf{x}(\mathbf{s}, t)$:

$$\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)'\beta \ , \tag{12}$$

where $\mathbf{x}(\mathbf{s}, t) \equiv (x_1(\mathbf{s}, t), \ldots, x_k(\mathbf{s}, t))'$ is a $k \times 1$ vector whose entries correspond to variables such as time of year, precipitation, soil properties, management practices at upstream dairies, spatial location, basin characteristics, distances to upstream dairies, and so forth. The coefficients β of this equation are fitted using weighted least squares. The use of a linear model here is a means to an end, the end being optimal spatio-temporal prediction of (log) nitrate concentration. More mechanistic, deterministic models are possible (e.g., White et al., 1992) but they do not allow quantification of the predictor's variability.

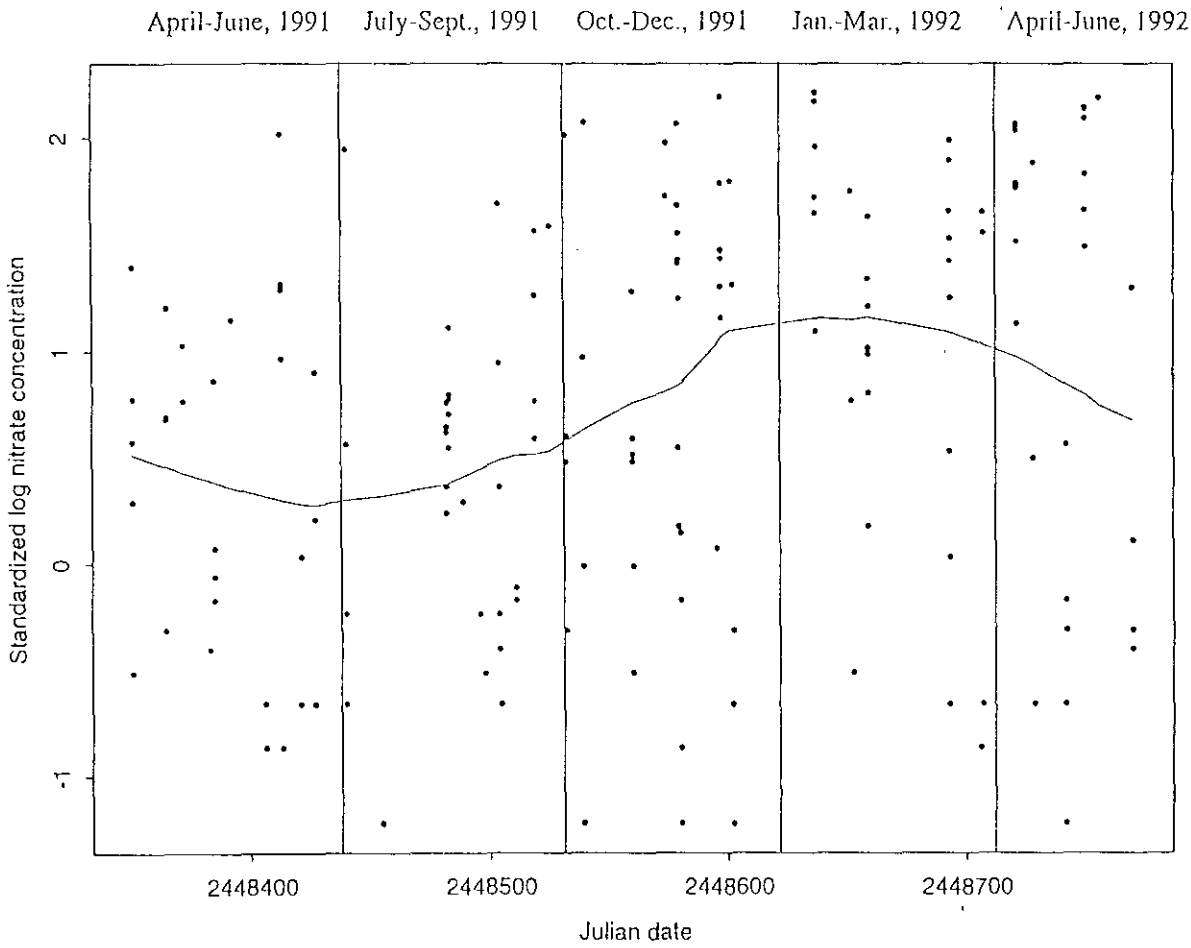


Figure 2. Plot of standardized log-nitrate concentrations versus time; the curve is the result of a Lowess filter (Cleveland, 1979) and the vertical lines delineate seasonal periods

The small-scale variation term is estimated by the residuals from the fitted large-scale variation:

$$\hat{\delta}(\mathbf{s},t) = Z(\mathbf{s},t) - x(\mathbf{s},t)'\hat{\beta} \; . \tag{13}$$

These residuals are treated as "data" and analyzed to characterize spatial and temporal dependence. This dependence, once characterized, is used to make predictions in space and time with known confidence.

An exploratory analysis indicated that the data are quite noisy but, when marginalized on either space or time, some dependencies are present. The exploratory analysis also indicated two sites, 4 and 12, that might potentially cause problems in more confirmatory analyses and, thus, should be watched carefully.

Variable Selection

The large-scale variation term for log nitrate concentrations was fitted using weighted least squares. The following regression-model selection criterion was used, following Ericksen, Kadane, and Tukey (1989). The 2^{17} possible linear models composed of the 17 explanatory variables given in the previous section were fitted in Splus. The selection of the final model was a two-stage process.In the first stage, all models were selected for which the ratio of each coefficient to its standard error was greater than two in absolute value. This step ensured that the coefficients of all of the explanatory variables in the model chosen were likely to be different from zero. Notice that, strictly speaking, it cannot be claimed that the coefficients were significantly different from zero due to spatial dependence among the regression errors. This would require knowledge of the errors' variance-covariance matrix, something that is unavailable at this phase of the analysis. Nevertheless, Ericksen, Kadane, and Tukey's selection criterion can still be used as a means of large-scale-variation model selection. The models identified in this first stage were compared in the second stage based on the criteria: small residual squared error and large R^2 value. The idea is to find a model with small error and high explanatory power. As expected, the model with the smallest residual squared error also had the largest R^2 value.

The models considered were initially fitted with weights equal to the weighted areas of the three-day areas of influence (equation (2)), reflecting the intuition that the larger the area the smaller the variability in (log) nitrate concentration. It is seen below that the final data weights chosen are different from these original weights, although they do retain this appealing feature.

