2006

Spatial statistical models that use flow and stream distance

Jay M. Ver Hoef
Alaska Department of Fish and Game, 1300 College Road, Fairbanks, AK 99701, USA

Erin Peterson
Department of Geosciences, Colorado State University, Fort Collins, CO, USA

David Theobald
Natural Resources Ecology Lab, Colorado State University, Fort Collins, CO, USA

Environ Ecol Stat (2006) 13:449–464
DOI 10.1007/s10651-006-0022-8

ORIGINAL ARTICLE

Spatial statistical models that use flow and stream distance

Jay M. Ver Hoef · Erin Peterson · David Theobald

Received: July 2005 / Revised: March 2006

Abstract We develop spatial statistical models for stream networks that can estimate relationships between a response variable and other covariates, make predictions at unsampled locations, and predict an average or total for a stream or a stream segment. There have been very few attempts to develop valid spatial covariance models that incorporate flow, stream distance, or both. The application of typical spatial autocovariance functions based on Euclidean distance, such as the spherical covariance model, are not valid when using stream distance. In this paper we develop a large class of valid models that incorporate flow and stream distance by using spatial moving averages. These methods integrate a moving average function, or kernel, against a white noise process. By running the moving average function upstream from a location, we develop models that use flow, and by construction they are valid models based on stream distance. We show that with proper weighting, many of the usual spatial models based on Euclidean distance have a counterpart for stream networks. Using sulfate concentrations from an example data set, the Maryland Biological Stream Survey (MBSS), we show that models using flow may be more appropriate than models that only use stream distance. For the MBSS data set, we use restricted maximum likelihood to fit a valid covariance matrix that uses flow and stream distance, and then we use this covariance matrix to estimate fixed effects and make kriging and block kriging predictions.

J. M. Ver Hoef
Alaska Department of Fish and Game, 1300 College Road, Fairbanks, AK 99701, USA

E. Peterson
Department of Geosciences, Colorado State University, Fort Collins, CO, USA

D. Theobald
Natural Resources Ecology Lab, Colorado State University, Fort Collins, CO, USA

Present Address:
J. M. Ver Hoef (✉)
National Marine Mammal Laboratory, 7600 Sand Point Way NE, Bldg 4 Seattle, WA 98115-6349, USA
e-mail: jay.verhoef@noaa.gov

Keywords Stream networks · Valid autocovariances · Geostatistics · Variogram · Block kriging

1 Introduction

Streams and rivers form one of the most important environmental resources in a nation. Clean water is vital for drinking, and it provides habitat for plants and animals. A lot of time and money has been spent to characterize and monitor streams and rivers (see, e.g., Torgersen et al. 2004; Yuan 2004). As with most environmental and ecological data, a sample must be taken from a possibly infinite population of values on a stream network. For example, sample units could be counts of fish from a small stream segment, or water quality samples. Often, the area of interest is larger than a single stream segment, encompassing a whole stream network. The goals of data collection from streams may be varied, but often include (1) predicting at unsampled locations, (2) predicting an average or total for a stream segment or a whole stream network, and (3) estimating relationships between the response variable and other covariates. Our overall goal is to develop spatial statistical models for stream networks that can accomplish these goals.

A general formulation to handle all of the goals listed above is a spatial linear model. Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the dimension of $\mathbf{Y}$ is an n by 1 vector. The relationship between the response variable and covariates is modeled through the design matrix $\mathbf{X}$ and parameters $\boldsymbol{\beta}$. The classical assumption is that the random errors ϵ are independent, so $\text{var}(\boldsymbol{\epsilon})$ is $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. In spatial statistics, the independence assumption is relaxed and values are allowed to be correlated, so, in general, $\text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$. When used for spatial prediction, this model is referred to as "universal" kriging (Cressie 1993, p. 151), with "ordinary" kriging being the special case where the design matrix $\mathbf{X}$ is a single column of ones. The general formulation of the covariance matrix $\boldsymbol{\Sigma}$ has too many parameters to estimate. Using assumptions like ergodicity and stationarity (Cressie 1993, p. 57), distance can be used to reduce the number of parameters. For example, a spherical autocovariance model is,

$$C(h;\theta_0,\theta_1,\theta_2) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0, \\ \theta_1\left[1 + \frac{1}{2}\left(\frac{h}{\theta_2}\right)^3 - \frac{3}{2}\frac{h}{\theta_2}\right] & \text{if } 0 < h < \theta_2, \\ 0 & \text{if } \theta_2 \le h, \end{cases} \tag{1}$$

where h is Euclidean distance. Thus, we have used distance to reduce the number of covariance parameters from $\frac{n(n+1)}{2}$ to 3. The key word here is "Euclidean."

When working with stream networks, we may not want to use Euclidean distance. An attractive alternative is to use stream distance. Transport of materials and movements of fish only occur within the stream network, so this may be a more appropriate distance metric when modeling autocovariance. Stream distance is defined as the shortest distance between two locations, where distance is computed only along the stream network. However, spatial autocovariance models developed for Euclidean distance may not be valid for stream distances.

As a simple example of what can go wrong, consider Fig. 1. Imagine an idealized stream network starting with a single node splitting into two, and then each node splitting into two more, etc., for $2^6 - 1 = 63$ nodes. Suppose that the stream segments

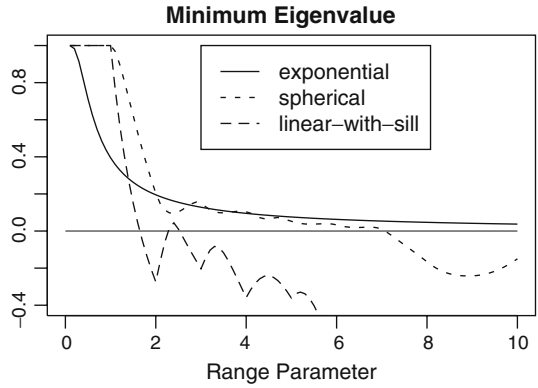

Fig. 1 An example of getting invalid covariance matrices when using stream distance for autocorrelation models that were developed for Euclidean distance

between nodes are all one unit long. Now consider the use of a standard autocovariance model when using stream distance, rather than Euclidean distance. Three different autocovariance models were used based on stream distance, one being the spherical model described in (1). Another is the exponential autocovariance model,

$$C(h;\theta_0,\theta_1,\theta_2) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0, \\ \theta_1 \exp(-h/\theta_2) & \text{if } 0 < h. \end{cases} \tag{2}$$

