Reproduction of Applications of Spatial Statistical Network Models to Stream Data

Course Project for **MTH 643A**

Submitted by: Aindrila Garai (221258)
Suchandra Sasmal (221439)
Ghazar Muzaffar (22103267)

Supervisor: Dr. Arnab Hazra

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Date of Submission:
November 14, 2023

Academic Year 2023–24

Abstract

Contributor: Ghazar

Streams and rivers are crucial ecosystems, supporting a substantial portion of Earth's biodiversity and providing vital ecosystem services for human populations. Accurately assessing the status and trends of these aquatic resources is essential for effective conservation and management. However, conventional statistical methods, originally designed for terrestrial applications, often prove inadequate when applied to the complex and dynamic nature of stream environments.

In response to this challenge, a novel class of spatial statistical models has emerged, specifically tailored for stream networks. These models, characterized by robust covariance structures, can accommodate diverse types of stream data, including water quality attributes, habitat conditions, and biological surveys. Employing appropriate probability distributions such as Gaussian, binomial, and Poisson, these models effectively address spatial autocorrelation, acknowledging the non-independence of measurements in clustered locations. Consequently, they contribute to a more nuanced understanding of stream datasets and enhance predictive capabilities.

The SSN package for R is noteworthy in this context, offering a suite of functions designed explicitly for modeling stream network data. This package seamlessly integrates geographic information systems data and introduces the Spatial Stream Network object class using spatial sp classes. Within the SSN package, functions facilitate the fitting of spatial linear models (SLMs) for the Spatial Stream Network object. These models incorporate unique covariance matrices that account for both distances and the topological configuration of stream networks. The versatility of the SSN package lies in its inclusion of various models, ranging from traditional Euclidean distance models to geostatistical models that consider the volume and direction of flowing water. The package encompasses Poisson and binomial families within a generalized linear mixed model framework, providing flexibility to address a spectrum of research questions. Additionally, the SSN package supports essential tasks such as plotting, diagnostics, and prediction. This includes the application of kriging methods for handling missing data or unobserved locations, as well as block kriging over sets of stream segments.

Contents

1 INTRODUCTION

Contributor: Ghazar

Rivers and streams are home to aquatic life forms and provide sustenance to mankind through food, transportation, recreation, and much more making them vital parts of Earth's ecosystems. However, the effects of climate change and continued human development threaten the management and protection of these essential waterways. We need trustworthy scientific data to conserve and manage these resources. Due to the increasing demand for such information, collecting new data through conventional methods is often cumbersome and requires a lot of expenditure in terms of manpower and monetary aid.

To tackle this challenge, scholars are progressively investigating novel methodologies to optimize the efficacy of current databases. The utilization of this approach has the potential to be both economically viable and productive, especially when collaborating with advancements such as remote sensing, bioinformatics, computation, and data storage, which have substantially augmented the quantity of stream data accessible in recent times. With time as more and more closely spaced observations are being taken, certain challenges arise from this increasing data density, it violates the underlying assumption of independence among observations, which is a fundamental of classical statistics. When this assumption is not satisfied, it can result in inaccurate estimation of parameters and unreliable statistical inferences. In the past, classical statistical analyses were formulated to offer probabilistic inferences regarding population parameters, such as means, totals, and proportions. They, however, did not take into account the spatial distribution of measurements.

To address the potential bias caused by spatial gradients, researchers developed stratified random sampling designs to handle the impact of spatial variation on these estimates. It was observed that spatial patterns and dependencies played a crucial role in organizing natural systems, thereby giving rise to the emergence of spatial statistics and spatial ecology, both of which have proven to be highly valuable in the study of terrestrial ecosystems. Streams exhibit notable distinctions from terrestrial ecosystems due to their formation of directed networks that facilitate the flow of energy, materials, and information through confined corridors within landscapes. Hence, it is important for statistical methods designed for stream data analysis and prediction to incorporate these characteristics to achieve optimal performance.

The field has witnessed significant advancements that have resulted in the emergence of spatial stream-network models (SSNMs) that integrate appropriate covariance structures adapted to stream networks. These models consider the branching structure, directed flow, longitudinal connectivity, and abrupt changes near tributary confluences that are fundamental to streams. Stream survey data can be effectively analyzed using Stream Survey Network Models (SSNMs), which offer diversity in their application to different types of data such as water chemistries, habitat conditions, and biological attributes. These models use statistical distributions such as Gaussian, binomial, and Poisson to accurately analyze and interpret the collected data. (SSNMs) are experiencing an increasing level of acceptance, and the development of cost-free software has been undertaken to streamline their integration with pre-existing databases.

We can utilize preexisting stream databases and improve the efficacy of decision-making processes, conservation initiatives, and resource management practices within stream ecosystems. The utilization of spatial models such as SSNMs in regions with well-established stream databases can provide valuable findings, enhance the accuracy of predictions, refine parameter estimates, and are inexpensive. In areas where stream data are scarce, the utilization of SSNMs (Spatial Stream Network Models) and their associated simulation techniques can provide valuable guidance in devising sampling strategies that are well-optimized for stream systems.

2 OBJECTIVE

Contributors: Aindrila, Suchandra

A novel spatial statistical model has been developed, focusing on robust covariance structures tailored for stream networks. This model demonstrates versatility by accommodating various types of stream data, such as water quality attributes, habitat conditions, and biological survey results. This adaptability is achieved through the application of suitable probability distributions, such as Gaussian, binomial, or Poisson, depending on the specific nature of the data under consideration. In this paper, we offer a concise overview of spatial autocorrelation, referring to the phenomenon of nonindependence among observations. We explore its impact on statistical inference and illustrate instances where the application of Spatial Stream Network Models (SSNMs) enhances inference for common datasets and research questions. Within the framework of the SSNM, a mixed-model autocovariance structure is being employed. This structure encompassed components such as exponential tail-up, exponential Euclidean, and linear-with-sill tail-down, as outlined in detail.

3 DATA

Contributors: Aindrila, Suchandra

The class of the core data structure in the SSN package is SpatialStreamNetwork which especially works with stream networks. It is based on the S4 object model. The S4 object is an improvement over the S3 object which contains the object name, column names, and the type of the data that will be stored in the columns. This adds safety to our code and prevents us from accidentally making naive mistakes. We get SpatialStreamNetwork by extending the SpatialLinesDataFrame class, with additional components. As SpatialStreamNetwork combines point and line features within a single S4 object, it is unique. The additional components stores spatial point data (SSNPoints), a matrix containing network coordinates of each line segment (network.line.coords), and a reference to the ".ssn directory".

The process of building spatial data is very lengthy and involves numerous steps for a stream network model. After the completion of preprocessing a “.ssn” directory is created and stores the data. The two shapefiles “edges”, “sites” are always in this directory that contains both the geometry and attribute information for the stream network itself and the locations at observed data points.

It has similarities with the SpatialPointsDataFrame class but is not a direct extension. The first thing to notice is that common slot names are prefixed with “point”. To tackle spatial point data for observation and prediction of locations a new class called SSNPoint is introduced in addition to the SpatialStreamNetwork class. Another key component of SSNPoint is the “network.point.coords” matrix, which contains network coordinates. To deal with multiple datasets of prediction sites an SSNPoints list is made such that it stores multiple SSNPoint objects, each with a unique identifier (ID) to distinguish them within the SpatialStreamNetwork object.

Here, we use real-life spatial data from the Middle Fork, a stream in Idaho, USA. It is available in the SSN package. This dataset is a subset of a larger dataset available here. The MiddleFork data folder contains the spatial, attribute, and topological information needed to construct a spatial stream network object using the SSN package. The folder contains five spatial datasets, namely: edges, sites, CapeHorn, Knapp and pred1km. In this project we will be taking “sites” as our observed dataset and “CapeHorn”, “Knapp” and “pred1km”. The following variables are used in the fitting of the model:

- Summer_mn: Overall summer mean termperature of the deployment.
- CUMDRAINAG: Cumulative drainage area (km2)
- AREAWTMAP: Area weighted mean annual precipitation (mm) at lowermost location on the line segment where the site resides
- SLOPE: Slope of the line segment (cm/cm) where the site resides
- ELEV_DEM: Elevation at the site based on a 30m DEM
- Deployment: Unique identifier of the site by year of deployment

The dataset contains two-stream networks, 45 observed data points, 220 prediction locations spaced 1 km apart, 1273 densely packed prediction locations on a stream called the Knapp, and an additional 654 densely packed prediction locations on a stream called CapeHorn.

