Development of a Statistical Model to Estimate the Likelihood of Observing Freshwater Mussels in Maryland Streams

Maryland
Department of
the Environment

May 7, 2021

Developed and distributed by:
Water Quality Standards Section
Environmental Assessment & Standards Program (EASP)
Water and Science Administration (WSA)
Maryland Department of the Environment
1800 Washington Boulevard
Baltimore, MD 21230
Phone: 410-537-3818

Principal Author: Timothy Fox
Water Quality Standards Section
Environmental Assessment & Standards Program

Acknowledgements

Reviewers

Matthew Ashton, Maryland Department Natural Resources, 580 Taylor Ave, Annapolis, MD

Yen-Der Chang, Maryland Department of the Environment, 1800 Washington Blvd, Baltimore, MD

Lisa Ochsenhirt, Maryland Association of Municipal Wastewater Agencies

Fred Pinkney, United States Fish and Wildlife Service, Chesapeake Bay Field Office, 177 Admiral Cochrane Dr., Annapolis, MD

Matthew Stover, Maryland Department of the Environment, 1800 Washington Blvd, Baltimore, MD

Contents

List of Tables

Introduction

Freshwater mussels are distributed throughout Maryland's surface waters and are uniquely sensitive to anthropogenic pollution and land-disturbing activities. Permitted activities and associated water quality standards must consider freshwater mussel sensitivities when protecting the general aquatic life of Maryland's surface waters. However, the Clean Water Act allows for the recalculation of water quality criteria when it can be shown that sensitive species (such as freshwater mussels) are absent from a delineated section of surface water (USEPA 2013). Data have shown that the distribution of freshwater mussels is not only limited by anthropogenic factors but also by natural abiotic stream characteristics. Therefore, protection of freshwater mussels may not be a necessary component of all surface waters.

Data from the Maryland Biological Stream Survey (MBSS) were used to develop a logistic regression model to predict the likelihood of observing freshwater mussels. The goal of the effort was to develop a model that could be used to determine if a given 75-meter MBSS sampling transect was highly unlikely to provide freshwater mussel habitat given certain variables collected within the transect. Five predictor variables were evaluated: average stream width, flow, gradient, upstream acreage, and ecoregion. A training dataset was used to validate the model. All of these variables are collected by MBSS during the spring and summer sampling seasons. Because *Corbicula fluminea* is a non-native, non-Unionid species, observations that included this species were removed from the analysis.

The model that was selected includes all five predictor variables with interaction terms. The signs of the predictor coefficients are consistent with the known habitat limitation of freshwater mussels. When the model is extrapolated to very low probabilities, it does a very good job of predicting when mussels are absent. An examination of the MBSS data showed that when the predictor variables estimate a probability below 0.03, mussels are extremely unlikely to be present.

Summary of Model Generation and Selection

This section outlines statistical tools and assumptions that were used to derive a model that best predicts when freshwater mussels are absent from a 75-meter sampling transect. An overview of this approach is summarized below:

- Multiple logistic regression was the technique selected to model the probability of freshwater mussels being observed.
- The MBSS dataset classifies an observation of the species *Corbicula fluminea* as a freshwater mussel observation. Because *Corbicula fluminea* is a non-native, non-Unionid species, samples that included this species were removed from the analysis.
- The repeatability of observing mussels in an 8-digit watershed for a given stream order was estimated by analyzing data from sites that were sampled on multiple occasions.
- A *training* multiple logistic regression model was generated using data from sites that were sampled prior to 2014. The model was then applied to estimate the probability of observing mussels in samples taken in the years 2014, 2015 and 2016. The estimated probabilities of observing mussels were compared to actual mussel observations.

- Results show that the logistic regression model performs adequately when attempting to identify streams that are unlikely to provide freshwater mussel habitat.

The Maryland Biological Stream Survey

The Maryland Biological Stream Survey (MBSS) was initiated by the Maryland Department of Natural Resources in 1993 and was Maryland's first probability-based stream sampling program intended to provide unbiased estimates of stream conditions with known precision at various spatial scales. The MBSS consists of a multi-stratification sampling design that ensures all 1st through 4th order non-tidal streams in the sampling frame have a non-zero probability of being sampled (Southerland et al, 2005).