Another model that is valid for distance in one dimension, but not Euclidean distance in two or more dimensions, is the linear-with-sill model,

$$C(h;\theta_0,\theta_1,\theta_2) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0, \\ \theta_1(1 - \frac{h}{\theta_2}) & \text{if } 0 < h < \theta_2, \\ 0 & \text{if } \theta_2 \leq h. \end{cases} \tag{3}$$

All three models $(1-3)$ were used in Fig. 1. The parameter θ_0 has been termed the "nugget" effect, and was set to zero, and the parameter θ_1, often called the partial sill, was set to 1 for all models in Fig. 1. The parameter that controls the amount of autocorrelation is θ_2, and it was allowed to vary. For each value of θ_2 for each model, a covariance matrix was determined based on stream distance among the 63 nodes in the example described earlier. The minimum eigenvalue as a function of θ_2 is shown in Fig. 1. Notice that negative eigenvalues mean that the covariance matrix is not positive definite, and hence not valid. Figure 1 demonstrates that the spherical and linear-with-sill models are not valid when using stream distance. Even so, Gardner et al. (2003) make kriging predictions using stream distance with a spherical autocovariance model, which is not generally valid. In Fig. 1, the exponential model seems valid, and this is clarified in Sect. 2.4. In general, we have been unable to find any literature on making kriging predictions using valid models based on stream distance.

One problem with developing spatial models for stream networks is that little is known about valid autocovariance models when using stream distances. A second problem is that stream distance alone may not be appropriate for modeling autocorrelation. Streams have flow, which is characterized by direction and volume. In general, we will use flow to mean direction, unless specifically stating flow volume. For some types of variables, such as stream chemistry values, we might want to consider models that do not allow autocorrelation between locations if the water at one location does not flow into another. For example, Fig. 2 shows

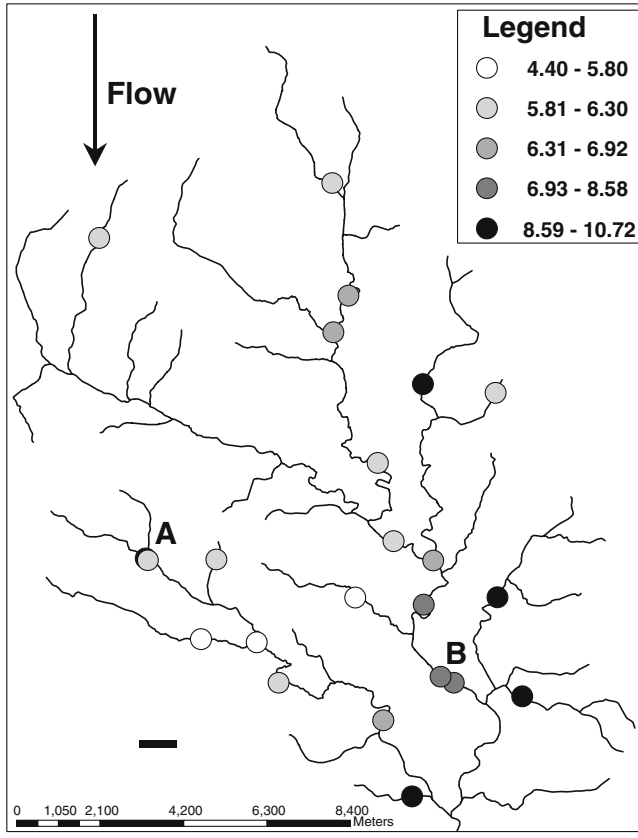

Fig. 2 An example of how flow affects stream chemistry values. Note that there are two locations in close proximity with overlapping circles near A

concentration measurements of SO_4 (in μeq/L) from a stream network in Maryland (http://www.dnr.state.md.us/streams/mbss/). In Fig. 2, there are locations at stream junctions, such as in the area labeled A, that are very close spatially, even when using stream distance, but the values are quite different because they are on two stream segments that do not share flow. The values downstream of junctions are intermediate in value from their upstream segments, as one would expect when the waters mix. In comparison, the values within a stream segment, such as the two values labeled with B in Fig. 2 are very similar.

To meet the earlier stated goal of developing spatial linear models on streams, the primary objective of this paper is to develop valid spatial autocovariance models that incorporate flow and use stream distance. Several papers have investigated the use of non-Euclidean distance for aquatic resources, including for streams and estuaries, but none has developed valid autocovariance models that incorporate stream distance and flow. Yuan (2004) discusses the problem, but uses Euclidean distance anyway because of the lack of valid models. Dent and Grimm (1999) and Torgersen et al. (2004) estimate a spherical covariance model based on stream distance in an exploratory fashion, but they do not attempt kriging predictions (which could have yielded negative variance estimates). Other approaches develop complicated mean structures

among locations, hoping to achieve independence among random errors. An example is the SPARROW approach (Smith et al. 1997). Non-Euclidean metrics have been tried in estuaries as well (Rathbun 1998; Little et al. 1997). A space-time model is developed in Cressie and Majure (1997). Curriero (1996) discusses an approach to the problem that uses metric dimensional scaling. The general idea is to use a matrix based on stream distance and map it into a low dimensional Euclidean space while trying to minimize distortion. Once we are in Euclidean space, we can use the usual spatial autocovariance models (if they are valid for that dimension). One problem with this method is that the addition of new prediction locations can change the model and all other predictions, even when the data have not changed. In addition, it does not include flow.

In this manuscript we use moving average constructions (also called kernel convolutions) to develop valid models for stream networks. In Sect. 2, we introduce these models, first developing a mathematical framework and notation, and then constructing random variables on streams and deriving their autocovariance. We concentrate on models that incorporate flow, but also show one based purely on stream distance. In Sect. 3, we use real data to fit a spatial linear model, where we use restricted maximum likelihood (REML) (see Cressie 1993, pp. 92–93 for REML applied to spatial models) to estimate the covariance parameters, and then estimate fixed effects and make kriging and block kriging predictions. We conclude with some discussion and future directions.