4 BACKGROUND

4.1 SPATIAL AUTOCORRELATION

Contributors: Ghazar

Spatial autocorrelation is a fundamental concept in the analysis of stream data. It measures the tendency of nearby measurements to be correlated, where measurements are dependent upon the degree of spatial separation between them. Streams often exhibit a common occurrence of positive autocorrelation, as measurements taken near each other tend to display a higher degree of similarity and dependence.

Two primary reasons that contribute to the occurrence of positive autocorrelation in streams are,

1. **Local Habitat Similarities:** Positive autocorrelation arises from the presence of comparable physical and ecological conditions in neighboring stream habitats. For example, the presence of comparable soil type, water depth, and temperature in adjacent pools can lead to increased fish populations, demonstrating the significance of conserving these microhabitats to promote biodiversity.
2. **Turbulent Stream Flows:** Varying stream velocities and turbulence within streams give rise to discrete microenvironments. Species that have adapted to turbulent environments have the potential to flourish in such areas, resulting in the formation of clusters characterized by shared attributes. Conversely, negative autocorrelation implies that nearby measurements exhibit greater dissimilarity. This can occur, for instance, when territorial fish exclude others from their immediate vicinity, leading to negative spatial correlations at specific distances.

Spatial autocorrelation when considered and used in spatial stream-network models (SSNMs), contributes to enhanced predictive accuracy, deeper ecological understanding, and the development of focused conservation strategies for stream ecosystems and is important for the analysis in stream ecosystems.

One common way to describe and visualize spatial autocorrelation is through semivariance analysis. Semi-variance measures the average variation between measurement values separated by a particular distance. The semi-variance for a given distance *"lag h"* is calculated using the empirical estimator formula:

$$\gamma(h) = 0.5\frac{1}{N(h)} \sum_{||x_i - x_j \in c(h)||} [z(x_i) - z(x_j)]^2 , \tag{1}$$

Here, $\gamma(h)$ is the semivariance for distance *lag h*, $N(h)$ is the number of data pairs separated by the distance h, $||x_i - x_j||$ is the distance between locations x_i and x_j, $c(h)$ represents the interval around h, and $z(x_i)$ is the data value at location x_i.

A plot of semivariance values against distance is called a semivariogram. When there is no spatial autocorrelation in a dataset, the semivariogram shows no trend. In the presence of positive autocorrelation, semivariance values are small near the origin and increase with greater distances. If nonzero semivariances occur at very short distances, it is referred to as the "nugget effect", representing spatial variation at resolutions finer than the sampling grain and measurement errors. The semivariogram may eventually reach an asymptote known as the "sill", indicating the variance or dissimilarity in uncorrelated data. The distance where the sill is reached is called the "range", signifying how quickly spatial autocorrelation decays with distance.

When classical statistical techniques are applied to data with spatial autocorrelation, they assume that each measurement is independent and contains unique information. However, in the presence of spatial autocorrelation, measurements are not independent, and the database contains redundant information. Failing to account for this redundancy can lead to overly liberal tests of statistical significance, resulting in an increased likelihood of type I errors. Parameter estimates may also be biased because measurements are spatially clustered, leading to over- or under-representation of conditions in certain areas.

In-stream research, it is advisable to employ models that explicitly accommodate spatial autocorrelation. These models take into account the spatial relationships between measurements, allowing for more accurate and reliable statistical inferences. Ignoring spatial autocorrelation can lead to the loss of valuable information and a reduction in the accuracy and validity of statistical analyses in stream ecosystems.

4.2 A COVARIANCE MODEL

Contributors: Aindrila, Suchandra

A spatial statistical model is essentially an expansion of the fundamental linear model commonly which are commonly used. The linear model is given by the form $y = \beta X + \epsilon$, where the response variable y is an $n \times 1$ vector, with n representing the total number of observations. This model consists of the response variable and predictors and their relationship is given through the design matrix $\boldsymbol{X}$ and parameters β. It is considered nonspatial because it assumes that the random errors, ϵ, are independent, with a variance of $\sigma^2 I$, where I is the $n \times n$ identity matrix, and does not vary according to different locations and remains constant.

	Nonspatial Model Covariance			Spatial Model Covariance		
Site	I	II	III	I	II	III
I	σ_{11}	0	0	σ_{11}	σ_{12}	σ_{13}
II	0	σ_{22}	0	σ_{21}	σ_{22}	σ_{23}
III	0	0	σ_{33}	σ_{31}	σ_{32}	σ_{33}

Table 1: Example Covariance Matrices for Nonspatial Statistical Model and Spatial Statistical Model.

In a spatial statistical model, this independence assumption is withdrawn, allowing values to manifest correlation, so $var(\epsilon) = \Sigma 1$, which is not constant throughout the domain. The general expression of the covariance matrix Σ involves numerous parameters that can be quite challenging to analyze and estimate. To describe the spatial relationships among elements within Σ, the model estimates the nugget, sill, and range parameters through an autocovariance function. This approach facilitates the estimation of covariance between any two sets of locations while reducing the number of parameter estimates for Σ from $n(n+1)/2$ to just 3.

A single autocovariance function usually estimates three parameters are estimated: the nugget effect, the partial sill, and the range parameter.

- The nugget effect captures variability that occurs at a scale finer than the closest measurements, as well as measurement error Cressie et al. [2006].
- The partial sill represents the variance of the autocorrelated process without the nugget effect.
- The range describes how quickly autocorrelation decreases with distance Cressie [2015].

In a covariance mixture, a partial sill and range parameter are estimated for each model, as well as, an overall nugget effect, and these parameters determine the relative influence that each component will have on the mixture given in Ver Hoef and Peterson [2010]. So, there is no need to determine a priori which covariance models to include. The covariance mixture provides a flexible approach that can be used to capture complex and multi-scale spatial patterns often found in stream datasets.

In stream networks, the traditional Euclidean distance between 2 points cannot be used as a straight-line measurement in a flat, two-dimensional space, which doesn't take into consideration the network's topology or the direction of flow. It cannot capture the spatial correlation that exists in stream networks. Hence we would be using along channel distance to model covariance. Along-channel distance takes into consideration the network's topology and the natural flow of water within the channel. It accounts for the cumulative distance traveled by moving along the stream's flow direction, as opposed to a straight-line measurement (Euclidean distance) between two points. However, euclidean distance can be a more appropriate measure when spatial patterns in stream data are believed to be affected by factors linked to the terrestrial landscape and the atmosphere, such as climate, geology, or land cover. Hence, a model that combines both autocovariance functions and Euclidean distance can be used which can efficiently model and take into account all kinds of spatial dependence in stream networks.

4.3 MODELS

Contributors: Aindrila, Suchandra

Novel stream network models introduced Hoef et al. [2006] utilizing moving averages that focus on stream network characteristics to develop innovative modeling techniques. These models utilize stream distance metrics, which quantify the shortest distance along the stream network between two locations (such as those outlined by Dent and Grimm [1999]).

The spatial stream network model depends on the spatial correlation between sites. Two locations have a flow-connected relationship if water flows from an upstream location to a downstream location. In contrast, a flow-unconnected relationship exists when two locations share a common junction downstream, but are not flow-connected. The dichotomous nature of stream flows introduces a new set of models, created using moving average constructions rather than Euclidean distance. The moving average construction as described by Ver Hoef and Peterson [2010] is,

$$C_u(r_i, s_j|\boldsymbol{\theta}_u) = \begin{cases} \pi_{i,j} C_t(h|\boldsymbol{\theta}_u), & \text{if } r_i \text{ and } s_j \text{ are flow-connected.} \\ 0, & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected,} \end{cases} \tag{2}$$

where r_i and s_j denote two locations on a stream network, h is the stream distance between them and where $\pi_{i,j}$ are weights due to branching characteristics of the stream.