Initially, the model

$$\mathbf{Z} = X\beta + \delta \tag{14}$$

was fitted, where the components of $\mathbf{Z}$ are made up of $Z(\mathbf{s}_i, t_k)$, the 157 standardized log nitrate concentration measurement at site $\mathbf{s}_i$ on day t_k; see the subsection on the original data base. The 157×17 matrix X is made up of 17 explanatory variables (described in the subsection on the derived data base) each obtained at the 157 location/time combinations. The 17×1 vector of regression coefficients are to be fitted assuming

$$\text{var}(\delta) = D\sigma^2 \,, \tag{15}$$

where σ^2 is unknown (to be estimated) and D is a 157×157 diagonal matrix with positive, possibly unequal, elements on the diagonal.

The weighted least squares fit of β is

$$\hat{\beta}_{wls} \equiv (X'D^{-1}X)^{-1}X'D^{-1}\mathrm{Z} \; ; \tag{16}$$

the residuals are

$$\hat{\delta} \equiv \mathrm{Z} - X\hat{\beta}_{wls} \; ; \tag{17}$$

and the standardized residuals are

$$\hat{\nu} \equiv D^{-1/2}\hat{\delta} \equiv D^{-1/2}\mathrm{Z} - D^{-1/2}X\hat{\beta}_{wls} \; . \tag{18}$$

The model-selection procedure described above considers subsets of the full model obtained by deleting explanatory variables. The initial weights used correspond to assuming

$$D = \mathrm{diag}(A_1^{-1}, \ldots, A_1^{-1}, A_2^{-1}, \ldots, A_{24}^{-1}) \; ,$$

where A_i is the weighted area for site i and the location of A_i^{-1} on the diagonal corresponds to the presence of a datum observed at site i.

Heteroskedasticity

Examination of residual plots indicated that the standardized residuals (given by equation (18)) from the initial models had heterogeneous variances. This heteroskedasticity was investigated as follows. From (14) and (15), write

$$D = \mathrm{diag}(A_1^{-k}, \ldots, A_1^{-k}, A_2^{-k}, \ldots, A_{24}^{-k}) \; ; \tag{19}$$

the initial weighted regression corresponds to the choice of $k = 1$. Thus, by fitting the weighting parameter k in (19), the initial choice of $k = 1$ can be confirmed or modified. Because of multiple samples (over time) at each site, the sample variance of residuals (see equation (17)) S_i^2 can be computed for the i-th site and the relation,

$$\log(S_i^2) = a + b\log(A_i) \; ,$$

can be fit by least squares, yielding estimates $\hat{a}$ and $\hat{b}$. Then an estimate of k would be $\hat{k} = -\hat{b}$; a value of $\hat{b}$ close to -1 would support our initial choice of $k = 1$.

This analysis was carried out and a value of $\hat{k} \cong 0.16$ was obtained, confirming the decrease in variability with larger weighted three-day area of influence but indicating that the

influence of those sites with larger areas should be considerably reduced. Therefore, the model (14), (15) was refit with D given by (19) and $k = 0.16$. A repeat of the model-selection procedure described in the previous subsection yielded the results listed below in the format given by Splus:

```
Residual Standard Error = 0.7228,  Multiple R-Square = 0.5291
N = 157,  F-statistic = 11.396 on 14 and 142 df, p-value = 0
```

	coef	std.err	t.stat	p.value
Intercept	−3.3178	1.6834	−1.9708	0.0507
dbo/1000000	−20.6250	4.7281	−4.3622	0.0000
lpa	60.1685	15.1205	3.9793	0.0001
as	1.3653	0.6651	2.0529	0.0419
hpa	0.0705	0.0170	4.1597	0.0001
shc	−0.0614	0.0159	−3.8709	0.0002
sv	0.2655	0.0563	4.7191	0.0000
wam1	−1.6712	0.2800	−5.9693	0.0000
wam2	−47.8895	11.5019	−4.1636	0.0001
wam3	−16.8877	6.0289	−2.8011	0.0058
wam4	14.9645	7.4018	2.0217	0.0451
wam5	−8.2808	4.0821	−2.0285	0.0444
wam6	−23.2975	6.0935	−3.8233	0.0002
wam8	9.8388	2.3597	4.1696	0.0001
wam9	−26.6024	5.6473	−4.7106	0.0000

(20)

The standardized residuals $\hat{\nu}$ from this model (see equation (18)) were examined using various diagnostic plots, a normal probability plot, and a stem-and-leaf plot; they were judged symmetric with a tendency to be heavier tailed than the normal distribution. In the next section, where spatio-temporal dependence in these residuals is modeled, a weighted-least-squares criterion is used to fit the variogram estimator and consequently the influence of heavy tails is down weighted.

It is tempting to use (20) for purposes beyond that for which it is intended, such as concluding that certain management practices *cause* high/low nitrate concentrations at nearby, downstream locations. The fitted regression given by (20) is merely a description of the large-scale variation of the data and should not be "over interpreted." It is a means to an end, the end being optimal prediction (with known precision) of log nitrate concentration in streams of the watershed; see the section on spatio-temporal prediction.

Small-Scale Variation

The small-scale variation is modeled as a zero mean *stochastic* process; in what follows, we shall refer to small-scale variation interchangeably as spatio-temporal dependence. The model (11), (12) is equivalent to,

$$Z(\mathbf{s},t) = \mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta} + \delta(\mathbf{s},t) \ . \tag{21}$$

The large-scale variation is determined once the coefficient $\boldsymbol{\beta}$ are; the previous section is concerned with estimating $\boldsymbol{\beta}$ efficiently using weighted least squares. Thus, with the large-scale variation estimated by

$$\mathbf{x}(\mathbf{s},t)'\hat{\boldsymbol{\beta}}_{wls} \ , \tag{22}$$

one can define the residuals

$$\hat{\delta}(\mathbf{s},t) \equiv Z(\mathbf{s},t) - \mathbf{x}(\mathbf{s},t)'\hat{\boldsymbol{\beta}}_{wls} \ , \tag{23}$$

or

$$\text{"residual"} = \text{"data"} - \text{"fit"} \ .$$

The residual is an estimate of the error (or small-scale variation) and should have zero mean (under the form of the model that was fitted). However, it does not have constant variance (i.e., it is heteroskedastic). Consider a location $\mathbf{s}$ (not necessarily a sampling site) on a stream in the watershed. Define

$$A(\mathbf{s}) \equiv \text{weighted area of the three-day area of influence upstream} \ , \tag{24}$$

where the weighted area is defined in an analogous fashion to (2). Indeed, in the notation of (2), $A_i \equiv A(\mathbf{s}_i)$, where $\mathbf{s}_i$ is the location of the i-th sampling site. Then, according to (18),

$$\hat{\nu}(\mathbf{s},t) \equiv (A(\mathbf{s})^{0.16})^{1/2}\hat{\delta}(\mathbf{s},t) = A(\mathbf{s})^{0.08}\{Z(\mathbf{s},t) - \mathbf{x}(\mathbf{s},t)'\hat{\boldsymbol{\beta}}_{wls}\} \ , \tag{25}$$

can be modeled as having mean zero and constant variance. It is to this standardized residual process that a (intrinsically stationary) spatio-temporal dependence model will be fitted.