2 Moving average constructions

Barry and Ver Hoef (1996) show that a large class of autocovariances can be developed by creating random variables as the integration of a moving-average function over a white-noise random process,

$$Z(s) = \int_{-\infty}^{\infty} g(x - s|\boldsymbol{\theta}) W(x) \mathrm{d}x, \tag{4}$$

where $W(x)$ is a white noise process and $g(x|\boldsymbol{\theta})$ is called the moving-average function and it is defined on $\mathcal{R}^1$. We are free to choose the moving average function, but it must have finite volume in order to create a stationary process. Typically, we choose functions centered on 0, where most of their mass occurs as well. Several examples are given in Table 1. The moving-average construction allows a valid autocovariance to be expressed as,

$$C(h|\boldsymbol{\theta}) = \begin{cases} \int_{-\infty}^{\infty} (g(x|\boldsymbol{\theta}))^2 \mathrm{d}x + \nu_j^2 & \text{if } h = 0, \\ \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta}) g(x - h|\boldsymbol{\theta}) \mathrm{d}x & \text{if } h > 0, \end{cases} \tag{5}$$

where we assume that the integrals exist. We allow a discontinuity ν_j^2 at $h = 0$, which is the "nugget" effect in geostatistical terminology (see Cressie 1993, p. 59), and was labeled θ_0 in (1–3). There is increasing use of these moving average models to construct valid autocovariances when confronted with new problems, such as multivariate models (Ver Hoef and Barry 1998; Ver Hoef et al. 2004), and nonstationary models (Higdon 1998; Higdon et al. 1999; Fuentes 2002). We will use the moving average construction to build valid models for streams, and these models will also account for water flow.

Table 1 Moving average functions and their corresponding autocorrelation functions

Name	Moving average function	Autocorrelation function
Linear with sill	$g(x) = 1$ $I(0 \le x \le 1)$	$\rho(h) = \begin{cases} 1-h & \text{if } 0 \le h < 1 \\ 0 & \text{if } 1 \le h \end{cases}$
Spherical	$g(x) = 1 - x$ $I(0 \le x \le 1)$	$\rho(h) = \begin{cases} 1 - \frac{3}{2}h + \frac{1}{2}h^3 & \text{if } 0 \le h < 1 \\ 0 & \text{if } 1 \le h \end{cases}$
Mariah	$g(x) = \frac{1}{x+1}$ $I(0 \le x)$	$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\ln(h+1)}{h} & \text{if } 0 < h \end{cases}$
Exponential	$g(x) = e^{-x}$ $I(0 \le x)$	$\rho(h) = e^{-h}$ if $0 \le h$

2.1 Mathematical framework and notation

To construct random variables similar to (4), but based on stream networks and flow, we first develop a mathematical framework and notation. We assume that the scale of the problem is such that we can depict a stream segment as a line, and the lines of a stream form a network, as in Fig. 3. We take these to be segments of the real line, and we consider a location downstream to be a lower real number than a location farther upstream. The whole network will have a single most-downstream location, which we set to 0. Any location on a stream network can be connected by a continuous line to the lowest point in that network, and hence distance from the lowest point is simply the length of that line. We define this as "distance upstream." In a stream network, there will be a finite number of stream segments, and we index them arbitrarily with $i = 1, 2, \ldots$. In a branching stream network many locations will have the same distance upstream, so in order to uniquely define each location and keep track of distance upstream, we denote each location as x_i, which is the distance upstream on the i^{th} stream segment. We will denote the most downstream location on the i^{th} segment as l_i, and we will denote the most upstream location as u_i, where u_i could be ∞. In order to consider the widest range of models, if there are no more stream segments upstream of a stream segment (e.g., segment $i = 2$ in Fig. 3), then we will consider a segment such as this to be defined on (l_i, ∞); in Fig. 3, this is denoted as an arrow.

Let the whole set of stream segment indices be denoted as I. The index set of stream segments upstream of x_i, excluding i, will be $U_{x_i} \subseteq I$. In Fig. 3 for example, $U_{s_3} = \{4, 5, \ldots, 13\}$, $U_{t_8} = \{9, 10\}$, and $U_{r_{11}} = \{12, 13\}$. Besides working with point locations, we also work with whole segments, so we define the index set of stream segments upstream of segment i, excluding i, as $U_{[i]} \subseteq I$. It will also be useful to define $D_{x_i} \subseteq I$ as the index set of all stream segments downstream of x_i into which x_i flows, including the segment containing x_i. In Fig. 3 for example, $D_{s_3} = \{1, 3\}$, $D_{t_8} = \{1, 3, 5, 6, 8\}$ and $D_{r_{11}} = \{1, 3, 5, 11\}$. Similarly, the index set of stream segments downstream of segment i, including i, is denoted $D_{[i]} \subseteq I$.

Using these definitions, we can say that two locations, s_i and t_j, on a stream network are "flow-connected" if $D_{s_i} \cap D_{t_j} = D_{s_i}$ or D_{t_j}. In a similar way, we can obviously define flow-connected between a location and a stream segment, or between two stream segments. For example, in Fig. 3, s_3 and t_8 are flow-connected because $D_{s_3} \cap D_{t_8} = D_{s_3}$, whereas t_8 and r_{11} are not flow-connected because $D_{r_{11}} \cap D_{t_8} \neq D_{r_{11}}$ nor D_{t_8}. Finally, consider