The moving average functions are,

Name	Moving Average Functions
Linear with Sill	$g(x\|\boldsymbol{\theta}_u) = \theta_1 \boldsymbol{I}(0 \leq x/\alpha_u \leq 1)$
Spherical	$g(x\|\boldsymbol{\theta}_u) = \theta_1(1 - x/\alpha_u)\boldsymbol{I}(0 \leq x/\alpha_u \leq 1)$
Exponential	$g(x\|\boldsymbol{\theta}_u) = \theta_1 e^{-x/\alpha_u}\boldsymbol{I}(0 \leq x)$
Mariah	$g(x\|\boldsymbol{\theta}_u) = \theta_1 \frac{1}{1+x/\alpha_u}\boldsymbol{I}(0 \leq x)$

Table 2: Moving Average Functions

In addition, a spatial stream-network model may be fit using a mixed-covariance structure, which is based on a combination of two or more autocovariance models.

4.3.1 TAIL-UP MODEL

A tail-up model is a moving average function when it starts at some location and is non-zero only upstream of that location. It is a very specific class of statistical models. Utilizing moving averages Tail-up models analyze the tails of time series. TU model points upstream resulting in spatial correlation are restricted to flow-connected locations.

To identify the extreme ends of the distribution Tail-up models are very necessary. It is specially used in trends, patterns, or anomalies in data. It helps to detect rare events or outliers in time series data. Let h denotes their separation distance via the stream network. For the tail-up model, putting the functions given in table 2 in the integration given by

$$C_t(h|\boldsymbol{\theta}_u) = \int_h^\infty g(x|\boldsymbol{\theta}_u) g(x - h|\boldsymbol{\theta}_u)\, dx, \tag{3}$$

gives the following forms for the tail-up model:

- Tail-up Linear-with-Sill Model,

$$C_t(h|\theta_u) = \sigma_u^2 \left(1 - \frac{h}{\alpha_u}\right) I \left(\frac{h}{\alpha_u} \leq 1\right),$$

- Tail-up Spherical Model,

$$C_t(h|\boldsymbol{\theta}_u) = \sigma_u^2 \left(1 - \frac{3}{2}\frac{h}{\alpha_u} + \frac{1}{2}\frac{h^3}{\alpha_u^3}\right) I \left(\frac{h}{\alpha_u} \leq 1\right),$$

- Tail-Up Exponential Model,

$$C_t(h|\boldsymbol{\theta}_u) = \sigma_u^2 \exp{(-\frac{3h}{\alpha_u})},$$

- Tail-up Mariah Model,

$$C_t(h|\boldsymbol{\theta}_u) = \begin{cases} \sigma_u^2 \left(\dfrac{log(1 + \dfrac{90h}{\alpha_u})}{\dfrac{90h}{\alpha_u}}\right), & \text{if } h > 0. \\ \sigma_u^2, & \text{if h=0.} \end{cases},$$

- Tail-up Epanechnikov Model,

$$C_t(h|\boldsymbol{\theta}_u) = \frac{\sigma_u^2(h - \alpha_u)^2 f_{eu}(h;\alpha_u)}{16\alpha_u^5} I \left(\frac{h}{\alpha_u} \leq 1\right),$$

  where $f_{eu}(h;\alpha_u) = 16\alpha_u^2 + 17\alpha_u^2 h - 2\alpha_u h^2 - h^3$, $I(.)$ is the indicator function, $\sigma_u^2 > 0$ is partial sill and $\alpha_u > 0$ is the range parameter and $\theta_u = (\sigma_u^2, \alpha_u)^T$. A distance is called the effective range at which autocorrelation reaches 0.05. For the exponential and Mariah models, when h is the range parameter autocorrelation is approximately 0.05.

At stream confluences, Spatial weights ($\pi_{i,j}$) are used in the TU model allowing more weight to be allocated to data on tributaries thought to have a stronger influence downstream. Spatial-weighting schemes for TU stream-network models are often based on catchment area but can be calculated based on any ecologically relevant variable that is available for every line segment in a stream dataset. The weight for the segment downstream of the junction is the sum of the two upstream weights when two segments converge at a junction. Thus an an additive function is created when the segment moves downstream. We denote the value of the additive function as $\Omega(x)$ for some point x on the stream network. For two flow-connected points where r_i is downstream from s_j, the weights are,

$$\pi(i,j) = \sqrt{\frac{\Omega(s_j)}{\Omega(r_i)}}$$

4.3.2 TAIL-DOWN MODEL

A tail-down model is a moving average function when it starts at some location and is non-zero only downstream of that location. The spatial correlation of the TD model is permitted between both flow-connected and flow-unconnected locations because the moving-average function for the TD models points in the downstream direction.

For tail-down models, we have two situations flow-connected and flow-unconnected. When two sites are flow-unconnected, let a denote the shorter of the two distances to the common downstream junction and b denote the longer of the two distances.

For the flow-connected sites, we have,

$$C_c(h|\boldsymbol{\theta}_u) = \int_{-\infty}^{-h} g(-x|\boldsymbol{\theta}) g(-x-h|\boldsymbol{\theta})\, dx \tag{4}$$

where $h = s_2 - r_1 > 0$, s_2 is upstream of r_1 stream, $g(-x|\boldsymbol{\theta})$ is a unilateral tail-down function with nonzero values only on the negative side of 0. For $b \geq a$,

$$C_n(a,b|\boldsymbol{\theta}_u) = \int_{-\infty}^{-b} g(-x|\boldsymbol{\theta}) g(-x-(b-a)|\boldsymbol{\theta})\, dx \tag{5}$$

where b indicates the distance of the other location to the same junction and for two flow-unconnected locations, we use a to indicate the distance from one location to the nearest junction downstream of which it shares flow with the other location.

If two sites are flow-connected, h denotes their separation distance via the stream network. By putting the function in Table 2 in (4) and (5), the function $C_t(h|\boldsymbol{\theta}_u)$ can take the following forms for the TD model,

- Tail-Down Linear-with-Sill Model, $b \geq a \geq 0$,

$$C_d(a,b,h|\boldsymbol{\theta}_d) = \begin{cases} \sigma_d^2 \left(1 - \frac{h}{\alpha_d}\right) I\left(\frac{h}{\alpha_u} \leq 1\right), & \text{if flow-connected.} \\ \sigma_d^2 \left(1 - \frac{b}{\alpha_d}\right) I\left(\frac{b}{\alpha_u} \leq 1\right), & \text{if flow-unconnected.} \end{cases}$$

- Tail-Down Spherical Model, $b \geq a \geq 0$,

$$C_d(a,b,h|\boldsymbol{\theta}_d) = \begin{cases} \sigma_d^2 \left(1 - \frac{3}{2}\frac{h}{\alpha_d} + \frac{1}{2}\frac{h^3}{\alpha_d^3}\right) I\left(\frac{h}{\alpha_d} \leq 1\right), & \text{if flow-connected.} \\ \sigma_d^2 \left(1 - \frac{3}{2}\frac{a}{\alpha_d} + \frac{1}{2}\frac{b}{\alpha_d^3}\right) \left(1 - \frac{b}{\alpha_d}\right)^2 I\left(\frac{h}{\alpha_d} \leq 1\right), & \text{if flow-unconnected.} \end{cases}$$

- Tail-down Exponential Model,

$$C_d(a,b,h|\boldsymbol{\theta}_d) = \begin{cases} \sigma_u^2 \exp^{-\frac{3h}{\alpha_u}}, & \text{if flow-connected.} \\ \sigma_u^2 \exp^{-\frac{3(a+b)}{\alpha_u}}, & \text{if flow-unconnected.} \end{cases}$$