The Variogram and Its Estimator

Spatio-temporal dependence can be characterized through the variogram,

$$2\gamma(h,u) \equiv \text{var}\{(A(\mathbf{s}')^{0.08}\delta(\mathbf{s}',t') - A(\mathbf{s})^{0.08}\delta(\mathbf{s},t)\} \ , \tag{26}$$

which is assumed to be a function only of the "lags" h and u, and where the space-lag interval $h \equiv \|\mathbf{s}' - \mathbf{s}\|$ is the stream distance between sampling sites, one downstream from the other, and

the time-lag interval $u \equiv |t' - t|$ is the time between samples. The intrinsic stationarity assumption in (26) allows 2γ to be estimated (e.g., Cressie, 1993, Section 2.4). Define

$$N(h,u) \equiv \{(\mathbf{s}_i,t_i),(\mathbf{s}_j,t_j) : \|\mathbf{s}_i - \mathbf{s}_j\| = h \text{ and } |t_i - t_j| = u\} , \tag{27}$$

where $\{(\mathbf{s}_i,t_i) : \ i = 1,\ldots,157\}$ are the locations of the monitoring sites and the times that samples were taken. The classical (method-of-moments) estimator is $2\hat{\gamma}$ with

$$2\hat{\gamma}(h,u) \equiv \sum_{N(h,u)} \{\hat{\nu}(\mathbf{s}_i,t_i) - \hat{\nu}(\mathbf{s}_j,t_j)\}^2/|N(h,u)| , \tag{28}$$

where $|N(h,u)|$ denotes the number of pairs in $N(h,u)$. The spatial predictors discussed in the next subsection require a valid variogram model that is obtained by fitting such a model to (28).

The spatial variogram is estimated by dividing the interval $[0, \max\{\|\mathbf{s}_i - \mathbf{s}_j\| : \text{for all } i,j\}]$ into mutually exclusive and exhaustive intervals, or lags. (Recall that here "space" is one-dimensional, along streams of the upper North Bosque watershed.) A variogram estimate is then calculated for each of the lags from all pairs of points whose separation distance falls within the interval for that lag. For spatio-temporal modeling, this must be done in the temporal dimension, as well.

The choice of lag boundaries is quite important and is often difficult. Lag boundaries should be chosen to allow sufficient averaging in the estimator (28); Journel and Huijbregts (1978, p. 194) recommend that there be at least 30 distinct pairs of points in each lag. However, it is also important to fit the variogram carefully near the origin. This often means that lag intervals near the origin should be narrower, even if it results in fewer pairs than is desirable. In the present case, the first lag in each of the time and space dimensions is a true zero lag. That is, there are pairs of samples that are collected at the same sampling site (for space), or on the same day (for time). Table 1 gives the lags that were used for variogram fitting. The numbers of pairs $|N(h_l,u_m)|$ in each of the 6×6 lags (the Cartesian product of space lag $\times$ time lag) do not always meet the 30-pair criterion. Of particular note is lag (h_0,u_0), which has only one pair, and lag (h_0,u_5), which has only two pairs. Special care is taken with these lags when fitting the variogram model; see below.

Figure 3 shows the classical variogram estimator (28) as a function of space lag h and time lag u and it shows clearly that if one fixes the space lag at $h = h_0(= 0)$, there is a strong temporal dependence. Similarly, if one fixes $u = u_0$, spatial dependence is exhibited, although this is not quite as clearly demonstrated. As one moves away from the origin in both space and time, the

Table 1. Intervals used to divide space and time lags

Space Lag l	Interval	Midpoint of h_l	Time Lag m	Interval	Midpoint of u_m
0	0 m	0	0	0 days	0
1	(0, 6172.69] m	3086.347	1	(0, 20] days	10
2	(6172.69, 12345.39] m	9259.041	2	(20, 40] days	30
3	(12345.39, 18518.08] m	15431.735	3	(40, 60] days	50
4	(18518.08, 24690.78] m	21604.429	4	(60, 80] days	70
5	(24690.78, 30863.47] m	27777.123	5	(80, 100] days	90

surface seems to fluctuate somewhat. However, the dip in the variogram estimates near the origin indicates the presence of spatial and temporal dependence.