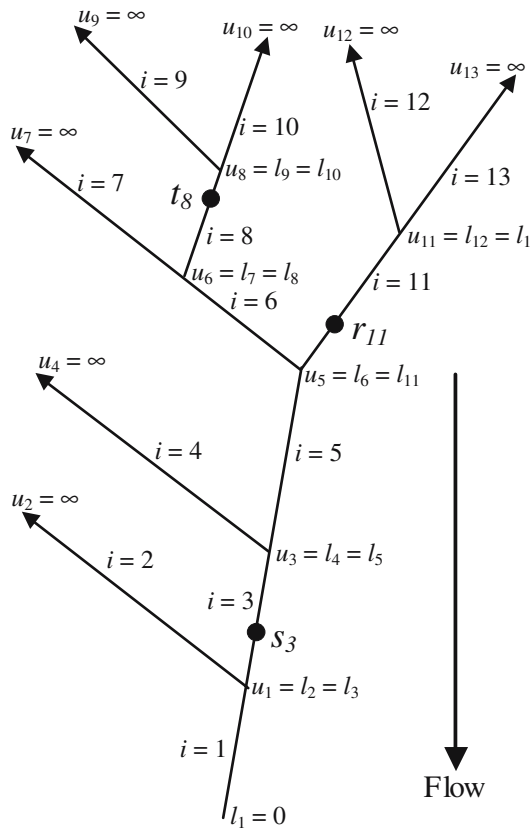

Fig. 3 Example stream network, with 13 stream segments labeled with i. The arrows at the upper end of stream segments indicate that they are assumed to have infinite length. Three locations on the network, s_3, t_8, and r_{11} are shown with solid circles

$$B_{s_i,t_j} \equiv \begin{cases} \overline{(D_{s_i} \cap D_{t_j})} \cap (D_{s_i} \cup D_{t_j}) & \text{if } s_i \text{ and } t_j \text{ are flow-connected,} \\ \emptyset & \text{otherwise,} \end{cases}$$

which can be thought of as the set of stream segments *between* two locations, including the segment for the upstream location but excluding the segment for the downstream location. In an obvious way we can also define $B_{s_i,[j]}$ and $B_{[i],[j]}$. In Fig. 3 for example, $B_{s_3,[8]} = \{5, 6, 8\}$. Note that $B_{s_i,t_i} = \emptyset$ because s_i and t_i are on the same segment, even though they are flow-connected.

We are now in position to define stream distance, whose common-sense meaning is the shortest distance between two locations on a stream network. We define this as

$$d(s_i, t_j) \equiv \begin{cases} |s_i - t_j| & \text{if } s_i \text{ and } t_j \text{ are flow-connected,} \\ (s_i - u) + (t_j - u) & \text{otherwise,} \end{cases}$$

where $u = \max\{u_k : k \in D_{s_i} \cap D_{t_j}\}$. For example, in Fig. 3, $d(s_3, t_8) = |s_3 - t_8|$ while $d(t_8, r_{11}) = (t_8 - u_5) + (r_{11} - u_5)$.

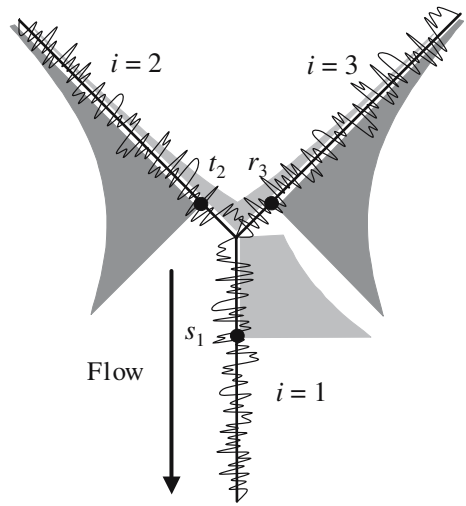

Fig. 4 Constructing random variables by using moving averages. Three stream segments, labeled $i = 1, 2, 3$ are shown. The wavy lines represent the white noise processes, and the shaded areas represent the moving average functions. Three locations, s_1, t_2, and r_3 are shown with solid circles

2.2 Constructing random variables

Next, we will build random variables like that given in (4), but for our stream network. To help develop the ideas, consider Fig. 4. The white noise process $W(x)$ in (4) is depicted as the wavy line around each line segment. For a location, such as that given by s_1, the moving average function is shown as the shaded function along the line segment. This moving average function could go in both directions, up and down the stream in relation to flow. In this paper, we primarily consider functions that only go upstream, as shown in Fig. 4.

The moving average would be a standard construction for a single line segment that was continuous from $-\infty$ to ∞ on the real line, such as for time series models. However, for stream networks, the line segments split into two. Hence, we split the moving average into two parts, as shown in Fig. 4, with one part going up segment $i = 2$ and the other going up segment $i = 3$. We will do the integral in (4) piecewise, summing up all segments that contain the moving average function $g(x|\boldsymbol{\theta})$. Because of the "upstream" construction, we only need to take the integrals for the segments that are in i and U_{s_i}. If we want the random variable $Z(s_i)$ to be stationary, we need to take some care in the way that $g(x|\boldsymbol{\theta})$ gets split as we go upstream. That is, suppose that upstream of the variable defined at t_2 there are no further splits, whereas upstream of r_3 there are many splits. If we use no weighting at all, then the total area under $g(x|\boldsymbol{\theta})$ will be much greater for r_3 than for t_2. From (5), the variance of a random variable is $\int_{-\infty}^{\infty}(g(x|\boldsymbol{\theta}))^2 \mathrm{d}x + \nu_j^2$, so the variance of r_3 would be greater than for t_2.

The solution is to use weighting, and the weighting can also incorporate flow volume. In the absence of any flow volume characteristics, we can simply weight each split in the moving average function by $\sqrt{1/2}$. On the other hand, suppose that the flow volume from one upstream segment is larger than the flow volume for the other upstream segment, then we might want to weight according to flow volume. It is often difficult to measure flow volume for each stream segment, but a proxy variable such as the stream order can be used; or the area of each basin could be used, which can be obtained from digital elevation maps (DEMs) in a Geographical Information System

(GIS). In Fig. 4, if we weight the segment $i = 2$ with ω_2 and the segment $i = 3$ with ω_3, where $\omega_2 + \omega_3 = 1$, then we maintain stationarity of the variances by weighting with $\sqrt{\omega_2}$ and $\sqrt{\omega_3}$.

