Accounting for Euclidean and Hydrologic Distance in a Geostatistical Model of a Fluvial Fish Distribution

Gregory B. Anderson[1] and Jian Huang[1]

[1]Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

Introduction

- Geostatistical models account for spatial dependence in the random error of linear models, and usually provide more accurate predictions in species distribution models.
- Common geostatistical approaches based on Euclidean distance may not capture patterns of aquatic species produced by stream connectivity and the fluvial hierarchy.
- To guarantee ecological and statistical feasibility, Peterson and Ver Hoef[1] proposed an approach of combining Euclidean and fluvial covariance in geostatistical models.
- In this study, we utilized a mixed-model moving-average approach to predict the spatial distribution of a fish species (*Percina crassa*) in the upper Yadkin River
- We present how spatial pattern and autocorrelation vary under different covariance structures based on several proximity matrices (Euclidean and different hydrologic distances), and we compare a series of mixed models representing non-competing hypothesis regarding spatial patterns using leave-one-out cross-validation (LOOCV).

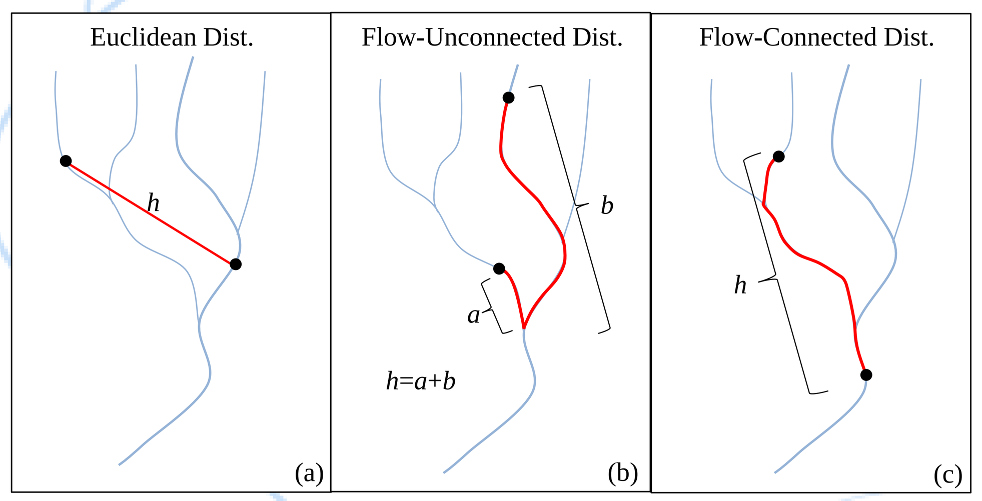

Figure 1. Euclidean (a), flow unconnected (b), flow connected distances (c) used for Euclidean, Tail-Down, and Tail-Up autocovariance models, respectively. Total hydrologic distance (*h*) can be partitioned into downstream distance (*a*) and upstream distance (*b*).

Methods

Data Sources

- Records for the upper Yadkin River system were retrieved from the NC Division of Water Quality website, consisting of 146 unique sampling localities.
- A stream network was created using elevation, flow accumulation, and flow direction rasters from the National Hydrograph Dataset Plus (Figure 2).

Model Description

- The geostatistical model was formulated as a generalized linear mixed model:

$$\Pr(\boldsymbol{y}=1) = \text{logit}^{-1}(\boldsymbol{X\beta} + \boldsymbol{Z}_E\boldsymbol{\alpha}_E + \boldsymbol{Z}_D\boldsymbol{\alpha}_D + \boldsymbol{Z}_U\boldsymbol{\alpha}_U + \boldsymbol{Z}_N\boldsymbol{\alpha}_N).$$