MDE derived a logistic regression model using data from approximately 1300 sites collected by the MBSS. The MBSS collected these data from randomly selected stream locations throughout Maryland with each station consisting of a 75-meter stream transect. The MBSS database has recorded mussel observations along with several other habitat and abiotic factors at hundreds of locations.

Overview of Logistic Regression

Logistic regression models are used to quantify a relationship between one or more predictor variables and a categorical response variable. In the case of "binary" logistic regression, we can estimate the probability of two possible disjoint outcomes given certain predictors. Binary logistic regression models have been widely used to predict the probability of observing biota, and the purpose of this exercise was to develop a logistic regression model that predicts the probability of observing freshwater mussels given certain abiotic predictors. Ideally, this model will allow the Department to identify streams that have a low probability of observing mussels and therefore could permit the use of recalculated water quality criteria (and associated water quality-based effluent limits (WQBEL) based on those criteria) that did not include mussels in the derivation. The logistic regression model derived from the MBSS data used the probability of mussel observations as a response variable and the following abiotic factors as predictors:

- Log Stream Gradient (Slope)
- Log Discharge Rate (Flow)
- Log Average Stream Width (Width)
- Log Upstream Catchment Area (Area)

But what probability should be considered a "low probability of observing a freshwater mussel"? To provide sufficient protection to freshwater mussels, MDE is proposing to use the 0.03 threshold to identify streams that are unlikely to provide mussel habitat. More specifically, MDE proposes that if the logistic regression model predicts a less than 0.03 probability of observing a mussel, the discharger may have the option to use a WQBEL based on a site-specific criterion that does not include freshwater mussels.

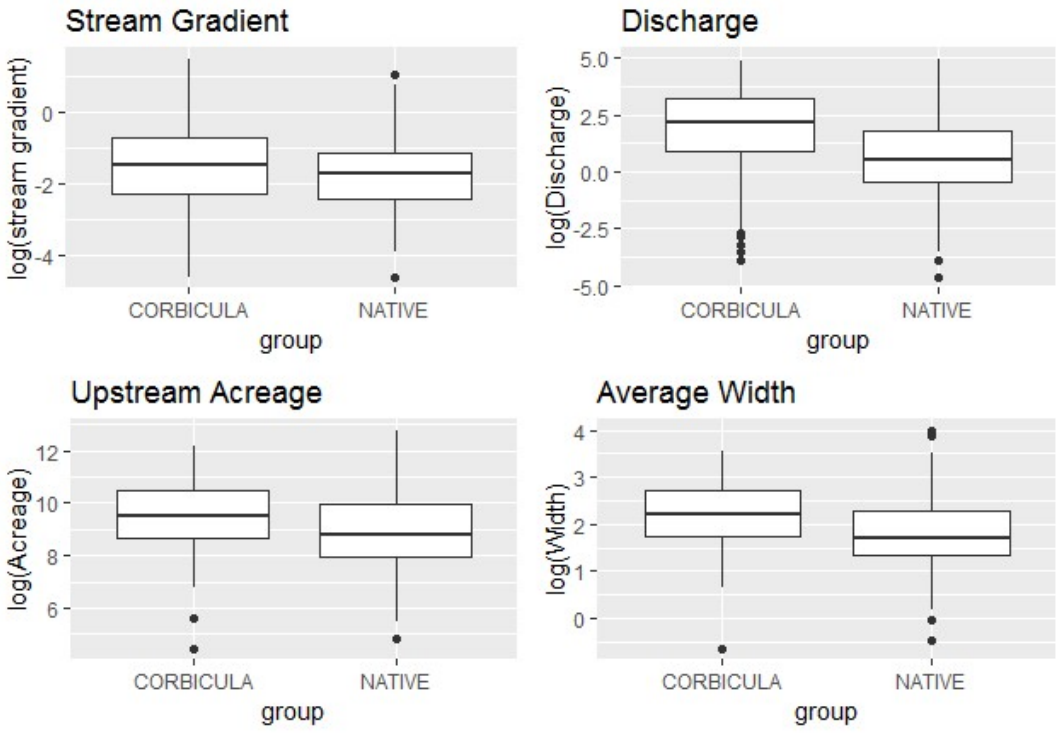
Removing non-Unionid Bivalves

In this analysis, *C. fluminea* were removed, because it was possible that the presence of a non-native bivalve may be a confounding factor limiting the distribution of native mussels. Studies have shown that

the presence of *C. fluminea* displaces or extirpates native mussel populations (McMahon 1983; Hakenkamp and Palmer 1999; Cohen et al. 1994; Lauritsen 1986; Beaver et al. 1991; Hakenkamp et al 2001; Phelps 1994). Because MDE wanted to reduce the amount of possible biotic confounding factors, sampling events that showed the presence of *C. fluminea* (and no other native mussel) were not used in model derivation. Although the distribution of native mussels in Maryland differs slightly from *C. fluminea* in Maryland, there is significant overlap (Figure 1).