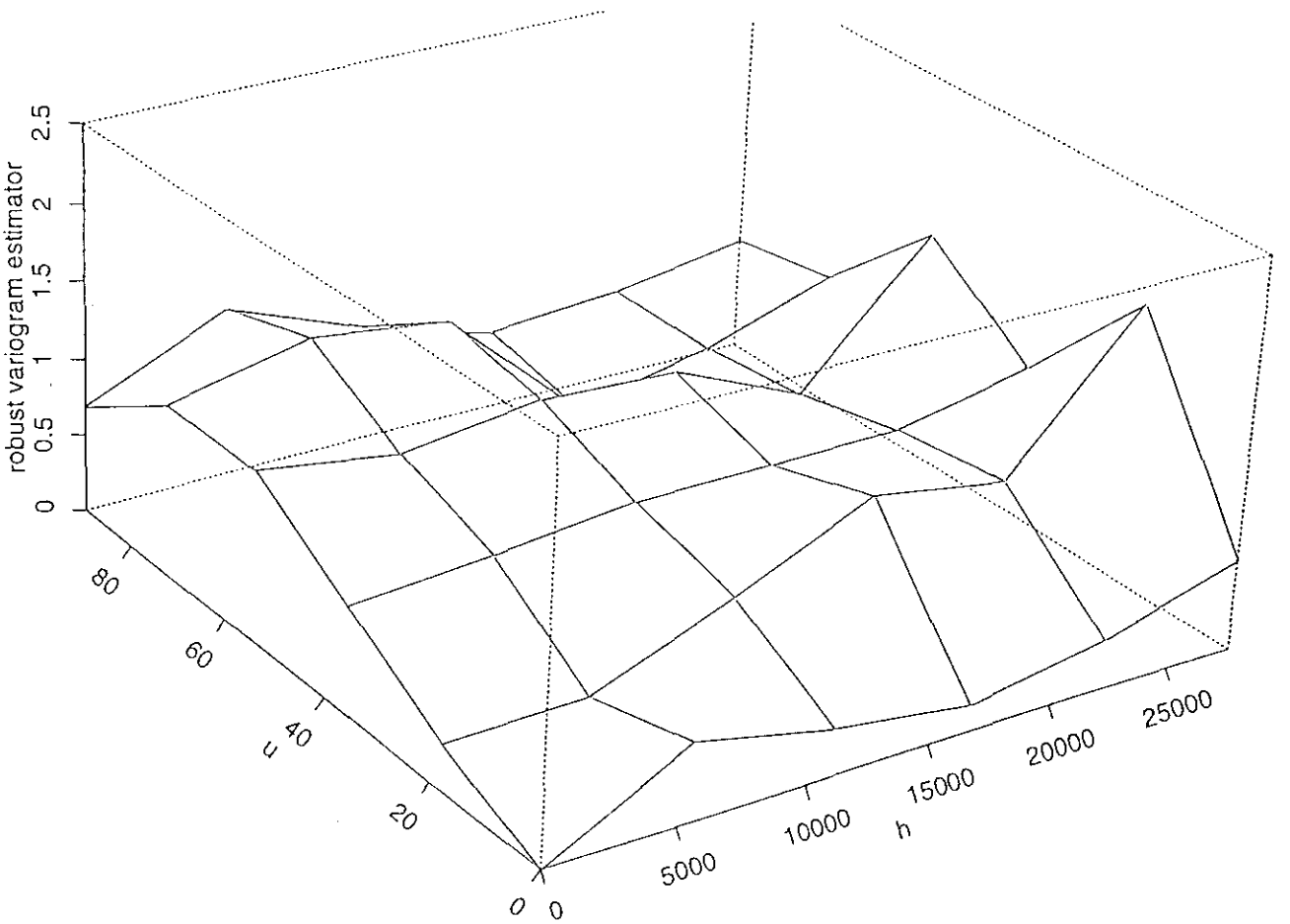


Figure 3. Surface based on empirical variogram estimates

The variogram (26) must be conditionally negative-definite (e.g., Cressie, 1993, p. 60); however, its estimators typically are not. Therefore, in order to carry out valid spatial prediction (with positive mean-squared prediction errors), a valid variogram model must be fit to the estimates shown in Figure 3. The proposed model is

$$2\gamma(h,u;\ \theta) = \begin{cases} 0 & , \quad h=0 \text{ and } u=0 \\ c_1+\theta_1(1-e^{-hb_1}) & , \quad h\neq 0 \text{ and } u=0 \\ c_2+\theta_2(1-e^{-ub_2}) & , \quad h=0 \text{ and } u\neq 0 \\ c_1+\theta_1(1-e^{-hb_1})+(c_2+\theta_2(1-e^{-ub_2})) & , \quad h\neq 0 \text{ and } u\neq 0 \end{cases} \tag{29}$$

where the components of $\theta \equiv (c_1, c_2, \theta_1, \theta_2, b_1, b_2)'$ are all non-negative parameters to be fitted.

This variogram model is constructed as the sum of a one-dimensional (along a stream network) exponential variogram and a one-dimensional (time) exponential variogram. It is easy to see that the covariance function associated with this variogram model is positive semi-definite. However, because it is not strictly positive-definite, there are particular arrangements of sample sites that can cause singularities in the kriging system (Myers and Journel, 1990; Rouhani and Myers, 1990). This arrangement of sample sites did occur during prediction but it is not difficult to handle it; see the end of the next subsection.

The parameters of the variogram model (29) were fitted to the empirical estimates (28) using the following steps. First, we concentrated on estimating (c_2, θ_2, b_2) by fixing $h = 0$ and fitting a one-dimensional (time) exponential variogram model to the variogram estimates obtained from (28). When h is fixed at 0, there are sufficient pairs of points for averaging (in the variogram estimator) with narrower lag intervals (seven intervals from 0 to 84 days in steps of 12 days and the eighth interval from 84 to 100 days) than in Table 1, so these narrower lags were used.

The variogram model was fitted using the weighted-least-squares criterion given by Cressie (1993, p. 99). That is, minimize with respect to $\theta = (c_2, \theta_2, b_2)'$ the following criterion:

$$\sum_{m=1}^{8} w(u_m;\ \theta)\{\hat{\gamma}(0,u_m) - (\gamma(0,u_m;\ \theta))\}^2 \ , \tag{30}$$

with weights $w(u_m;\ \theta)$ given by

$$\frac{|N(0,u_m)|}{\{\gamma(0,u_m;\ \theta)\}^2} \ , \tag{31}$$

where $N(0,u_m)$ is given by (27). The criterion (30) was minimized using a grid-search algorithm written in Splus. The resulting parameter estimates are $(\hat{c}_2, \hat{\theta}_2, \hat{b}_2) = (0, 1.100, 0.045)$.

Next, the parameter c_1 was estimated by fixing $u = 0$ and examining the resultant spatial variogram estimates. The variogram estimate at $h = 0$ is 0.004, which is very close to 0 (see

Figure 3). Even though only one pair of points was used in this estimate, evidence from the estimated variogram values for h near 0 and the fitting of the temporal variogram at $h = 0$, corroborated the estimated value of $\hat{c}_1 = 0$.

Finally, the two parameters (θ_1, b_1) was fitted to the empirical variogram estimates shown in Figure 3 by first fixing c_2, θ_2, b_2, c_1 at their estimates and then minimizing with respect to (θ_1, b_1):

$$\sum_{m=0}^{5}\sum_{l=1}^{5} w(h_l, u_m; (\theta_1, b_1))\{\hat{\gamma}(h_l, u_m) - \gamma(h_l, u_m; (\theta_1, b_1))\}^2 , \tag{32}$$

with weights $w(h_l, u_m; (\theta_1, b_1))$ given by

$$\frac{|N(h_1, u_m)|}{\{\gamma(h_l, u_m; (\theta_1, b_1))\}^2} . \tag{33}$$

In (32) and (33), $\gamma(h_l, u_m; (\theta_1, b_1))$ is given by (29) with the previously fitted parameters fixed at their respective estimates.

The resulting parameter estimates are $(\hat{\theta}_1, \hat{b}_1) = (0.4920, 0.0019)$, which, along with $(\hat{c}_1, \hat{c}_2, \hat{\theta}_2, \hat{b}_2) = (0, 0, 1.100, 0.045)$, are substituted into (29) to yield the fitted model. This model is displayed in Figure 4 with fitted values $(\hat{c}_1, \hat{c}_2, \hat{\theta}_1, \hat{\theta}_2, \hat{b}_1, \hat{b}_2) = (0, 0, 0.4920, 1.100, 0.0019, 0.045)$.

Optimal Prediction on the Stream Network

The data base shows 157 observations of (log) nitrate concentrations at 17 stream sampling sites during a 15-month period. What log concentration could be expected anywhere along the stream network and on any day, given rainfall amounts at the six precipitation gauges shown in Figure 1 and management practices at the upstream dairies? The statistical model fitted above allows this question to be answered with known confidence.