- Tail-down Mariah Model,

$$C_d(a,b,h|\boldsymbol{\theta}_d) = \begin{cases} \sigma_d^2 \left(\dfrac{log(1+90h/\alpha_d)}{90h/\alpha_d}\right), & \text{if flow-connected,h} > 0. \\ \sigma_d^2, & \text{if flow-connected,h=0.} \\ \sigma_d^2 \left(\dfrac{log(1+90a/\alpha_d) - log(1+90b/\alpha_d)}{90h/\alpha_d}\right) & \text{if flow-unconnected,h} \neq 0. \\ \sigma_d^2 \left(\dfrac{1}{90a/\alpha_d + 1}\right) & \text{if flow-unconnected,h} = 0. \end{cases}$$

- Tail-down Epanechnikov Model, $b \geq a \geq 0$,

$$C_d(a,b,h|\boldsymbol{\theta}_d) = \begin{cases} \dfrac{\sigma_u^2 (h-\alpha_u)^2 f_{eu}(h;\alpha_u)}{16\alpha_u^5} I\left(\dfrac{h}{\alpha_u} \leq 1\right), & \text{if flow-connected.} \\ \dfrac{\sigma_d^2 (h-\alpha_d)^2 f_{eu}(a,b;\alpha_u)}{16\alpha_d^5} I\left(\dfrac{b}{\alpha_d} \leq 1\right), & \text{if flow-unconnected} \end{cases},$$

where $f_{eu} = 16\alpha_d^3 + 17\alpha_d^2 b - 15\alpha_d^2 a - 20\alpha_d a^2 - 2\alpha_d b^2 + 10\alpha_d ab + 5ab^2 - b^3 - 10ba^2, \sigma_d^2 > 0$ and $\alpha_d > 0$ and $\theta_d = (\sigma_d^2, \alpha_d)^T$.

The weights for the tail-down model are similar to the tail-up model. As the model depends on the point farthest from the junction; i.e., b, and so a is the shorter of the two distances, a does not appear in the tail-down linear-with-sill model, but is used indirectly.

4.3.3 SPATIAL LINEAR MIXED MODEL

Let X be a design matrix of fixed effects with parameters β with Y as the response variable. Then the general spatial linear model in the SSN package used,

$$Y = X\beta + z_u + z_d + z_e + W_1\gamma_1 + \ldots + W_p\gamma_p + \epsilon,$$

where the z_u vector is spatially autocorrelated random variables that use a tail-up autocovariance and z_d contains spatially-autocorrelated random variables with a tail-down autocovariance. Here var$(z_u) = \sigma_u^2 R(\alpha_u)$ and $R(\alpha_u)$ is a correlation matrix. It depends on the range parameter α_u, var$(z_d) = \sigma_d^2 R(\alpha_d)$, z_e contains spatially-autocorrelated random variables with a Euclidean distance autocovariance, var(z_e) $= \sigma_e^2 R(\alpha_e)$, W_k is a design matrix for random effects γ_k; $k = 1, \ldots, p$ with var(γ_k) $= \sigma_k^2 \mathbf{I}$, and ϵ contains independent random variables with var$(\epsilon) = \sigma_0^2 \mathbf{I}$. When used for spatial prediction, this model is referred to as "universal" kriging with "ordinary" kriging is a special case where the design matrix X is a single column of one. The general covariance matrix used in the SSN package is,

$$\text{cov}(Y) = \sigma = \sigma_u^2\,\mathrm{R}(\alpha_u) + \sigma_d^2\,\mathrm{R}(\alpha_d) + \sigma_e^2\,\mathrm{R}(\alpha_e) + \sigma_1^2\,\mathrm{W}_1\,\mathrm{W}_1^T + \ldots + \sigma_p^2\,\mathrm{W}_p\,\mathrm{W}_p^T + \sigma_0^2\,\mathrm{I} \quad (6)$$

SSN Package uses MLE to estimate β and some subset of covariance parameters.

4.4 SEMIVARIOGRAM

Contributors: Aindrila, Suchandra

The semivariogram depicts the spatial autocorrelation of the measured sample points. Once each pair of locations is plotted, a model is fit through them. At a certain distance, the model levels out. The distance where the model first flattens out is known as the range. Sample locations separated by distances closer than the range are spatially autocorrelated, whereas locations farther apart than the range are not. The semivariogram and covariance functions quantify the assumption that things nearby tend to be more similar than things that are farther apart. Semivariogram and covariance both measure the strength of statistical correlation as a function of distance.

The value that the semivariogram model attains at the range (the value on the y-axis) is called the sill. The partial sill is the sill minus the nugget. Theoretically, at zero separation distance (lag = 0), the semivariogram value is 0. However, at an infinitesimally small separation distance, the semivariogram often exhibits a nugget effect, which is some value greater than 0. For example, if the semivariogram model intercepts the y-axis at 2, then the nugget is 2.

The nugget effect can be attributed to measurement errors or spatial sources of variation at distances smaller than the sampling interval or both. Measurement error occurs because of the error inherent in measuring devices. Natural phenomena can vary spatially over a range of scales. Variation at microscales smaller than the sampling distances will appear as part of the nugget effect. Before collecting data, it is important to gain some understanding of the scales of spatial variation.

The process of modeling semivariograms and covariance functions fits a semivariogram or covariance curve to your empirical data. The goal is to achieve the best fit, and also incorporate your knowledge of the phenomenon in the model. The model will then be used in your predictions. The semivariogram is defined as $(s_i, s_j) = 0.5var(Z(s_i) - Z(s_j))$,

If two locations, s_i and s_j, are close to each other in terms of the distance measure of $d(s_i, s_j)$, you expect them to be similar, so the difference in their values, $Z(s_i) - Z(s_j)$, will be small. As s_i and s_j get farther apart, they become less similar, so the difference in their values, $Z(s_i) - Z(s_j)$, will become larger. This can be seen in the following figure, which shows the anatomy of a typical semivariogram.

Notice that the variance of the difference increases with distance, so the semivariogram can be thought of as a dissimilarity function. The height that the semivariogram reaches when it levels off is called the sill. It is often composed of two parts: a discontinuity at the origin, called the nugget effect, and the partial sill; added together, these give the sill. The nugget effect can be further divided into measurement error and microscale variation. The nugget effect is simply the sum of measurement error and microscale variation and, since either component can be zero, the nugget effect can be composed wholly of one or the other. The distance at which the semivariogram levels off to the sill is called the range.

The covariance function is defined to be, $C(s_i, s_j) = Cov(Z(s_i), Z(s_j))$, Covariance is a scaled version of correlation. So, when two locations, s_i and s_j, are close to each other, you expect them to be similar, and their covariance (a correlation) will be large. As s_i and s_j get farther apart, they become less similar, and their covariance becomes zero. This can be seen in the following figure, which shows the anatomy of a typical covariance function. Relationship between semivariogram and covariance function. There is a relationship between the semivariogram and the covariance function: $(s_i, s_j) = \text{sill} - C(s_i, s_j)$,

4.5 TORGEGRAM

Contributors: Aindrila, Suchandra

It is clear that for tail-up and tail-down covariance models (and mixtures thereof) on stream networks, the correlation among responses may depend not only on total stream distance but also on flow connectedness, flow volume, and/or distances to a common junction. Thus, in contrast to Euclidean geostatistics, one empirical semivariogram is not adequate for characterizing spatial dependence on a stream network. Instead, no less than four distinct empirical semivariograms, each a modification of the Euclidean empirical semivariogram, may be necessary. We call this quadripartite collection of empirical semivariograms the Torgegram in honor of Christian Torgersen, a stream ecologist who was among the first to study spatial dependence on streams.

Traditionally, the semi-variogram is constructed using Euclidean distance between the observation sites. However, for our stream network application, it is better to characterize the semi-variance in terms of stream distance rather than Euclidean distance. In addition, because some variables are impacted by the flow connectivity structure of the stream network (e.g., there is no water transport between unconnected locations as well as in the upstream direction), it is important to separate the measured distances into two categories. First, distances between flow-connected sites (e.g., the water flows directly from the upstream site location to the downstream site location) are equal to the length of the stream segments connecting the two sites. Second, distances between flow-unconnected sites are equal to the sum of the lengths of the two-stream segments that connect each site to the nearest downstream connecting junction. Because of this distance categorization, we obtain two semi-variograms that when plotted simultaneously are called the Torgegram (Zimmerman and Ver Hoef 2017). If the spatial dependence of the variable of interest is not impacted by network connectivity, the flow-connected and flow-unconnected portions of the Torgegram should be similar; otherwise, they could be quite different.