For a stream network then, the construction that is equivalent to (4) is

$$Z(s_i) = \int_{s_i}^{u_i} g(x_i - s_i|\boldsymbol{\theta})W(x_i)\mathrm{d}x_i + \sum_{j\in U_{si}} \left(\prod_{k\in B_{si,[j]}} \sqrt{\omega_k} \right) \int_{l_j}^{u_j} g(x_j - s_i|\boldsymbol{\theta})W(x_j)\mathrm{d}x_j. \quad (6)$$

For example, suppose we use the exponential moving average function from Table 1 for s_3 in Fig. 3. Then

$$\begin{aligned} Z(s_3) = &\int_{s_3}^{u_3} e^{-(x_3-s_3)}W(x_3)\mathrm{d}x_3 + \sqrt{\omega_4}\int_{u_3}^{\infty} e^{-(x_4-s_3)}W(x_4)\mathrm{d}x_4 \\ &+\sqrt{\omega_5}\int_{u_3}^{u_5} e^{-(x_5-s_3)}W(x_5)\mathrm{d}x_5 + \sqrt{\omega_5\omega_6}\int_{u_5}^{u_6} e^{-(x_6-s_3)}W(x_6)\mathrm{d}x_6 \\ &+\sqrt{\omega_5\omega_6\omega_7}\int_{u_6}^{\infty} e^{-(x_7-s_3)}W(x_7)\mathrm{d}x_7 + \cdots + \sqrt{\omega_5\omega_{11}\omega_{13}}\int_{u_{11}}^{\infty} e^{-(x_{13}-s_3)}W(x_{13})\mathrm{d}x_{13}, \end{aligned}$$

where recall that $\omega_4 + \omega_5 = 1$, $\omega_6 + \omega_{11} = 1$, etc.

2.3 Valid covariances based on stream distance and flow

From the definition in (6), we can use (5) to obtain valid autocovariance models for a stream network.

$$C(s_i, t_j|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ are not flow-connected,} \\ C_1(0) + \nu_j^2 & \text{if } s = t, \\ \prod_{k\in B_{s_i,t_j}} \sqrt{\omega_k} C_1(d(s_i, t_j)) & \text{otherwise.} \end{cases} \quad (7)$$

where $C_1(h) = \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta})g(x-h|\boldsymbol{\theta})\mathrm{d}x$ and recall that $d(s_i, t_j)$ is the stream distance between s_i and t_j on the stream network. The result in (7) might seem surprising at first. It basically says that we can use moving average models developed in one dimension, without any branching, as long as we use the proper weighting. Consequently, the final covariance matrix is quite easy to construct.

Consider any autocovariance function of distance in one dimension that can be developed using moving averages. Many commonly used models have moving average representations; some autocorrelation models are given in Table 1. Many of the models in Table 1 can be found in textbooks on geostatistics, such as Cressie (1993, p. 61) and Chiles and Delfiner (1999, p. 80). A general approach for working backwards from known, valid autocovariance models to the moving average function is given by Cressie and Pavlicova (2002). However, it is easy to create our own models. For example, consider $g(x) = 1/(x+1)$ so that $\int_{-\infty}^{\infty} g(x)g(x-h)\mathrm{d}x = \ln(h+1)/h$ for $h > 0$, and it is equal to 1 for $h = 0$. This model appears to be new, so we call it the MARIAH model (Moving Average Reciprocal 1 Add H), which is given in Table 1. From Table 1, a general method to develop a valid autocovariance from the autocorrelation function is by scaling the lag h, multiplying by a partial sill, and adding a nugget effect for $h = 0$; i.e.,

$$C_1(h|\boldsymbol{\theta}) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0, \\ \theta_1 \rho(h/\theta_2) & \text{if } h > 0. \end{cases} \quad (8)$$

Now we can develop a matrix $\mathbf{V}$ using stream distance for the functions given in Table 1 and (8). As noted earlier, this will not necessarily be a valid covariance matrix. However, if we use a weight matrix $\mathbf{A}$, we obtain valid autocovariance matrices by using the Hadamard (element-wise) product $\boldsymbol{\Sigma} = \mathbf{A} \odot \mathbf{V}$; the matrix $\mathbf{A}$ contains zeros whenever locations are not flow-connected, and when sites are flow-connected, $\mathbf{A}$ contains the square root of the percentage of flow volume (or other weightings) $\prod_{k \in B_{s_i,t_j}} \sqrt{\omega_k}$ that the downstream location receives from the upstream location, as given in (7). To maintain stationary variances, the two weightings should sum to one whenever there is a fork in the stream network. Assuming an exponential model in Table 1, for the example in Fig. 4, $\boldsymbol{\Sigma} = \mathbf{A} \odot \mathbf{V}$ is,

$$\begin{pmatrix} 1 & \sqrt{\omega_2} & \sqrt{\omega_3} \\ \sqrt{\omega_2} & 1 & 0 \\ \sqrt{\omega_3} & 0 & 1 \end{pmatrix} \odot \begin{pmatrix} \theta_0 + \theta_1 & \theta_1 e^{-d(s_1,t_2)/\theta_2} & \theta_1 e^{-d(s_1,r_3)/\theta_2} \\ \theta_1 e^{-d(s_1,t_2)/\theta_2} & \theta_0 + \theta_1 & \theta_1 e^{-d(t_2,r_3)/\theta_2} \\ \theta_1 e^{-d(s_1,r_3)/\theta_2} & \theta_1 e^{-d(t_2,r_3)/\theta_2} & \theta_0 + \theta_1 \end{pmatrix}.$$

2.4 Valid covariances based on stream distance only

So far, we have developed models by taking moving average functions that are asymmetric, defining them only on the positive real numbers, and then letting the "tails" of the functions run upstream only. Now, consider the case where the tails run downstream only. As long as the stream network has a single lowest point, all locations will have the possibility of non-zero covariance because part of the moving average functions could overlap, unlike the case when the tails run upstream. Because these models have the possibility of autocorrelation among all locations, they may be more appropriate for variables like fish abundance, because fish can swim upstream and could be less sensitive to flow.

We do not develop these models in detail here; they will be the subject of a future paper. We do mention one model that is easy to obtain, however, because it will make for an interesting comparison to the flow models. If we use the exponential moving average, as given in Table 1, but run the tails downstream, we obtain a model that only depends on stream distance. It is not difficult to show, for example, in Fig. 4, that if the stream distance between s_1 and t_2 is the same as the stream distance between t_2 and r_3, then $C(s_1, t_2) = C(t_2, r_3)$ if we use the exponential moving average with the tails running downstream. So far, we have only been able to show that this is true for the exponential moving average, and it is not true for the other models in Table 1. Thus, the exponential covariance model (2) is the only one that we know that is guaranteed to work using stream distance in the same way that we would use Euclidean distance. This also explains why the exponential model is valid in Fig. 1. Note that Curriero (1996) also found this to be the only valid model for networks such as roads.