Recall that the model fitted was:

$$\mathbf{Z} = X\hat{\beta}_{wls} + \hat{\delta} ,$$

where the residual process $\hat{\delta}(\mathbf{s}, t)$ serves as a proxy for the spatially-dependent error process $\delta(\mathbf{s}, t)$ defined in (21). The spatial and temporal dependence in the residual process has been characterized by the variogram $2\gamma(h, u)$, estimated in the previous subsection. Using the theory of geostatistics (e.g., Cressie, 1993, Part I), it is possible then to predict (krige) a value of Z at stream location $\mathbf{s}_0$, which is not necessarily a sampling site, at time t_0. It is also possible to quantify the precision of the predictor through its root mean-squared prediction error (kriging standard error).

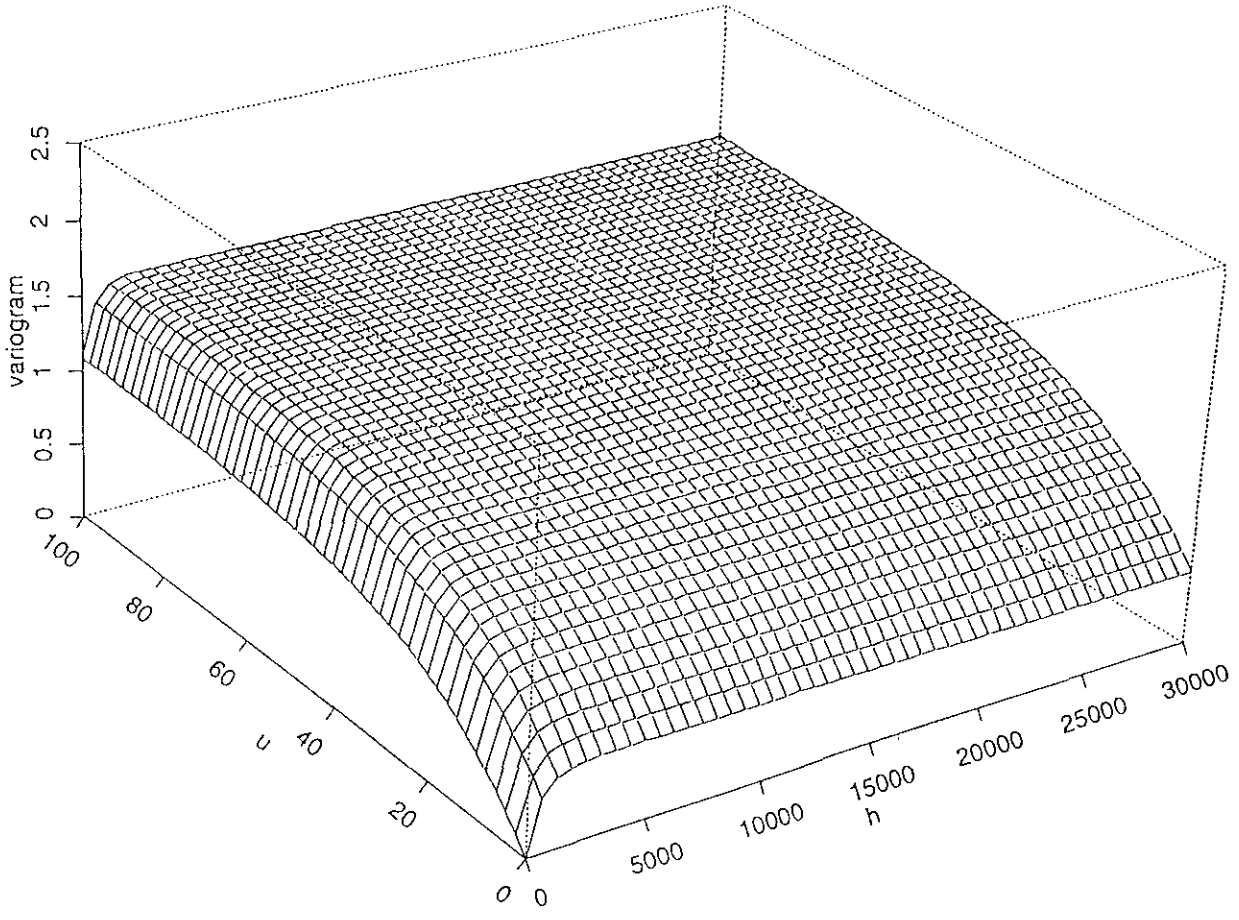


Figure 4. Surface of fitted variogram model.

The approach carried out here is to implement ordinary kriging on the residual process (small-scale variation) and then to add back the residual predicted value to the estimated mean process (large-scale variation) to give a final predicted value and its associated root mean-squared prediction error. The first stage involves predicting $\hat{\delta}(\mathbf{s}_0, t_0)$, the residual log nitrate concentration at a stream location $\mathbf{s}_0$ and a time t_0, from "data" $\{\hat{\delta}(\mathbf{s}_i, t_i) : i = 1, \ldots n\}$ or, equivalently, to predict $\hat{\nu}(\mathbf{s}_0, t_0)$ from "data" $\{\hat{v}(\mathbf{s}_i, t_i) : i = 1, \ldots n\}$, where n is the number of samples taken upstream or downstream from site $\mathbf{s}_0$ and within the past sixty days of time t_0, and is often much less than 157; see equations (23) and (25) for definitions. Write the residual and standardized

residual observations as $\hat{\delta}$ and $\hat{\nu}$, respectively.

Following the development in Cressie (1993, Section 3.2), define

$$\Gamma_O \equiv \begin{cases} \gamma(h_{ij}, u_{ij}) & , \quad i = 1, \ldots, n, \; j = 1, \ldots, n \\ 1 & , \quad i = n+1, \; j = 1, \ldots, n \\ 0 & , \quad i = n+1, \; j = n+1 \; , \end{cases} \tag{34}$$

where the remaining entries of Γ_O are filled in by symmetry. Also define

$$\gamma_O \equiv (\gamma(h_{01}, u_{01}), \ldots, \gamma(h_{0n}, u_{0n}), 1)' \; . \tag{35}$$