The Torgegram is a stream network version of the empirical semivariogram:

- For unilateral models and mixes thereof, correlations may depend not only on stream distance but also on flow connectedness, flow volume, and distances to a common junction. Thus, for diagnostic purposes, one empirical semivariogram is not adequate.
- Instead, four are needed. We call this quadripartite collection the Torgegram, in honor of stream ecologist Christian Torgerson.

Four components of the Torgegram :

1. The flow-unconnected stream-distance (FUSD) semivariogram: where the partition the site pairs on segments into the stream distance bins.
2. If $Y(.)$ is pure tail-up, then Y^{FUSD} is unbiased for the portion of its semivariogram, which is flat.
3. If $Y(.)$ is not pure tail-up but has an exponential semivariogram, then Y^{FUSD} is unbiased for the portion.
4. Otherwise, this component has no clear interpretation.

5 PREDICTION AND ESTIMATION

Contributors: Ghazar

5.1 PARAMETER ESTIMATION

The basic goal of analysis in ANOVA, ANCOVA, and regression is to estimate the parameters through which relationships are described between one or more predictor variables and a response variable. Multiple linear regression is generally used to estimate a vector of parameters. It describes how much a response variable changes relative to one unit change in a predictor variable, keeping other predictor variables constant. An SSNM Regression is used when the variables describe qualitative features such as fish or macroinvertebrate abundance, habitat conditions, or water quality. SSNM regression is a multiple linear regression with a proper auto covariance function implemented on network data. So. the estimates provided by SSNM Regression account for spatial structure in residual errors using an autocovariance function and often provide improved parameter estimates.

5.2 PREDICTION AT UNSAMPLED LOCATIONS

Stream and river networks stretch over a long range of kilometers which increases the cost of direct measurements. we can predict a response variable at the unsampled locations by multiplying the model parameter estimates with the values of predictors. Semi-continuous maps showing the status of a stream attribute can be developed if predictions are made at points placed systematically throughout a network. SSNMs make predictions similarly but also use information from the autocovariance function to improve predictive accuracy near measurement sites. Those predictions are generated using the universal kriging equation which has two parts; a prediction based on the linear regression model and an adjustment for spatial autocorrelation. Thus, the model predictors set the mean process for an initial prediction, which is then adjusted based on any information that nearby measurements provide. Predictions may also be generated using ordinary kriging basely solely on values of the response variable, but this assumes that the average condition of the response variable remains constant across the study region.

6 SIMULATION STUDY

Contributors: Aindrila, Suchandra

To demonstrate the efficacy of using the new proposed model for stream network data, an extensive simulation is being conducted using Ver Hoef et al. [2014]. The first step is to create a spatial stream network object. Next, simulate an autocorrelated response variable for this object. These functions are used to test methods and compare different sampling methods.

Here, we discuss four types of design functions,

- poissonDesign(lambda): It considers a Poisson process with the rate of occurrence of points on a per-network basis. lambda is the rate os poisson process.
- hardCoreDesign(n, inhibitionregion): This design generates n randomly distributed points on each network and removes points until the remainder is at least the distance between every point and the next most downstream point.
- systematicDesign(spacing): It generates a set of regular points that is important for block design. It is a very complicated design.
- binomialDesign(n): binomial design specifies the number of points (n). It is generated from a uniform distribution across each network.

In our implementation for networks, every point lies at a constant distance from the next most downstream point. So, the points just upstream of a junction may be very close to each other. The points may not appear to be equally spaced visually. Here, we use binomial design for the observed data points and a systematic design for the predicted data points to generate an SSN object using the SSN package in RStudio. Then, we plot the observed and predicted points on a spatial map in Figure 1.

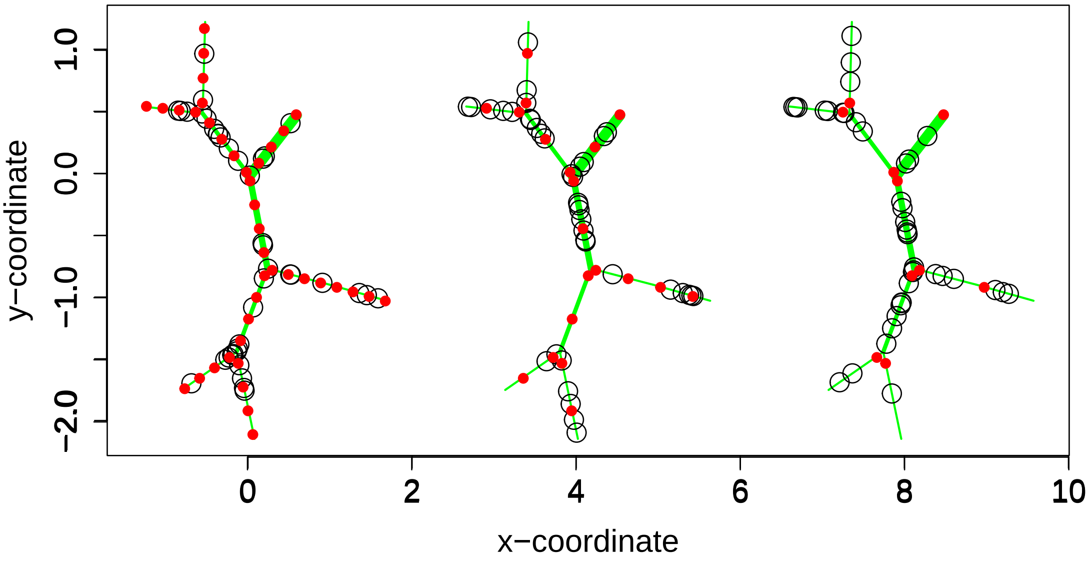

Figure 1: Spatial map of simulated observed and predicted points.

In the 1 green line represents the simulated three-stream network data. The black circle is the observed points that have been created using binomial design. Red dots are prediction points created by systematic design. The green lines get thinner in the up and downstream.

Then, the distance matrix between all the points is calculated. From the created SSN object. Our next task is to extract the data frames for the observed and predicted locations. Then continuous covariates, categorical covariates, and random effects for SSNM regression are created. Both the observed and prediction data frames have the same set of columns for any random or fixed effects.

Next, we use the SimulateOnSSN function to simulate data on the spatial stream network object created. The output of the above function is a list with three objects. Two of these objects are given in the following table.

Xnames	Coefficient
Intercept	10
X1	1
X2	0
F12	-2
F13	0
F14	2

Table 3: True Simulated Coefficients.

Covariance Model	Parameter	Estimate
Exponential Tail-Up	Parsill	3.0
Exponential Tail-Up	Range	10.0
Linear Sill Tail-Down	Parsill	2.0
Linear Sill Tail-Down	Range	10.0
Exponential Euclidean	Parsill	1.0
Exponential Euclidean	Range	5.0
RE1	Parsill	1.0
RE2	Parsill	0.5
Nugget	Parsill	0.1

Table 4: True Simulated Covariance Parameters.

The next task is to simulate response data from the created SSN object Figure 1 shows the spatial map of the simulated data from the SSN object.

