2-2008

Improving Predicted Distribution Models for Riverine Species: An Example from Nebraska

Scott P. Sowa
University of Missouri, 4200 New Haven Road, Columbia, MO

Gust Annis
University of Missouri, 4200 New Haven Road, Columbia, MO

Michael E. Morey
University of Missouri, 4200 New Haven Road, Columbia, MO

A. Garringer
University of Missouri, 4200 New Haven Road, Columbia, MO

AQUATIC

Improving Predicted Distribution Models for Riverine Species: An Example from Nebraska

Scott P. Sowa[1], Gust Annis[1], Michael E. Morey[1], and A. Garringer[1]

Introduction

Modeling species distributions is in most instances, we believe, better if perceived as an exercise in modeling spatial patterns in habitat conditions. This perspective forces the modeler to think about factors and processes that influence local habitat and also to account for as many of these factors as possible in the modeling process. Local habitat conditions in riverine ecosystems (for example, pH, temperature, turbidity, permanence of flow, depths, velocities, substrate, cover, primary production, etc.) are influenced by a wide array of factors and processes operating at multiple spatial and temporal scales (Matthews 1998; Fausch et al. 2002). However, of primary importance is the interplay of watershed and local conditions (Hynes 1975; Richards et al. 1996; Rabeni and Sowa 2002). For instance, local substrate conditions are influenced by water and sediment delivery which are largely determined by watershed conditions and also local geomorphic conditions (for example, channel gradient) that affect sediment transport (Jacobson and Pugh 1999).

Until recently it has been essentially impossible to quantify watershed conditions for thousands of streams segments across large geographic areas (for example, entire states). For this and other reasons, species distribution models developed for the Missouri Aquatic GAP Project were based on only a handful of local habitat variables (Sowa et al. 2007). This pilot project illustrated the importance and utility of these local variables for modeling the distribution of riverine biota, however, the resulting models had relatively low accuracy. We recently completed a project, involving development of statewide predicted distributions for fishes of Nebraska, in which we were able to quantify both watershed and local conditions for essentially all stream segments in the state and use them in the modeling process. Results from this project, which is the focus of this article, provide a specific example of how using both watershed and local variables for modeling the distribution of riverine biota can significantly improve model accuracy.

Methods

Methods used to develop the predicted distribution maps for fishes of Nebraska were essentially the same as those used in the Missouri Aquatic GAP Project (Sowa et al. 2005; Sowa et al. 2006). For the sake of brevity we will focus mainly on those elements of the methods that we believe led to improved accuracy of the Nebraska models compared to those of Missouri.

Species Data and Range Maps

We obtained 6,623 fish community collection records from the Nebraska Game and Parks Commission (NGPC) and the Nebraska Department of Environmental Quality (NDEQ). Collections made between 1857 and 2001, include 2,914 distinct stream segments and contain 41,130 species occurrence records for the 100 fish species that occur in Nebraska.

Using ArcGIS 9.1 (ESRI 2005), each collection was geographically linked to the 1:100,000 11-digit Hydrologic Unit (HU) coverage. Digital range maps, based on 11-digit HUs, were constructed for each species, submitted for professional review, and revised as necessary.

[1]Missouri Resources Assessment Partnership, School of Natural Resources, University of Missouri, 4200 New Haven Road, Columbia, MO 65201.

GIS Base Layer for Predictive Modeling

Each collection also was geographically linked to the Nebraska 1:100,000 valley segment type (VST) coverage (Sowa et al. 2005), which served as the base layer for developing the predicted distribution models. The finest resolution ("linear spatial grain") of our predictions was the stream segment. Within Nebraska there are 62,941 individual stream segments in the VST coverage with an average length of 2.0 km.

Predictor Variables

Eight local and 14 watershed variables were used as potential predictors (Table 1). Local variables were quantified for all 62,941 stream segments following the methods of Sowa et al. (2007) and represent the same variables used to predict species distributions in the Missouri Aquatic GAP Project. Watershed variables were quantified for all but 323 segments of the Missouri River due to a lack of time and money needed to quantify physiographic conditions throughout the enormous watersheds of these segments (for example see: Figure 1) (Sowa et al. 2006).