3 Example

As an example, we use the data shown in Fig. 2. These data are concentration measurements of SO_4 (in μeq/L) from a stream network in Maryland, as described earlier. There were 23 sites. Ideally, we would have more locations, but this data set provides a useful illustration.

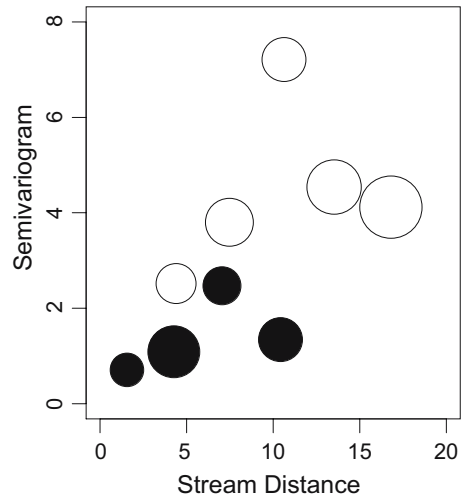

Fig. 5 Empirical semivariogram for data shown in Fig. 2. The solid circles show the semivariogram among locations that are flow-connected, and the open circles show the semivariogram among locations that are not flow-connected. Only lags with > 15 pairs are shown, and the sizes of the circles are proportional to the number of data pairs that are averaged for each value

We begin with an exploratory analysis of the data. In Fig. 5, we plot the empirical semivariogram for the data from Fig. 2, using the classical estimator as given, for example, in Cressie (1993, p. 75). We use two modifications: (1) we use stream distance rather than Euclidean distance, and (2) we separate those pairs of locations that are flow-connected from those that are not. Figure 5 shows that there is a strong difference in the empirical semivariogram between flow-connected locations and those that are not flow-connected. Both sets of points increase with increasing distance, indicating a possible trend, but locations that are flow-connected are lower overall than those that are not flow-connected. This is consistent with the idea that those locations that are flow-connected are more autocorrelated. Note that there is no weighting for flow volume in the variogram, so it would not be appropriate to use this for estimating covariance parameters, as is often done in classical geostatistics.

For weights on each stream segment, we used a digital elevation map (DEM) and National Hydrography Dataset (NHD) to compute the hydrologic basin for each stream segment, and then used a GIS to calculate the area of that basin. For each fork in the network, each segment was weighted in proportion to the area of its stream basin, where the weights summed to one. A GIS was also used to compute the stream distances. In addition to the 23 data locations, we created 433 prediction locations evenly spaced throughout the stream network. We included one covariate in the analysis; the distance upstream from the lowest point.

First, we use the exponential model that incorporates flow. We used REML to estimate the three covariance parameters contained in $\mathbf{\Sigma = A \odot V}$, which were $\widehat{\theta_0} = 0.387$, $\widehat{\theta_1} = 3.55$, and $\widehat{\theta_2} = 2217$, where the range is in kilometers. We then used the fitted covariance matrix to estimate the fixed effects. This is termed "empirical" best linear unbiased estimation (EBLUE), and is often used in software such as SAS (Littell et al. 1996). A table of fixed effects estimates and other relevant information, similar to what is produced in linear model software, is given in Table 2. There is evidence of a decreasing trend in SO_4 with distance upstream.

Next, we fit both constant mean models and models with the distance upstream covariate with all of the covariance models in Table 1, and include the unweighted exponential model based purely on stream distance. In Table 3, we compare models

Table 2 Fixed effects table for example data

Effect	Estimate	Standard error	Degrees of freedom	*t* value	Prob *t*
Intercept	9.07	0.984	21	9.22	< 0.0001
Distance upstream	−0.1337	0.0537	21	−2.488	0.0213

Table 3 Criteria for comparing models for example data

Model	AIC[a]	BIC[b]	RMSPE[c]
Constant mean			
Linear-with-sill	90.270	94.812	1.599
Spherical	90.276	94.818	1.600
Mariah	90.386	94.928	1.612
Exponential	90.324	94.866	1.606
Exponential[d]	95.409	99.951	1.652
With distance-upstream covariate			
Linear-with-sill	87.354	93.031	1.446
Spherical	87.354	93.031	1.446
Mariah	87.354	93.031	1.446
Exponential	87.354	93.031	1.446
Exponential[d]	95.684	101.362	1.642

[a] Akaike Information Criteria
[b] Bayesian Information Criteria
[c] Root-Mean-Squared-Prediction-Errors
[d] Model based on pure distance, without flow weightings

based on AIC (An Information Criteria, Akaike 1973), BIC (Bayesian Information Criteria, Schwarz 1978) and root-mean-squared-prediction errors (RMSPE) from cross-validation (for a description of cross-validation, see Cressie 1993, p. 101). Because we are comparing models with different fixed effects, we used maximum likelihood when computing AIC and BIC, but used restricted maximum likelihood for RMSPE. Table 3 indicates that there is little difference among the models that use flow. Also notice that for AIC, BIC, and RMSPE, all of the models using flow volume weights are much better than the exponential model based purely on stream distance. As in Table 2, AIC, BIC, and RMSPE indicate that distance upstream is a useful covariate. Also notice from Table 3 that the models with the distance upstream covariate have exactly the same AIC, BIC, and RMSPE values for all flow-based covariances. This occurs when $\mathbf{\Sigma} = \mathbf{A} \odot \mathbf{V} = \mathbf{A} \odot (\theta_0 \mathbf{I} + \theta_1 \mathbf{1}\mathbf{1}')$; i.e., all autocorrelation models have the ability to have nearly pure autocorrelation of one among all locations. In this case, most of the spatial structure in the covariance matrix is captured by the flow weightings in $\mathbf{A}$.

Now consider models with the distance upstream covariate; it is interesting to compare parameters for the exponential covariance model, both with and without the flow volume weightings. The partial sill and nugget effect for the flow model are 3.546 and 0.387, respectively, while for the pure distance case they are 1.996 and 1.217. Clearly, from Fig. 2, there is heterogeneity near the confluence of streams, and in general, this is apparent from Fig. 5. The pure distance model accounts for this with a larger nugget effect, whereas the flow model handles the heterogeneity through the weights.