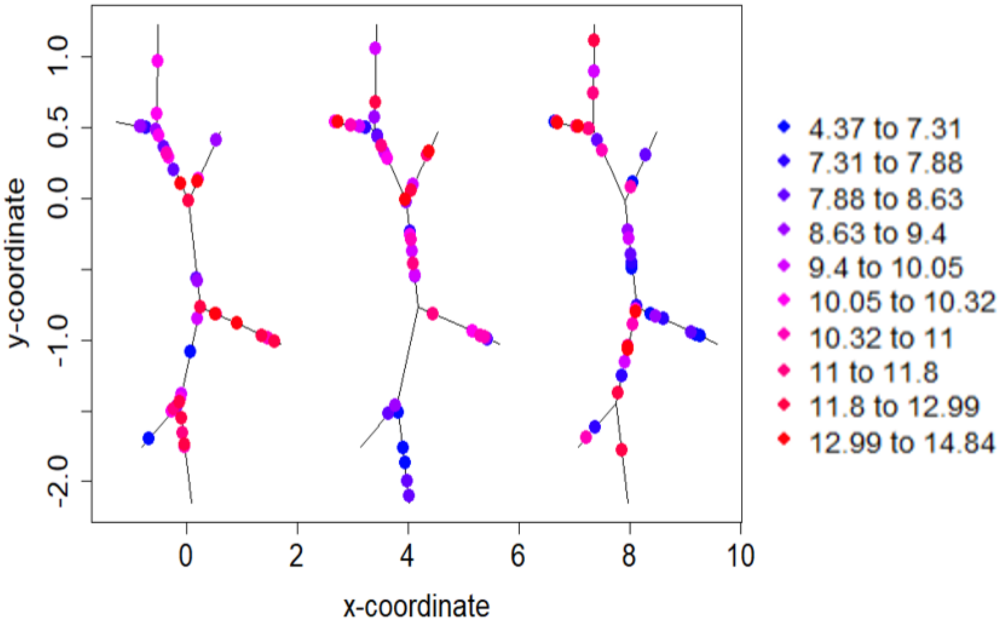

Figure 2: Values for the simulated response variable were simulated observed locations. The simulated data are colored according to their autocorrelated response.

Figure 2 shows the simulated response values at the locations. The simulated values are then stored in the data frames. Now, those values are to be extracted and then to this data, we fit a mixed-model autocovariance structure which consists of exponential tail-up, exponential Euclidean, and linear-with-sill tail-down components.

Parameter	Estimate	Standard Error	t value	p-value
Intercept	10.10931	0.57436	17.601	2×10^{-16}
X1	0.97884	0.05645	17.339	2×10^{-16}
X2	0.05756	0.05908	0.974	0.332
F11	0.00000	NA	NA	NA
F12	-1.59593	0.17191	-9.283	2×10^{-16}
F13	-0.04262	0.16355	-0.261	0.795
F14	2.07169	0.16563	12.508	2×10^{-16}

Table 5: Coefficient Estimates of SSNM Regression

We can observe that the fixed-effects estimates and the coefficients specified in the SimulateOnSSN function are close to each other. So, the model estimates the coefficients quite well.

Covariance Model	Parameter	Estimate
Exponential Tail-Up	Parsill	2.268
Exponential Tail-Up	Range	14.835
Linear Sill Tail-Down	Parsill	0.344
Linear Sill Tail-Down	Range	1.042
Exponential Euclidean	Parsill	0.264
Exponential Euclidean	Range	1.433
RE1	Parsill	0.263
RE2	Parsill	0.251
Nugget	Parsill	0.158

Table 6: Estimation of Covariance Parameters of SSNM Regression.

The covariance parameters are not estimated as well as coefficients. The reason behind poor estimation is the long distance from their simulated values. The actual covariance matrix can be close to the original covariance matrix, especially near the origin.

In the final stage, the predictions from the fitted model are compared to the actual simulated values.

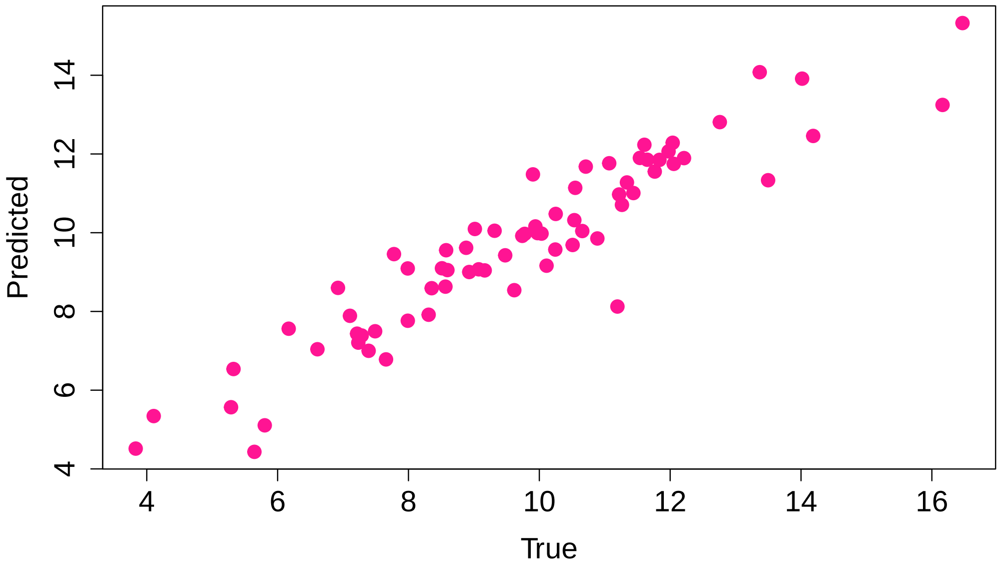

Figure 3: A comparison of true simulated data and predictions at locations after true values have been removed.

Figure 3 shows that the predicted values corresponding to the true value of the response almost falls on the $x = y$ line, validating the fact that the predicted values are very close to the true values. It shows excellent prediction accuracy for the model which includes both covariates and autocorrelation.

7 DATA ANALYSIS

Contributors: Aindrila, Suchandra

There are several variables in the Middle Fork data. The names of the variables in the “names(mf04p)” for each observed and prediction data set in mf04p.

7.1 EXPLORATORY ANALYSIS

We plot the observed and the predicted locations in the dataset in figure 4.

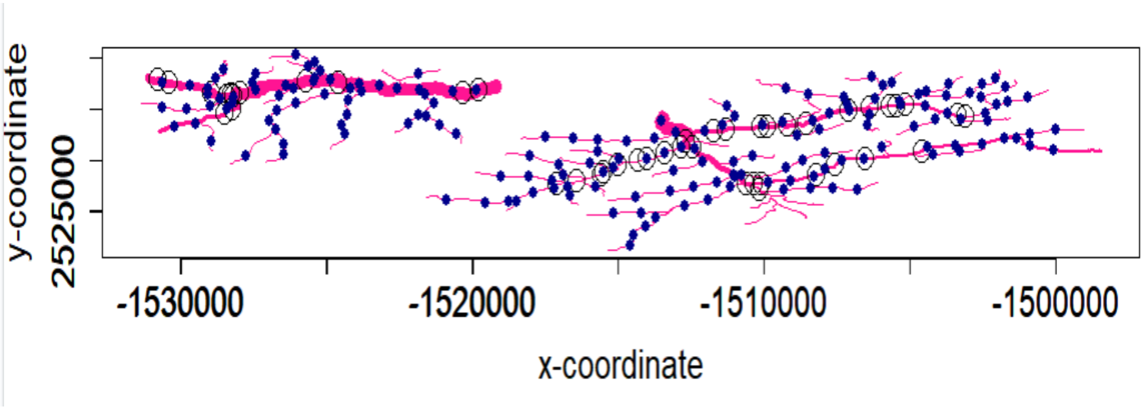

Figure 4: Spatial Map of observed and predicted locations. The pink lines represent the stream. The black points are the observed locations. The blue smaller points are the predicted locations.

Next, we plot the observed "Summer_mn" temperatures on the stream network in figure 5.