Table 1. Descriptions for the 23 local and watershed predictor variables.

Local variable	
Flow	Binary variable that differentiates perennial and intermittent flow.
Temp	Binary variable that differentiates cold and warm water streams.
Linkr10	A ten category description of stream size based on Shreve Link magnitude (Shreve 1966).
sdiscr_2c	Binary variable that differentiates stream segments that flow into either the same size stream or a larger stream.
grdseg10	A ten category designation of stream segment gradient (m/km).
neb_geol	A 14 category variable designating the surficial geology through which each stream segment flows.
stxt4cat	A 4 category variable designating the general soil texture class through which each stream segment flows.
drn_grp	A 5 category variable designating the major drainage group in which a given stream segment occurs.
Watershed variable	
avegrd10	Average gradient of all stream segments in the watershed.
hyda_p	Percent of watershed containing Hydrologic Soil Group A placed into ten categories.
hydb_p	Percent of watershed containing Hydrologic Soil Group B placed into ten categories.
hydc_p	Percent of watershed containing Hydrologic Soil Group C placed into ten categories.
hydd_p	Percent of watershed containing Hydrologic Soil Group D placed into ten categories.
stxt01_p	Percent of watershed containing Soil Surface Texture Class 1 (Sand) placed into ten categories.
stxt02_p	Percent of watershed containing Soil Surface Texture Class 2 (Loamy sand) placed into ten categories.
stxt03_p	Percent of watershed containing Soil Surface Texture Class 3 (Sandy loam) placed into ten categories.
stxt04_p	Percent of watershed containing Soil Surface Texture Class 4 (Silt loam) placed into ten categories.
stxt06_p	Percent of watershed containing Soil Surface Texture Class 6 (Loam) placed into ten categories.
stxt08_p	Percent of watershed containing Soil Surface Texture Class 8 (Silty clay loam) placed into ten categories.
stxt09_p	Percent of watershed containing Soil Surface Texture Class 9 (Clay loam) placed into ten categories.
stxt11_p	Percent of watershed containing Soil Surface Texture Class 11 (Silty clay) placed into ten categories.
stxt12_p	Percent of watershed containing Soil Surface Texture Class 12 (Clay) placed into ten categories.

Figure 1. Map of Nebraska streams showing percentage of the watershed for each stream segment that contains soils classified as Hydrologic Soil Group A.

Statistical Methods

Models were constructed with version 14 of the Classification Tree add-on of SPSS version 14.0 (SPSS, Inc. 2005). The specific modeling algorithm we used was Exhaustive CHAID, which is a modification of CHAID (Kass 1980) developed by Biggs et al. (1991). We generated species-specific input datasets containing a row for each of 6,623 collection records, a column for the binary species response variable (1=present, 0=absent), and columns for each of the 23 predictor variables.

We set the minimum number of collections allowable in a parent node equal to 10 percent and the number allowable in a child node equal to 1 percent of the total occurrence records for each species. We set the alpha level for splitting and merging equal to 0.05 and used the Bonferroni alpha adjustment to account for the increased likelihood of a Type One error associated with multiple comparisons (Bonferroni 1935).

The above methods were used to model distributions of most fish species. Alternative methods were used for species having too few occurrence records in order to generate a model and those species that do not occur outside of the Missouri River mainstem (Sowa et al. 2006).

Model Outputs

Probability of Occurrence

Each terminal node in a classification tree model provides a probability of occurrence for a given species under a certain set of conditions. These probabilities can be applied to an independent dataset using the suite of if/then model statements generated by SPSS. For each species we applied the resulting if/then statement model to the attribute table of the statewide 1:100,000 VST coverage (Figure 2). This process produced a column in the attribute table for that particular species which provides the probability of occurrence for each of the 62,618 stream segments in the state. However, all stream segments falling outside the professionally-reviewed geographic range were converted to zero probability.