In order to make predictions, we will include the distance upstream covariate and use the fitted covariance matrix using REML for the exponential model using flow; although, recall from Table 3, it really does not matter which autocorrelation model is used. Once we have a valid covariance matrix, the usual universal kriging equations can be used (see, e.g., Cressie 1993, p. 123). The predictions are shown in Fig. 6. Notice in Fig. 6 that we see some of the usual properties from kriging. The prediction standard errors are smallest near the data values, and predictions change gradually within stream segments. We can see a general trend upstream with decreasing values. However, there are several interesting and unusual features in Fig. 6. First, look at the predictions on the stream segment labeled A (a close-up is provided with an inset). Notice that the predictions are lower than the two nearest values on the main stream, and lower than all of the predictions on the main stream segment, even though no data occurred on the segment labeled A. The model recognizes the fact that, relative to the mouth of stream segment A, the downstream value is lower than the upstream

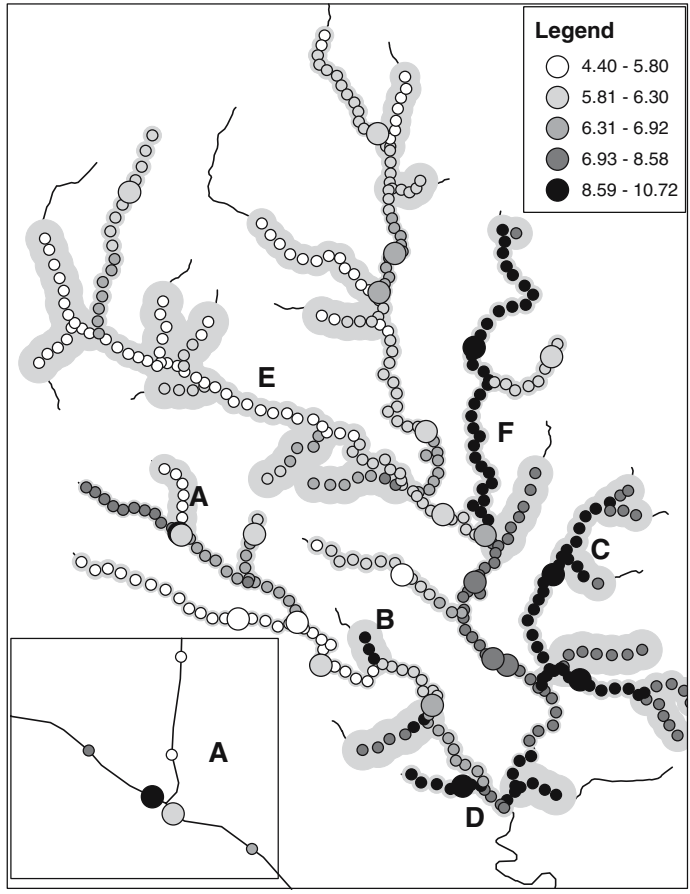

Fig. 6 Predictions for the example data in Fig. 2. The Observed locations are shown with large circles and predicted locations are shown with smaller circles; both are shaded according to their observed or predicted values. The width of the gray shading behind the circles is proportional to the prediction standard errors. Thus, areas with wider shading have less precision

value. We suppose that it is likely that stream segment A has added water with low values. In fact, the stream flow from segment A is rather small, so the predictions are quite low. The prediction standard errors are large on stream segment A, but it is interesting to see that the predictions are able to use the changing values in the stream segment into which A flows to adjust the predicted values. The same thing happens to the stream segment labeled with B, except here the predicted values are greater than the main stream because the observed values increase when going from the upstream location to the downstream location on the main stream.

For another interesting property, notice that the prediction standard errors get very large at the fork labeled C, whereas they are smaller at the fork labeled D. The observed location at C is just below the fork, whereas for D it is just above the fork. When the observed location is just below a fork, as in C, there is no way of knowing whether that value is the mixture of streams with high and low values, or any combination of mixing. This lack of knowledge is reflected in the prediction standard errors. Also, there is more flow volume coming from the northern fork at C, so the prediction standard errors are a bit lower here, reflecting the fact that it would probably have more influence over the observed value. These properties have obvious and important implications for sampling designs on stream networks.

Finally, in Fig. 7 we simulate data near a stream fork. In general, kriging predictions are "smooth," changing gradually between observed locations. However, notice that because of the construction using asymmetric moving averages with tails that only run upstream, there are prediction discontinuities at stream junctions in Fig. 7. This is an important property for predicting stream chemistry values and fits our understanding of how flow affects observations and predictions.

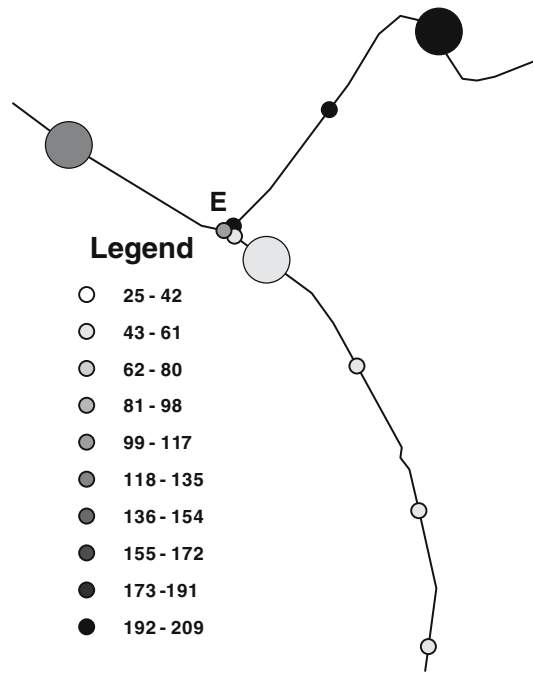

Fig. 7 A close-up of simulated predictions. Observed locations are shown with large circles and predicted locations are shown with smaller circles; both are shaded according to their observed or predicted values. Notice the discontinuity of predictions at the stream fork

In the prediction of a resource, or the monitoring of pollution, we often want to predict the total or average amount along a stream segment, in addition to making a map of point predictions. When using covariates, this is also known as universal block kriging (see, e.g., Cressie 1993, p. 151). Because we have a valid, fitted covariance matrix, we can easily perform universal block kriging on stream segments. From Fig. 6, we block krige the stream segment E and the segment labeled with F. The predicted average for segment E is 5.51 with a prediction standard error of 1.07, while the predicted average for segment F is 9.88, with a prediction standard error of 0.62.