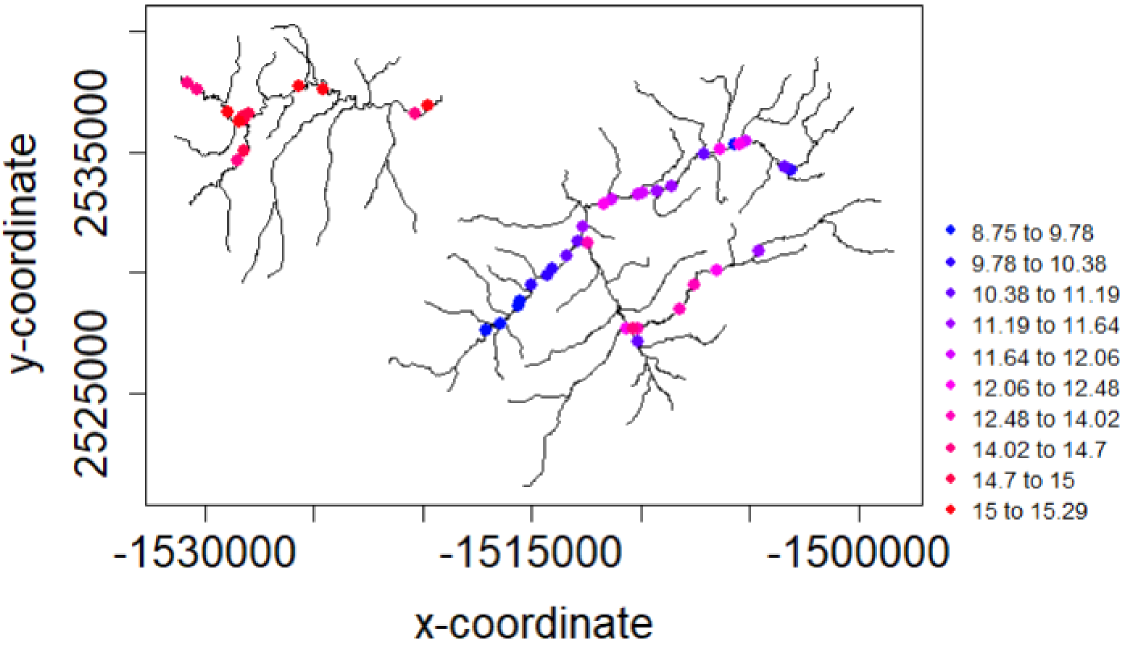

Figure 5: Observed "Summer_mn" temperatures in the "Middle Fork" data set

In the northwest corner of figure 5, the "Summer_mn" values of streams are high, whereas the rest of the values in the eastern part are relatively low.

The torgegram of the mean summer temperature for the Middle Fork data set is plotted in figure 6.

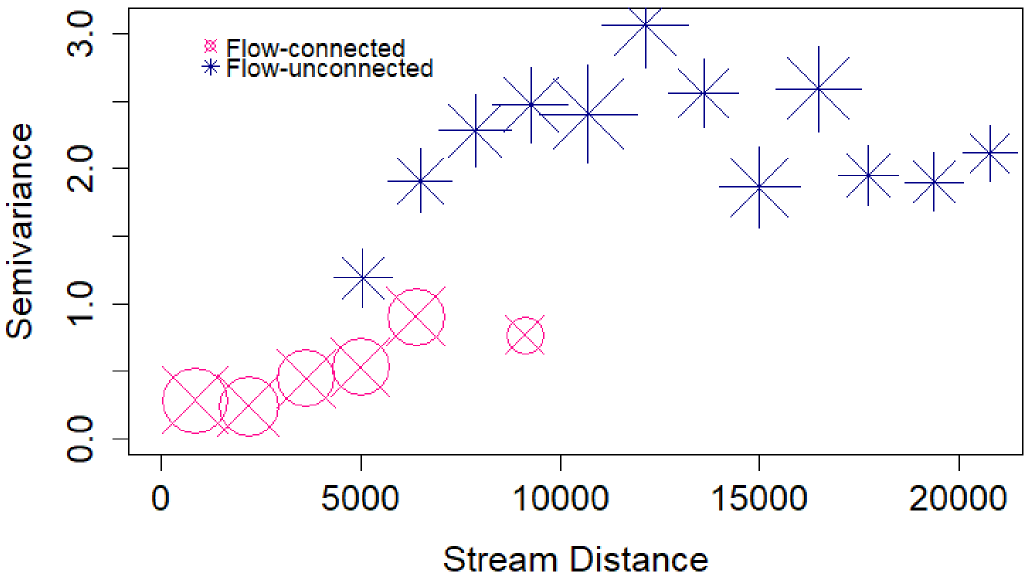

Figure 6: Torgegram of the mean summer temperature for the Middle Fork data set.

We can observe that the semi-variance is lower for the flow-connected stream networks than for the flow-unconnected streams. So, we can interpret that the correlation is low for the flow-unconnected streams, which is true b. The semi-variance for flow-unconnected streams is over three times higher than for flow-connected streams.

7.2 RESULTS

Our next task is to fit our proposed model to the "Summer_mn" temperature as the response variable. The results are shown in 7 and 8.

Parameter	Estimate	Standard Error	t value	p-value
Intercept	74.40356	13.29143	5.598	2×10^{-16}
ELEVDEM	-0.01408	0.010129	-1.390	0.1727
SLOPE	-1.65581	16.642689	-0.099	0.9213
CUMDRAINAG	-0.00178	0.006957	-0.256	0.7990
AREAWTMAP	-0.01018	0.004197	-2.427	0.0202
MAXELEVSMO	-0.01155	0.007932	-1.456	0.1539
ratio	-0.24231	0.164037	-1.477	0.1481
afvArea	-0.37740	2.769106	-0.136	0.8923

Table 7: Coefficient Estimates of SSNM Regression

Covariance Model	Parameter	Estimate
Exponential Tail-Up	Parsill	1.72144
Exponential Tail-Up	Range	117791.44082
Linear Sill Tail-Down	Parsill	0.09310
Linear Sill Tail-Down	Range	117786.55788
Exponential Euclidean	Parsill	0.01466
Exponential Euclidean	Range	111804.83124
Nugget	Parsill	0.00115

Table 8: Estimation of Covariance Parameters of SSNM Regression.

From the table 7 and 8, we can interpret standard errors of some variables are quite high. Also, most of the p-values are greater than the critical value. This implies that our model may not be a good fit for the dataset.

Figure 7 shows the predicted values for the variable "Summer_mn" using the tail-up model.

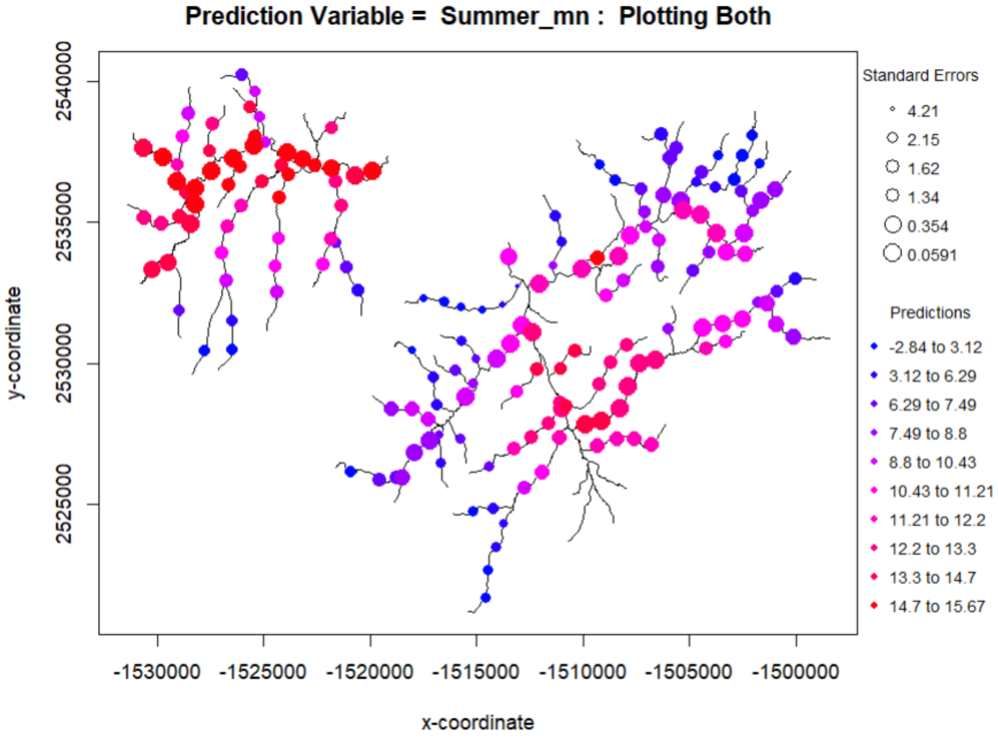

Figure 7: Prediction using Tail Up Model

Figure 8 shows the predicted values for the variable "Summer_mn" using the tail-down model.