  where $\boldsymbol{X\beta}$ models the 1st-order variation, $\boldsymbol{\alpha}_E, \boldsymbol{\alpha}_D, \boldsymbol{\alpha}_U$ represent spatially autocorrelated random effects (SARE) based on Euclidean, total (Tail-Down) and flow connected hydrologic distance (Tail-Up), respectively, and $\boldsymbol{\alpha}_N$ represents additional variation not accounted for by the model.
- The covariance matrix of each random effect ($\boldsymbol{\Sigma}_\alpha$) can then be broken into an autocorrelation matrix (**R**) or the identity matrix (**I**) that is scaled by the individual variance component (${\sigma_\alpha}^2$), where ${\sigma_\alpha}^2$ represents the partial sill for each random effect, and ${\sigma_\alpha}^2{}_N$ represents the nugget effect.
- The scaled autocorrelation function ${\sigma_\alpha}^2\mathbf{R}$ can then be written as an auto-covariance function $C_\alpha(\boldsymbol{s}, \boldsymbol{s}'|\boldsymbol{\theta})$, and different covariance models can be applied.

Data Analysis

- Data analysis was performed using the R package SSN[2].
- Fixed effects (X, Y, watershed area [W] and W^2) were selected using AIC, and random effects were selected with LOOCV based on the area under the ROC curve (AUC).
- For Euclidean distance (isotropic and anisotropic): Exponential, Gaussian, and Spherical models were fit. For Tail-Up and Tail-Down models: Exponential, Linear-sill, and Spherical models were fit.