The subscripted lags in (34) and (35) are given by $h_{ij} \equiv \|\mathbf{s}_i - \mathbf{s}_j\|$ and $u_{ij} \equiv |t_i - t_j|$; $0 \leq i, j \leq n$. The ordinary kriging predictor is given by,

$$p_\nu(\mathbf{s}_0, t_0) \equiv \sum_{i=1}^{n} \lambda_i \hat{\nu}(\mathbf{s}_i, t_i) \equiv \boldsymbol{\lambda}' \boldsymbol{\nu} \; , \tag{36}$$

subject to the unbiasedness restriction,

$$\sum_{i=1}^{n} \lambda_i = 1 \; . \tag{37}$$

Using the method of Lagrange multipliers, the optimal coefficients in (36) are given by

$$(\lambda_1, \ldots \lambda_n, m)' = \Gamma_O^{-1} \gamma_O \; , \tag{38}$$

where m is the Lagrange multiplier that preserves the unbiasedness condition (37). The root mean-squared prediction error (i.e., kriging standard error) of the optimal predictor (36), (38) is given by

$$m_\nu(\mathbf{s}_0, t_0) = \{\gamma_O' \Gamma_O^{-1} \gamma_O\}^{1/2} \; . \tag{39}$$

From (25), the ordinary kriging predictor of $\hat{\delta}(\mathbf{s}_0, t_0)$ is

$$p_\delta(\mathbf{s}_0, t_0) = A(s_0)^{-0.08} p_\nu(\mathbf{s}_0, t_0) = A(s_0)^{-0.08} (\boldsymbol{\lambda}' \hat{\boldsymbol{\nu}}) \; , \tag{40}$$

where $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_n)'$ is obtained from (38) and the kriging standard error is

$$m_\delta(\mathbf{s}_0, t_0) = A(s_0)^{-0.08} m_\nu(\mathbf{s}_0, t_0) \; . \tag{41}$$

From (23), one may finally conclude that the optimal predictor of $Z(\mathbf{s}_0, t_0)$ is

$$p_Z(\mathbf{s}_0, t_0) = \mathbf{x}(\mathbf{s}_0, t_0)' \hat{\boldsymbol{\beta}}_{wls} + p_\delta(\mathbf{s}_0, t_0) \; . \tag{42}$$

The kriging standard error is, ignoring the smaller order effect due to the estimation of β,

$$m_Z(\mathbf{s}_0, t_0) = m_\delta(\mathbf{s}_0, t_0) \ , \tag{43}$$

which is the same as (41). As a consequence, an approximate 95% prediction interval for $Z(\mathbf{s}_0, t_0)$ is

$$p_Z(\mathbf{s}_0, t_0) \pm 2m_Z(\mathbf{s}_0, t_0) \ . \tag{44}$$

As noted in the previous subsection, because of the form of the variogram model (equation (29)), the matrix given by (34) can have singularities that make its inverse indeterminate. This situation, which is caused by particular configurations of sample locations (in space and time), did occur in this analysis. It was corrected for by jittering or perturbing the time coordinate u by a small amount. This is scientifically meaningful, because temporal readings are only precise to the nearest 24 hours and so the exact time the recording was taken is unknown. Matrices were then invertible and prediction could proceed.

Spatio-Temporal Prediction in the Upper North Bosque Watershed

Predictions were made at three sampling sites, sites 7, 11, and 12, on a date (June 10, 1992) that was beyond the period of record used to develop the model but for which samples were available. This was done in order to validate the predictive capability of the model. For all sites (7, 11, and 12), the observed values for the standardized log-nitrate concentration fall within their approximate 95% prediction intervals given by (44). Table 2 shows the observed and predicted values.

Table 2. Observed and predicted values for June 10, 1992

Site	Observed NO_3 (mg/l)	Observed Stzd Log	Predicted Regression	Predicted Error	Predicted Stzd Log	Pred Std Error	Interval (low end)	Interval (high end)
7	2.78	0.3022	0.7678	0.3132	1.081	0.6483	-0.2155	2.3776
11	3.77	0.8016	1.0678	0.7345	1.8023	0.7011	0.4001	3.2044
12	4.44	1.1300	1.5969	0.7271	2.324	0.6161	1.0919	3.5562

In this table, *observed* $NO_3(mg/l)$ is the observed nitrate concentration; *Observed Stzd Log* is the standardized log of the observed nitrate concentration; *Predicted Regression* is the first component of (42), due to the explanatory variables and estimated regression parameters; *Predicted Error* is the second component of (42), due to the spatio-temporal prediction of residual values; *Predicted Stzd Log* is the sum of the previous two values, namely the predicted value of

the standardized log-nitrate concentration given by (42); *Pred Std Error* is given by (43); and *Interval (low end)* and *Interval (high end)* are the ends of the approximate 95% prediction interval given by (44).

Because the concentration observed at site 12 for June 10 fell very near the end of the 95% prediction interval and because of the site's close proximity to the Stephenville, TX waste-water-treatment plant, cross-validation was carried out for site 12 on each of the dates for which data were available. Cross-validation was conducted by predicting the log-nitrate concentration without using the sample collected at the site on the date for which prediction was carried out. The results are shown in Table 3. The predictions for twelve of the thirteen dates fell within the 95% prediction interval. Furthermore, on seven of the thirteen dates, the prediction was lower than the observed value, while on the remaining six dates, the prediction was higher than the observed value. The conclusion from this is that the model is neither consistently underpredicting nor consistently overpredicting the actual value at site 12.

Table 3. Results of cross-validation for site 12

Date	Observed Stzd Log	Predicted Regression	Predicted Error	Predicted Stzd Log	Pred Std Error	Interval (low end)	Interval (high end)
4/4/1991	1.3909	1.5241	-0.4343	1.0898	0.4666	0.1565	2.0231
5/8/1991	0.8590	1.5241	-0.4325	1.0916	0.463	0.1655	2.0177
6/5/1991	2.0147	1.5241	0.1468	1.6709	0.4641	0.7426	2.5992
7/2/1991	1.9413	1.2622	-0.4258	0.8364	0.466	-0.0955	1.7683
8/14/1991	1.1051	1.2622	0.2427	1.5049	0.4522	0.6005	2.4092
9/4/1991	1.6890	1.2622	-0.4622	0.8	0.4472	-0.0945	1.6944
10/2/1991	2.0114	1.786	-0.5624	1.2236	0.4472	0.3293	2.1179
11/18/1991	2.0640	1.786	-0.2764	1.5096	0.4534	0.6028	2.4163
12/6/1991	2.1861	1.786	-0.1572	1.6288	0.454	0.7209	2.5368
1/15/1992	2.1670	2.0479	0.3557	2.4036	0.448	1.5075	3.2996
3/11/1992	1.9894	2.0479	0.0447	2.0925	0.4489	1.1948	2.9903
4/8/1992	2.0333	1.5241	0.6357	2.1598	0.4489	1.2621	3.0575
5/7/1992	2.1375	1.5241	0.6846	2.2087	0.4488	1.3111	3.1064

Predictions were also made at nineteen non-sampling sites, located from just upstream of site 11 to just downstream of site 12, which represents a stream distance of approximately 10 km. This was done in order to illustrate the predictive capabilities of the model at locations for which samples were not taken. All data either upstream or downstream from the prediction site and recorded within 60 days of the prediction date were used in the kriging equations (42), (43), and (44). Consequently, the number of data n used for prediction varied according to prediction site and prediction date. Figure 5 illustrates the locations of the three sampling sites and the nineteen

additional sites, numbered 50 through 69, at which predictions were made.