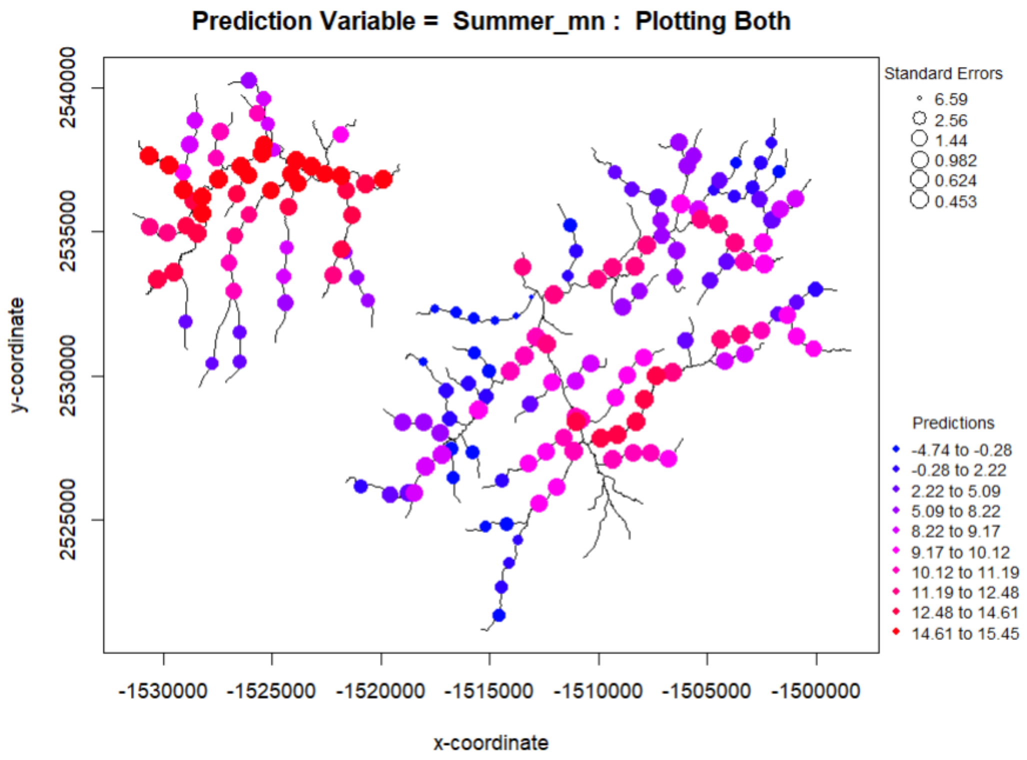

Figure 8: Prediction using Tail Down Model

Figure 8 shows the predicted values for the variable "Summer_mn" using a tail-exponential model.

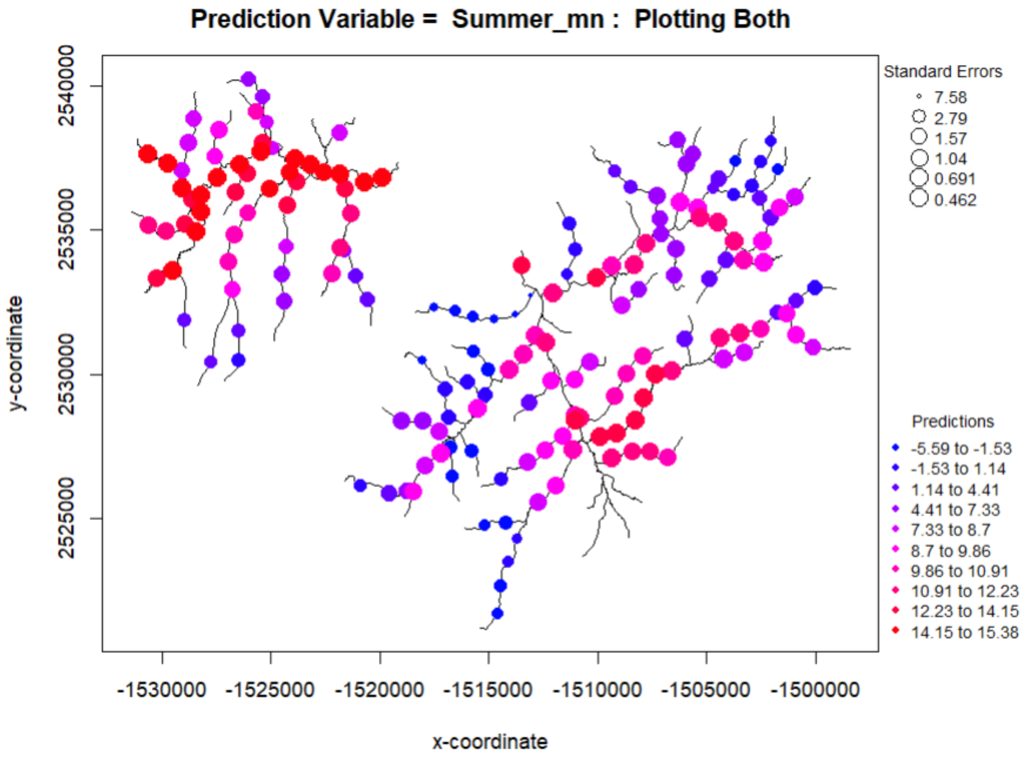

Figure 9: Prediction using Tail Exponential Model

For the tail down and tail exponential model, we can see that the standard errors for predicting "Summer_mn" temperatures are relatively low. The red dots represent higher values of the response while the blue dots represent the lower values. For all three models, the northwest corner of the plots predicts a high value of the response as in the observed data. In the middle of the eastern part, "Summer_mn" temperature is high and decreases gradually as we move towards the tail of the streams.

8 CONCLUSION

Contributors: Ghazar

Acknowledging the limitations inherent in relying solely on lumped models, the imperative of employing specialized analytical methodologies for streams, characterized by distinct spatial dimensions, comes to the forefront. compared to terrestrial systems, streams manifest a complex network structure therefore special stream network tools and methodologies are necessary for advanced analysis of the stream-related patterns and the conservation of streams.

A spatially dependent method offers the prospect of unraveling dynamic relationships within stream networks, thereby transcending the limitations inherent in traditional methodologies. Spatial techniques serve as a link to bridge existing gaps in the comprehension of streams, facilitating decisions and actions that are well-informed. The integration of innovative tools not only amplifies our capacity to comprehend the intricacies of stream dynamics but also empowers the formulation of targeted strategies conducive to their sustainable management and conservation. Such advancements, supported by empirical data, underscore a vision wherein streams are comprehensively understood, judiciously managed, and conserved for the mutual benefit of ecosystems and the communities reliant upon them. Indeed, the incorporation of spatial techniques in stream network analysis stands as a beacon guiding us toward a more robust understanding of these vital ecosystems, thereby enhancing our ability to make informed decisions and enact effective conservation measures.

References

Noel Cressie, Jesse Frey, Bronwyn Harch, and Mick Smith. Spatial prediction on a river network. *Journal of agricultural, biological, and environmental statistics*, 11: 127–150, 2006.

Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 2015.

Jay M Ver Hoef and Erin E Peterson. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, 105 (489):6–18, 2010.

Jay M Ver Hoef, Erin Peterson, and David Theobald. Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, 13:449–464, 2006.

C Lisa Dent and Nancy B Grimm. Spatial heterogeneity of stream water nutrient concentrations over successional time. *Ecology*, 80(7):2283–2298, 1999.

Jay Ver Hoef, Erin Peterson, David Clifford, and Rohan Shah. Ssn: An r package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, 56: 1–45, 2014.