Presence

Calculating richness or diversity measures requires explicit yes or no statements about species presence, which are not provided with a continuous probability of occurrence. In many instances, modelers deem a species as being present at locations where it has greater than 50 percent probability of occurrence (Csuti and Crist 1998). However, due to sampling biases and inefficiencies, species with low detection probabilities rarely have occurrence probabilities greater than 50 percent and would therefore never be predicted as "present." In fact, most fish species modeled in this project have maximum occurrence probabilities below 50 percent (Sowa et al. 2006).

Figure 2. Map of predicted occurrence probabilities for the longnose dace (*Rhinichthys cataractae*) throughout Nebraska.

To overcome this problem we used the "relative-50%" rule developed by Sowa et al. (2005) to generate a binary presence/absence model for each species. Specifically, for each model we identified the terminal node having the highest occurrence percentage that also contained at least 50 collection records. We then multiplied this highest percentage by 0.5 and selected all terminal nodes with occurrence probabilities greater than or equal to this percentage (Figure 3). These selected segments were then attributed with a value of 1 to denote presence, while all other segments were attributed with a 0 in a separate attribute field for each species. Again, all segments outside of the geographic range of the species were attributed with a 0.

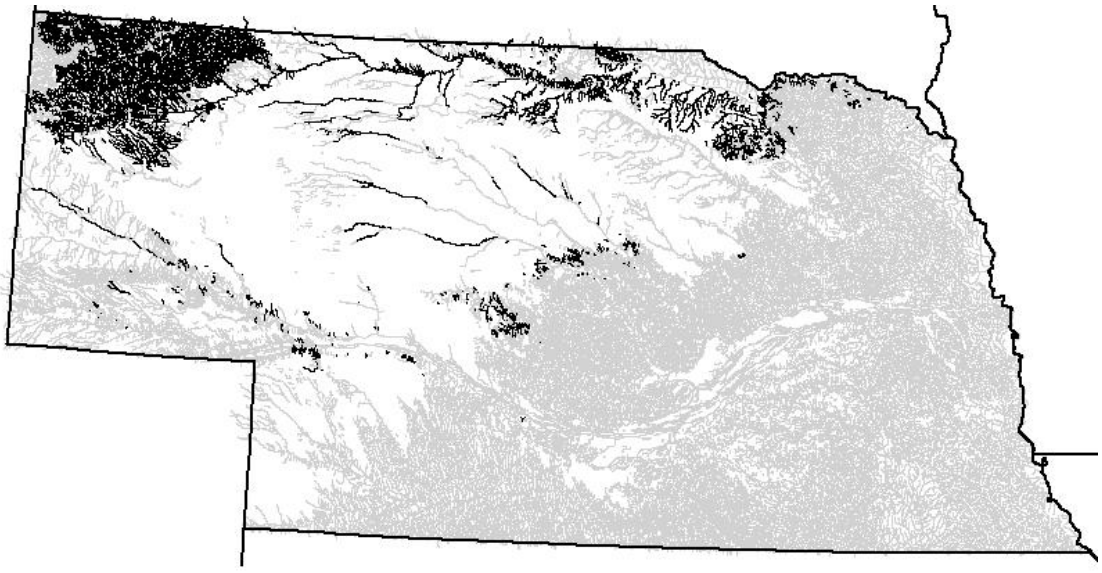

Figure 3. Map of predicted occurrence (in black) for the longnose dace (*Rhinichthys cataractae*) throughout Nebraska. Predicted occurrence was based on a relative 50 percent rule (see text). In this instance the highest occurrence probability, with sufficient samples, was 91 percent. This map shows all segments with occurrence probabilities greater than or equal to (0.5 times 91) or 45.5 percent. Overall accuracy of this model was 92 percent.

Results and Discussion

Lacking an independent dataset, we assessed the accuracy of the presence models against the input data used to create the models. For each species, we calculated omission (species occurs, but not predicted), commission (species predicted, but does not occur), and overall error rates. Species-specific error rates are provided in Sowa et al. (2006) and the average error rates across all 100 species are provided in Table 2. The overall error rate was only 8 percent (Table 2). This is significantly less than the 49 percent overall misclassification rate for fishes in the Missouri Aquatic GAP Project. Average omission (3 percent) and particularly commission (6 percent) error rates were also significantly lower than what was achieved in Missouri (Table 2) (MO: omission: 10 percent; commission: 48 percent).