Hydrologic Network

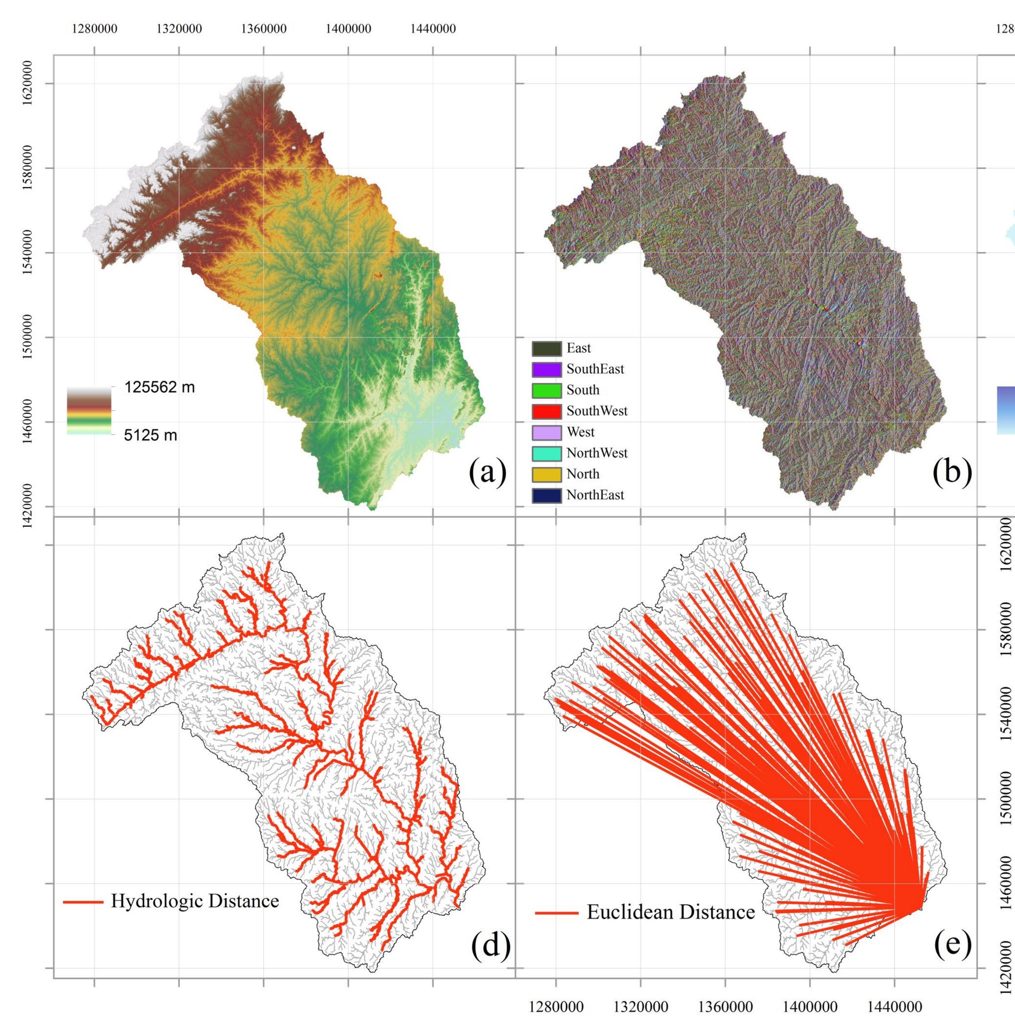

Figure 2. Derivation of hydrologic and Euclidean distance matrices. (a) elevation data from the national elevation dataset, (b) flow direction, (c) flow accumulation, (d) hydrologic distance to outlet, and (e) Euclidean distance to outlet.