4 Discussion and conclusions

We were able to meet the objectives stated in the introduction. We developed valid spatial covariance models that use flow and stream distance. For the example data set, the models using flow were much more appropriate than the pure stream distance model. The valid covariance matrix based on flow and stream distance allowed us to fit a spatial linear model to fixed effects and evaluate their importance. We also used the spatial linear model for kriging and block kriging predictions.

Many areas need further research in developing spatial models for stream networks. Exploratory graphics and diagnostic methods are important tools in selecting and evaluating models. Because of the weighting for flow models, the usual methods may not be appropriate. For example, the usual empirical variogram on residuals needs modification to account for weighting. As mentioned earlier, whole classes of models are yet to be developed for moving averages that have their tails running downstream rather than upstream, or in both directions. More functions can be developed for flow models as well. We used REML for covariance estimation, but Bayesian and other estimation methods need to be developed. The use of moving averages could allow for the use of nonstationary models, where we would allow the variance to change with flow. Space precludes additional development here, but these research areas are currently being investigated.

Acknowledgements Financial support for this work was provided by Federal Aid in Wildlife Restoration to the Alaska Department of Fish and Game, and the EPA STAR grant CR-82909501-0. Thanks to Noel Cressie, Jesse Frey, Andrew Merton and an anonymous reviewer for comments, to Ron Barry for many years of collaboration on moving average models and parts of Table 1, and to Scott Urquhart, Jennifer Hoeting, and Hariharan Iyer who invited this talk for the Graybill Conference in Ft. Collins in June, 2004.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp 267–281

Barry RP, Ver Hoef JM (1996) Blackbox kriging: spatial prediction without specifying the variogram. J Agricult Biol Environ Stat 1:297–322

Chiles J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. John Wiley and Sons, New York, p 695

Cressie N (1993) Statistics for spatial data, revised edition. John Wiley and Sons, New York, p 900

Cressie N, Majure JJ (1997) Spatio-temporal statistical modeling of livestock waste in streams. J Agricult Biol Environ Stat 2:24–47

Cressie N, Pavlicova M (2002) Calibrated spatial moving average simulations. Stat Model 2:267–279

Curriero F (1996) The use of non-euclidean distance in geostatistics. Ph.D. Thesis, Kansas State University
Dent CL, Grimm NB (1999) Spatial heterogeneity of stream water nutrient concentrations over successional time. Ecology 80:2283–2298
Fuentes M (2002) Spectral methods for nonstationary spatial processes. Biometrika 89:197–210
Gardner B, Sullivan PJ, Lembo Jr, AJ (2003) Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. Can J Fish Aquat Sci 60:344–351
Higdon D (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environ Ecol Stat 5:173–190
Higdon D, Swall J, Kern J (1999) Non-stationary spatial modeling. In Bayesian statistics 6, Oxford Univ Press, Oxford, 761–768
Littell RC, Milliken RC, Stroup WW, Wolfinger R (1996) SAS system for mixed models. SAS publishing, Cary, NC, p 656
Little LS, Edwards D, Porter DE (1997) Kriging in estuaries: as the crow flies, or as the fish swims? J Exp Mar Biol Ecol 213:1–11
Rathbun SL (1998) Spatial modeling in irregularly shaped regions: kriging estuaries. Environmetrics 9:109–129
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Smith RA, Schwarz GE, Alexander RB (1997) Regional interpretation of water-quality monitoring data. Water Resour Res 33:2781–2798
Torgersen CE, Gresswell RE, Bateman DS (2004) Pattern detection in stream networks: quantifying spatial variability in fish distribution. In: Nishida, T. (ed), Proceedings of the Second Annual International Symposium on GIS/Spatial Analyses in Fishery and Aquatic Sciences. Fishery GIS Research Group, Saitama, Japan
Ver Hoef JM, Barry RP (1998) Constructing and fitting models for cokriging and multivariable spatial prediction. J Stat Planning Inference 69:273–294
Ver Hoef JM, Cressie N, Barry RP (2004) Flexible spatial models for kriging and cokriging using moving averages and the fast Fourier transform (FFT). J Comput Graphical Stat 13:265–282
Yuan LL (2004) Using spatial interpolation to estimate stressor levels in unsampled streams. Environ Monit Assess 94:23–38

Biographical Sketches

Jay Ver Hoef completed most of this work as a biometrician for the Wildlife Conservation Division of the Alaska Department of Fish and Game. He is currently a statistician with the National Marine Mammal Laboratory in Seattle, where he serves as a statistical consultant and continues his statistical research interests. He is also an adjunct professor of statistics with the Mathematics Department of the University of Alaska, Fairbanks, and a fellow of the American Statistical Association. He received his B.S. in botany from Colorado State University, Fort Collins, his M.S. in botany from the University of Alaska, Fairbanks, and his Ph.D., a co-major in statistics and EEB (ecology and evolutionary biology), from Iowa State University, Ames, Iowa.

Erin Peterson is a Ph.D. candidate in the Department of Geosciences at Colorado State University, Fort Collins; she expects to receive her Ph.D. in earth resources in 2005. She is interested in developing GIS and spatial statistical methodologies that can be used to improve the efficiency of regional aquatic monitoring and assessment. She received her B.S. in conservation forestry from Michigan State University, East Lansing, and her M.S. in forestry from Colorado State University, Fort Collins. She has accepted a post-doctoral position at CSIRO-Australia (Commonwealth Scientific and Industrial Research Organisation) in the Mathematical and Information Sciences Division, where she will continue to pursue her research interests.

David Theobald is a scientist at the Natural Resource Ecology Laboratory and Department of Recreation and Tourism at Colorado State University. He is interested in assessing conservation threats, especially in the Rocky Mountain west. He received his B.A. in geography from the University of Colorado, Boulder, his M.A. in geography from the University of California, Santa Barbara, and his Ph.D. in geography from the University of Colorado, Boulder. He has recently written a book, GIS Concepts and ArcGIS Methods, and has contributed to The Atlas of the New West.