Considering that local habitat conditions in rivers and streams are significantly influenced by physiographic conditions in the watershed (Hynes 1975; Frissell et al. 1986), we believe the addition of 15 watershed variables as potential predictors was the most important factor leading to the improved accuracy of the models in Nebraska compared to Missouri. These watershed variables dominated our classification tree models, which contrasts with what Oakes et al. (2005) determined in a similar project that modeled fish distributions throughout the Big Blue River watershed in Kansas and Nebraska. However, as Wiens (1989; 2002) points out, such differences in the perceived importance of explanatory variables should be expected among studies when either the spatial grain or extent of the investigation differs. While the variables and spatial grain of our modeling efforts were similar to that of Oakes et al. (2005), the significantly larger spatial extent of our project (entire state vs. single watershed) covered a much wider range of physiographic conditions that influence stream habitat, which likely led to the increased predictive capabilities of the watershed variables in our models.

There were two other notable factors that also likely increased the accuracy of the models we developed for Nebraska. First, we had nearly twice as many collection records for Nebraska fishes (6,623) than we did for Missouri (3,723). All other things being equal, increasing the number of species occurrence records should increase model accuracy. Second, the collections for Nebraska covered a longer time frame (NE: 1857-2001; MO: 1900–1999) and had a substantially higher number of historical and reference-quality samples. Collections from highly disturbed locations will tend to decouple relations between species occurrence and natural features of the environment, which was the objective of our modeling efforts. The higher number of historical and reference-quality samples likely improved model accuracy.

Table 2. Average accuracy statistics for occurrence models developed for 100 Nebraska fishes.

	Average (percent)	Minimum	Maximum
Omission	3	0	19
Commission	6	0	33
Overall	8	0	38

Finally, we need to point out one last and very important difference between the models developed for Missouri and those developed for Nebraska. This difference does not pertain to the issue of accuracy, but rather the utility of the end products. The classification tree models we generated with the methods presented above are extremely complex. Manually applying hundreds of resulting if/then model statements (for a single model) to an independent dataset is essentially impossible to do for hundreds of species, not to mention doing this task without human error. Because of this problem, for Missouri we were only able to generate binary presence/absence attributes in the attribute file of the statewide VST coverage for each species, despite having models that provided occurrence probabilities.

Improvements in the SPSS software (SPSS 2005), since we modeled species distributions in Missouri, allow the resulting models to be applied to an independent dataset. This software advancement allowed us to attribute the Nebraska VST coverage with continuous probabilities of occurrence for each species. These continuous probabilities provide users with significantly more information on which to base decisions and greater flexibility in their use. In fact, we are currently working with the Nebraska Game and Parks Commission to use these occurrence probabilities to develop optimized sampling designs for locating additional populations of twelve at-risk fish species.

Predicted distribution models are a fundamental component of all GAP projects (Csuti and Scott, 1991), which is why the National Gap Analysis Program has been at the forefront of meeting this critical data need for conservation planning across the United States (Maxwell, 2006). GAP also has been a leader in addressing many research and technical issues surrounding this complex endeavor as evidenced by the number of peer-reviewed publications on this topic by gap practitioners (see http://gapanalysis.nbii.gov/). Considering the importance of species distribution data for resource planning and management (cf. Scott et al. 2002; Brooks et al. 2004; Pressey 2004), it is essential that we continually strive to develop the most accurate and precise distribution models possible.

Until recently it has been essentially impossible to quantify watershed conditions for tens of thousands of individual stream segments across large geographic areas. Fortunately, recent technological and methodological advancements have allowed us to overcome this obstacle, but it is still somewhat costly and time consuming to generate these watershed data. However, we believe that all future efforts to model the distributions of riverine biota across large regions should take the extra time, money and effort to incorporate watershed variables into the modeling process. The gains in model accuracy certainly outweigh the additional costs.

References Cited

Biggs, D., B. de ville, and E. Suen. 1991. A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics* 18: 49-62.

Bonferroni, C.E. 1935. Il calcolo delle assicurazioni su gruppi di teste. Pages 13-60 In *Studi in Onore del Professore Salvatore Ortu Carboni*. Rome: Italy.