Residual Autocorrelation

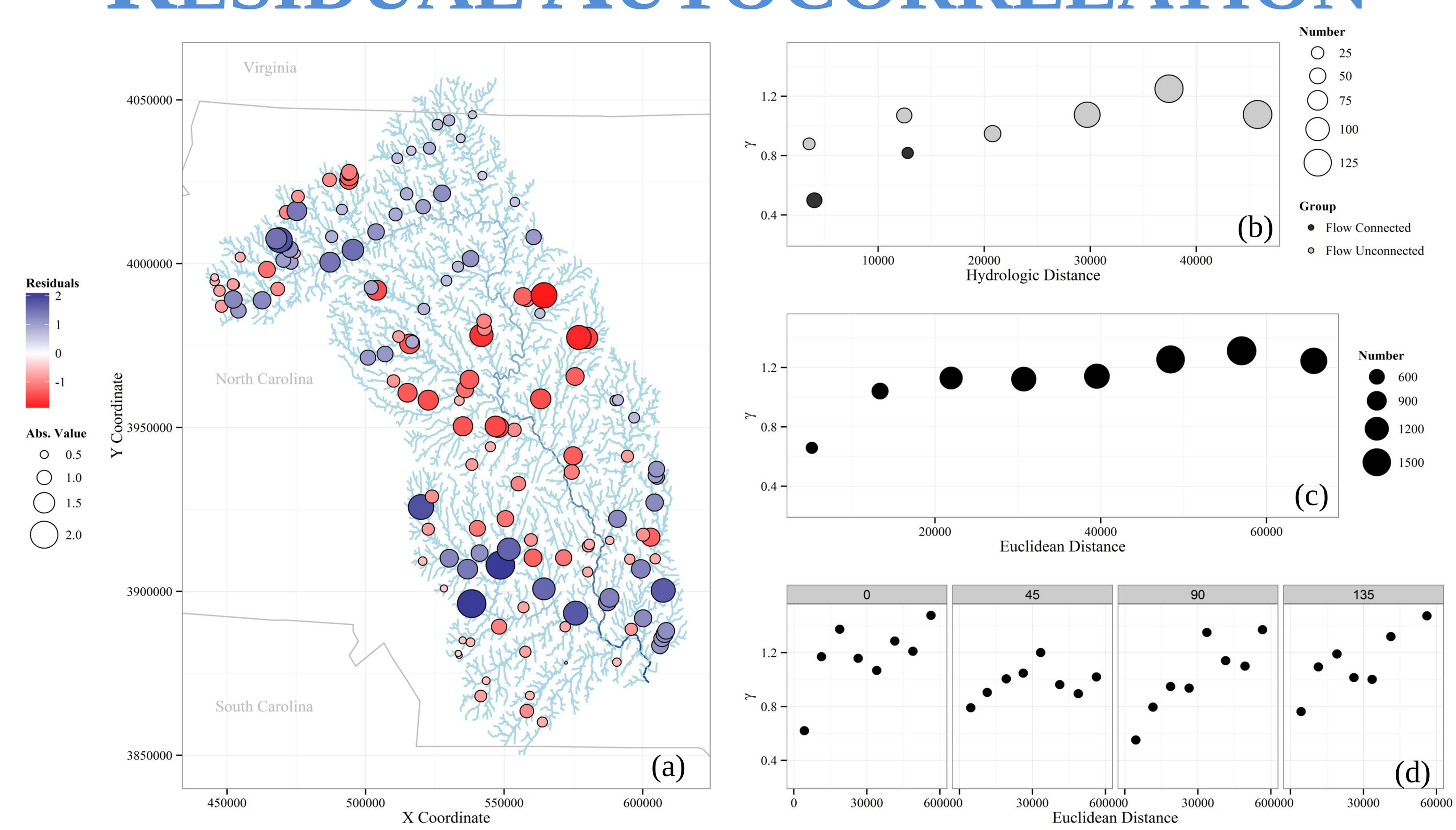

Figure 3. Residual autocorrelation from the GLM with no spatial effects. (a) map of residuals, (b) Torgegram of semivariance , (c) empirical semivariogram, and (d) empirical semivariogram by direction.