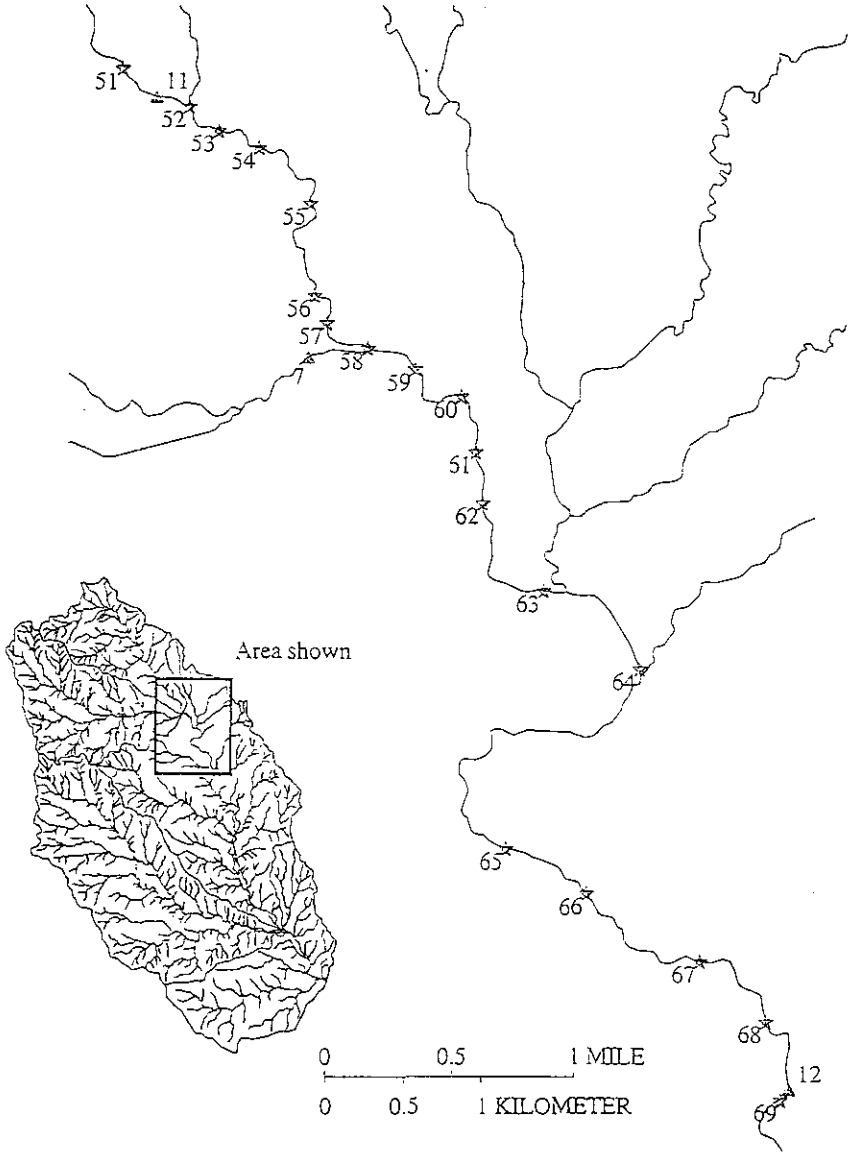


Figure 5. Locations of three sampling sites (numbered 7, 11, and 12) and nineteen prediction sites (numbered 51 through 69)

All significant explanatory variables (e.g., distance to basin outlet, number of dairies per acre) were calculated for each site and prediction was conducted for June 8, 1991; the results are presented graphically in Figure 6. Figure 6a shows the predicted values of standardized log-nitrate concentrations and Figure 6b shows the prediction standard error of the predictions, given by (43). The graduated symbols in these figures are calibrated as follows: The diameter of

the circle representing the smallest predicted value (-0.8837) is put equal to 0.1 inches and the diameter of the circle representing the largest value (2.6735) is put equal to 0.4 inches. Diameters in between are interpolated exponentially with an exponent of 0.75 chosen in accordance with Stevens' law of visual perception (Stevens, 1975).

Other dates were chosen for prediction, June 7 and 9, 1991, and June 10, 1992, as well as June 8, 1991. The full table of results, analogous to Tables 2 and 3, can be found in Majure (1995).