Brooks, T.M., G.A. da Fonseca, and A.S.L. Rodrigues. 2004. Species, data, and conservation planning. *Conservation Biology* 18:1682-1688.

Csuti, B. and P. Crist. 1998. Methods for assessing accuracy of animal distribution maps (version 2.01). In *A Handbook for Conducting Gap Analysis*. USGS Gap Analysis Program, Moscow, Idaho. Available online at: http://www.gap.uidaho.edu/handbook

Csuti, B. and J.M. Scott. 1991. Mapping wildlife diversity for gap analysis. *Western Wildlands* Fall: 13-18.

ESRI 2005. ArcGIS 9.1. Environmental Systems Research Institute, Redlands, CA.

Fausch, K.D., C.E. Torgersen, C.V. Baxter, and H.W. Li. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *Bioscience* 52(6):483-498.

Frissel, C.A., W.J. Liss, C.E. Warren, and M.D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. *Environmental Management* 10: 199-214.

Hynes, H.B.N. 1975. *The stream and its valley*. Verh. Int. Theor. Ang. Limnol. 19:1-15.

Jacobson, R.B. and A.L. Pugh. 1999. Riparian vegetation controls on the spatial pattern of stream-channel instability, Little Piney Creek, Missouri: U.S. Geological Survey Water-Supply Paper 2494.

Kass, G. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29: 119-127.

Matthews, W.J. 1998. *Patterns in Freshwater Fish Ecology*. Chapman and Hall, New York, NY.

Maxwell, J. 2006. Role of GAP data in state wildlife plan development: opportunities and lessons learned. In Maxwell et al., eds. *Gap Analysis Bulletin No. 14:4-10*. USGS/BRD/ Gap Analysis Program. Moscow, ID. Available online at http://www.gap.uidaho.edu/Bulletins/14/Maxwell.htm

Oakes, R.M., K.B. Gido, J.A. Falke, J.D. Olden, and B.L. Brock. 2005. Modelling of stream fishes in the Great Plains, USA. *Ecology of Freshwater Fish* 14:361-374.

Pressey, R.L. 2004. Conservation planning for biodiversity: Assembling the best data for the job. *Conservation Biology* 18:1677-1681.

Rabeni, C.F. and S.P. Sowa. 2002. A landscape approach to managing the biota of streams and rivers. Pages 114-142 In J. Liu and W. Taylor, eds. *Integrating Landscape Ecology Into Natural Resource Management*. Cambridge University Press. Cambridge, MA.

Richards, C., L.B. Johnson, and G.E. Host. 1996. Landscape-scale influences on stream habitat and biota. *Journal of Fisheries and Aquatic Sciences* 53 (Suppl. 1):295-311.

Scott, J.M., P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall, and F.B. Samson, eds. 2002. *Predicting Species Distributions; Issues of accuracy and scale*. Island Press, Washington, D.C.

Sowa, S.P., D.D. Diamond, R. Abbitt, G. Annis, T. Gordon, M.E. Morey, G.R. Sorensen, and D. True. 2005. *A Gap Analysis for Riverine Ecosystems of Missouri. Final Report*, submitted to the USGS National Gap Analysis Program. Moscow, ID. 1675 pp.

Sowa, S.P., G. Annis, M.E. Morey, and A. Garringer. July 2006. Developing predicted distribution models for fish species in Nebraska. Final report submitted to the USGS National Gap Analysis Program. Moscow, ID. 489 pp.

Sowa, S.P., G. Annis, M.E. Morey, and D.D. Diamond. 2007. A gap analysis and comprehensive conservation strategy for riverine ecosystems of Missouri. *Ecological Monographs* 77(3): 301-334.

SPSS Inc. (2005). *SPSS Base 14.0 for Windows User's Guide*. SPSS Inc., Chicago IL.

Wiens, J.A. 1989. Spatial scaling in ecology. *Functional Ecology* 3:385-397.

Wiens, J.A. 2002. Riverine landscapes: taking landscape ecology into the water. *Freshwater Biology* 47:501-515.