Spatial Prediction

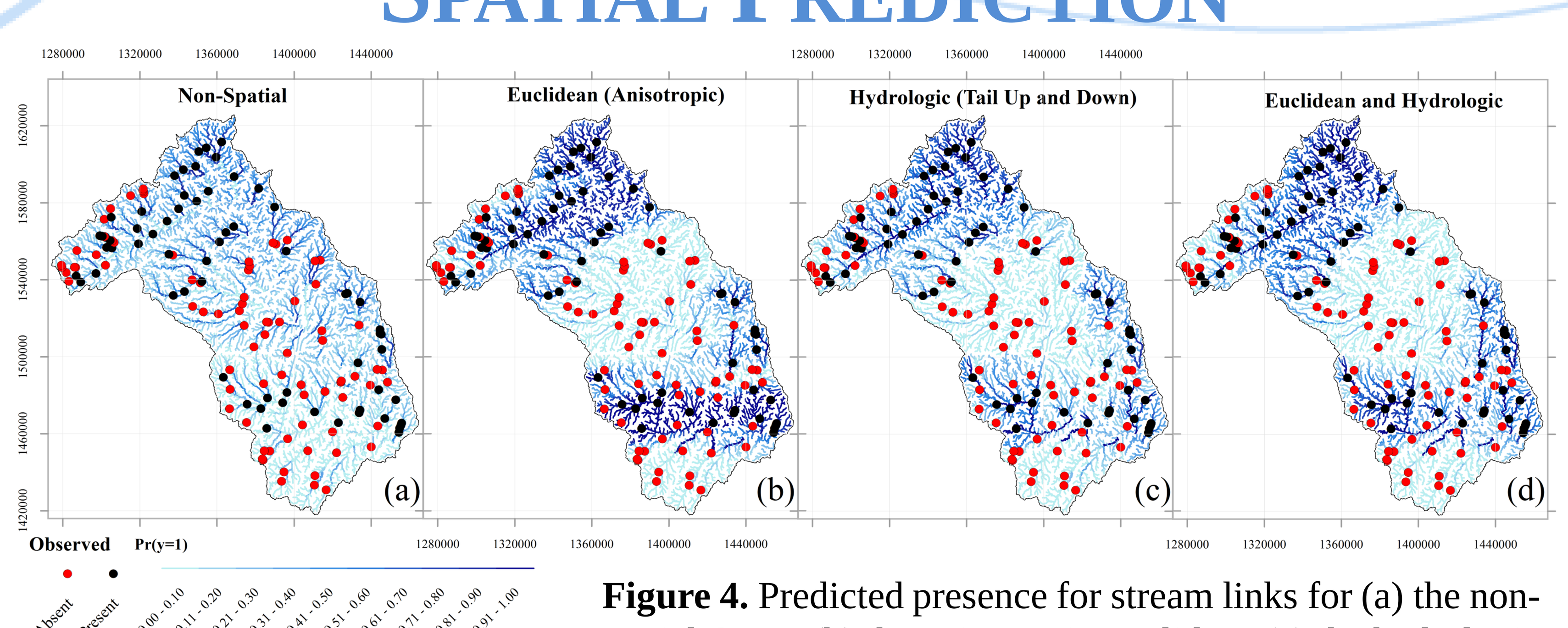

Figure 4. Predicted presence for stream links for (a) the non-spatial GLM, (b) the anisotropic Euclidean (c) the hydrologic (including total and flow-connected hydrologic distance) , and (d) all 3 covariance structures.

Results

- A global model containing all variables was selected for the fixed effects structure. Residuals from this model demonstrated autocorrelation patterns (Fig. 3)
- All models that included SARE had higher AUC and lower RMSPE values based LOOCV than the GLM (Table 1).
- Autocovariance structures that exclusively contained Euclidean autocorrelated random effects (isotropic and anisotropic) had better prediction (AUC) than structures that exclusively contained hydrologic random effects.
- Autocovariance structures that contained a mixture of Euclidean distance and hydrologic distance had better prediction and lower RMSPE (Table 1).
- Models that accounted flow connectivity along with other spatial structures had better prediction and lower RMSPE. However, models with ***only*** flow-connected covariance structures had poor prediction and high RMSPE.
- Models that included all covariance structures had the highest cross-validated prediction and lowest RMSPE.

Model Comparison

Table 1. Model structure, AUC, and RMSPE for the best model for each mixture type based AUC. Each covariance model is shown with the percent variance explained.

Mixture	Tail-Up	Tail-Down	Euclidean	AUC	RMSPE
TU/TD/EUC	Spherical (25%)	Exponential (27%)	Gaussian (40%)	0.880	0.368
TU/EUC	Linear-sill (47%)		Spherical (44%)	0.871	0.375
TD/EUC		Spherical (37%)	Gaussian (41%)	0.870	0.377
EUC (Aniso)			Spherical (57%)	0.863	0.384
EUC (Iso)			Exponential (59%)	0.854	0.390
TU/TD	Linear-sill (61%)	Spherical (28%)		0.853	0.388
TD		Linear-sill (40%)		0.842	0.399
TU	Exponential (85%)			0.816	0.416
Non-spatial				0.717	0.460

[1]TU stands for tail-up model (i.e., flow connected distance), TD stands for tail-down model (i.e., total hydrologic distance), and EUC stands for a Euclidean distance model.
[2] Anisotropic models were fit for autocovariance structures that only included Euclidean distance

Discussion and Conclusions

- The approach used in this study facilitated the addition of more ecologically relevant proximity matrices into geostatistical models.
- The comparison of the different models demonstrated that adequate specification of spatial dependence improves the accuracy of spatial predictions.
- Although hydrologic distance and connectivity are more ecological relevant for stream networks, Euclidean covariance structures had better prediction and explained a larger amount of variance (Table 1) when all three proximity matrices were used.
- Future improvements can be made to fluvial geostatstical models: (1) multivariate or co-kriging approaches, (2) incorporation of fluvial barriers (e.g., dams), and (3) incorporation into nonparametric approaches, including machine learning techniques.

References

[1] Peterson, E.E., and J. M. Ver Hoef. 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology **91**: 644-651

[2] Ver Hoef, J. M. and E.E. Peterson. 2010. A moving average approach for spatial statistical models of stream networks. JASA **105**: 6-18.