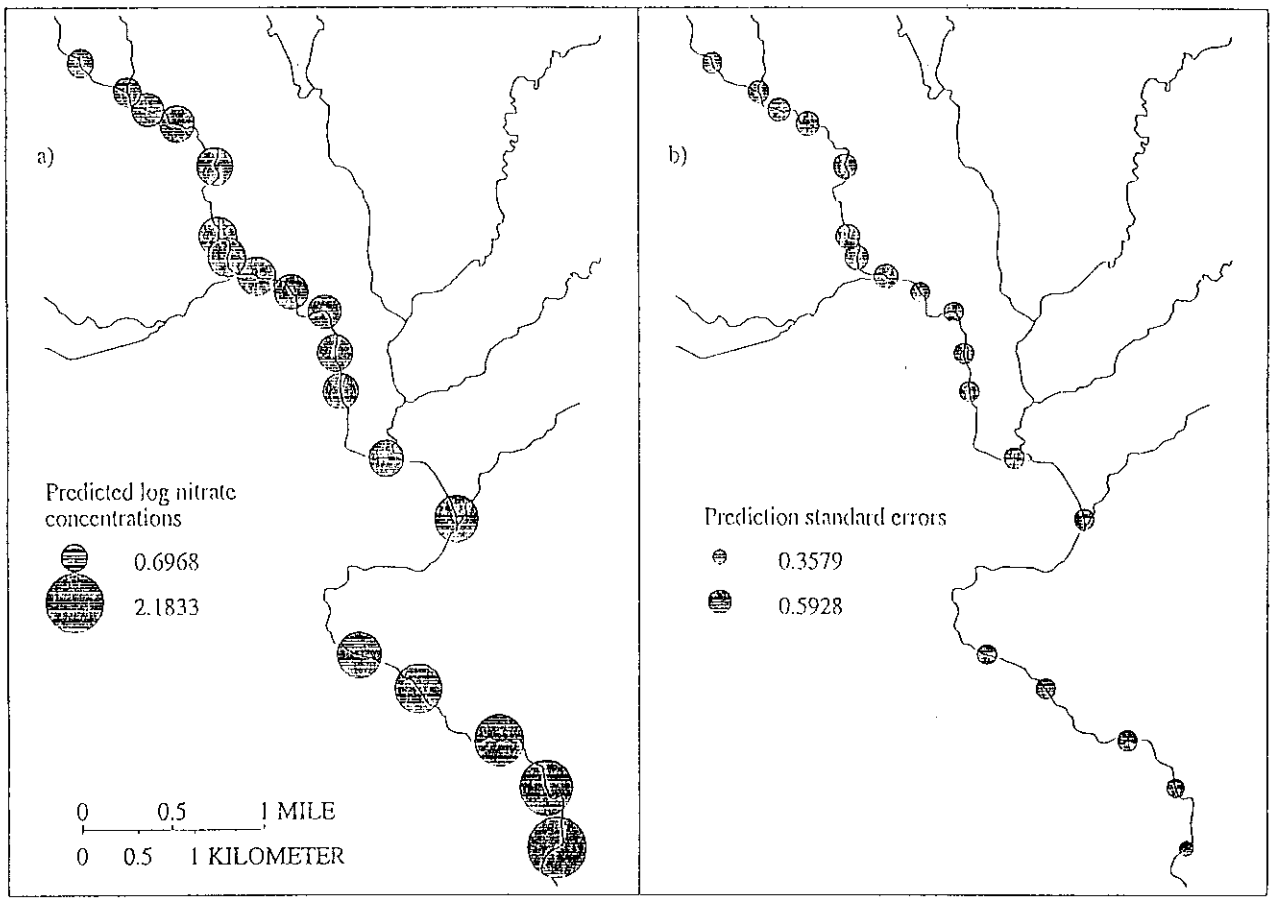


Figure 6. a) Predicted standardized log-nitrate concentrations
b) Prediction standard errors

Discussion

The spatio-temporal model presented allows the prediction of contaminant concentrations in space and time with known confidence. This model presents a novel approach to the use of GIS and statistical modeling to solve the spatio-temporal prediction problem.

One aspect of the results of the model that might be somewhat discouraging to some is the large variation of the predicted values. If, for example, one converts the prediction interval calculated for site 12 on June 10, 1992 (see Table 2) back into mg/l, the interval becomes 1.0094–48.8436, which is quite large. This variation can be largely explained by the low sampling density in space and time. An examination of the fitted variogram model indicates that, as the distance increases in either space *or* time from a prediction point to a sample point, the variogram values increase quickly to twice the overall variance of the data. Thus, when predicting, if there are no samples in close proximity (in space *and* time) to the prediction site, the prediction standard error will be large. In the current study, the sampling design of monitoring locations (in both space and time) was largely *ad hoc*, and certainly was not implemented with this kind of prediction exercise in mind. Having said this, the prediction intervals do provide important information, namely, away from the sample locations and times prediction is very difficult to accomplish with a high degree of precision.

An important assumption of the model is that the residual process defined in (4.5) has zero mean and is intrinsically stationary. That is, it represents only zero-mean, spatio-temporal statistical variation and not deterministic variation. However, there is one source of deterministic variation that is not explicitly accounted for, due to lack of appropriate data. Land-use distribution in the watershed is important for several reasons. First of all, it might have provided information on additional potential sources of contamination in the form of agricultural fields upon which chemical fertilizers are applied. Second, it has an important effect on the transport of materials (specifically nutrients) from the land surface to the surface-water system. Land use also has a dramatic affect on the rate of movement of water over and through a landscape; Maidment (1993) uses it to control flow velocity in the development of a spatially explicit hydrograph, which is essentially what we are emulating with the precipitation variable.

The spatio-temporal statistical model has appealing characteristics in spite of the problems described above. First, the model is statistical and, therefore, provides estimates of the precision of predictions. Large prediction intervals simply indicate that the quantity is difficult to predict accurately in time and space. Another appealing feature is the model's ability to take physical conditions into account through easy quantification of physical landscape characteristics in a GIS.

Finally, the model makes use of nearby samples (in space and time) to obtain the best predictor.

REFERENCES

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

Cressie, N. 1993. *Statistics for Spatial Data.* Wiley, NY.

Cressie, N. and J.J. Majure. 1996. Non-point-source pollution of surface waters over a watershed. In *Statistics for the Environment 3: Aspects of Pollution*, eds V. Barnett and K.F. Turkman. Wiley, Chichester, forthcoming.

Ericksen, E.P., J.B. Kadane, and J.W. Tukey. 1989. Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association*, 84, 927–944.

Hauck, L. 1993. Personal communication. Texas Institute for Applied Environmental Research, Stephenville, TX.

Jenson, S.K. and J.O. Dominque. 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric Engineering and Remote Sensing*, 54, 1593–1600.

Johengen, T.H. and A.M. Beeton. 1992. The effects of temporal and spatial variability on monitoring agricultural nonpoint source pollution. In Proceedings of the *National Rural Clean Water Program Symposium.* USEPA Seminar Publication, Washington, DC, 89–95.

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

Libby, L.W. and W.G. Boggess. 1990. Agriculture and water quality: Where are we and why? In *Agriculture and Water Quality International Perspectives*, eds J.B. Braden and S.B. Lovejoy. Lynne Riener Publishers, Boulder, CO, 1–37.

Long, C.M. 1992. Livestock waste pollution: A nationwide problem. *Internal Report.* Office of Policy, Planning and Evaluation, U.S. Environmental Protection Agency, Washington, DC.

Maidment, D.R. 1993. Developing a spatially distributed unit hydrograph by using GIS. *HydroGIS'93.* International Association of the Science of Hydrology Publication No. 211, 181–192.

Majure, J.J. 1995. A Spatio-Temporal Statistical Model of Pollutant Concentrations in Surface Waters. M.S. Thesis, Department of Statistics, Iowa State University, Ames, IA 50011.

Myers, D.E. and A.G. Journel. 1990. Variograms with zonal anisotropies and noninvertible kriging systems. *Mathematical Geology*, **22**, 779–785.

Rouhani, S. and D.E. Myers. 1990. Problems in space-time kriging of geohydrological data. *Mathematical Geology*, **22**, 611–623.

Stevens, S.S. 1975. *Psychophysics.* Wiley, NY.

TIAER, 1992. *Livestock and the Environment. Rethinking Environmental Policy, Institutions, and Compliance Strategies.* Interim Report to the Joint Interim Committee on the Environment, 72nd Texas Legislature. TIAER, Tarleton State University, Stephenville, TX.

White, D.A., R.A. Smith, C.V. Price, R.B. Alexander, and K.W. Robinson. 1992. A spatial model to aggregate point-source and non-point-source water-quality data for large areas. *Computers and Geosciences*, **18**, 1055–1073.