Removing sampling events from the dataset that observed *C. fluminea* reduced the number of sites classified as "no native mussels observed". Specifically, 310 *C. fluminea* sampling events were removed which reduced the sample dataset from 1619 to 1309. The number of native mussel observations in the model building dataset was not changed by removing *C. fluminea* observations because there were no sampling events that included both a native mussel observation along with a *C. fluminea* observation and included the four predictor variables. MDE concluded that 1309 samples were still sufficient to derive a robust statistical model and that removing confounding factors was vital to accurately predict the probability of mussel absence.

Figure 1: Distribution of Native Mussels and Corbicula fluminea by Abiotic Stream Predictor Variable



Assessing Repeatability of Mussel Observations

The objective of this analysis was to estimate the probability of a mussel observation occurring given that a mussel observation had occurred previously at a location. Another objective was to estimate the probability of no mussel observation occurring given that mussels were not previously observed at that

location. If the resulting estimated probabilities were relativity high, it would add some confidence that mussel observations are repeatable. A subset of stream segments was identified that were sampled in multiple years.

In the MBSS database, a new field was created that represented the first 13 characters of the "SITEYR" field. For example, ABPG-302-R-2000 was grouped by ABPG-302-R-20 to allow for multiple years to be grouped by this identifier. In this case, it was assumed that ABPG-302-R-2001, ABPG-302-R-2002, ABPG-302-R-2003, and so on represent sampling events occurring within the same 8-digit watershed and same stream order. In the case of sentinel sites, it was assumed that the sampling occurred on the same stream segment. A subset of all sites with 2 or more observations was formed to identify 8-digit watersheds that were visited repeatedly for a given stream order. A total of 286 sites were identified. Of these sites, 106 had mussel observations on every sampling event and 156 had no mussel observations on any site visits. Therefore 262 sites had consistent mussel observations results (92%). So, we can be reasonably confident that when mussel surveys take place at a given 8-digit watershed, the results are repeatable. It is important to note that this analysis did not provide information on the actual probability of mussels being absent from a site. However, this analysis did add some confidence that mussel observation results were consistent.

Model Building and Validation

The objective of this analysis was to determine if the predictors and model building method produced a logistic regression model that could effectively predict the probability of not observing freshwater mussels.

To accomplish this, a data splitting technique was employed. This technique splits the data set into two sets. The first set, called the model-building set or the *training sample*, is used to develop the test model. The second data set, called the validation or *prediction set*, is used to evaluate the reasonableness and predictive ability of the test model (Kutner et al 2004, 372). The MBSS dataset was divided into two subsets: a larger (training sample) dataset consisting of samples taken before 2014 and a smaller dataset (prediction set) consisting of sites taken on or after 2014. A total of 960 sampling events occurred in the larger dataset and 349 sites in the smaller dataset. Several candidate logistic regression models were derived using the larger dataset. The candidate predictors considered were:

- Log Stream Gradient (numeric variable)
- Log Discharge Rate (numeric variable)
- Log Average Stream Width (numeric variable)
- Log Upstream Catchment Area (numeric variable)
- Ecoregion dummy variables (Coastal, Piedmont, and Highland)

The best model was identified using the Akaike's Information Criterion (AIC). This criterion uses the sum of squared error, number of parameters in the model, and the sample size to balance error reduction with statistical model parsimony. When comparing different statistical models using this criterion, the model with the lowest AIC is considered the best. The following table displays the model with the best AIC of the different model options.

Table 1: Model Coefficients Derived Using Training Data

Model Derived Using Training data		
Coefficients	**Estimate**	**p-value**
Intercept	-8.65282	<0.00001
Log Stream Gradient	-1.98353	0.000195
Log Discharge Rate	-0.18772	0.220051
Log Average Stream width	1.91912	0.000980
Log upstream Acreage	0.37252	0.054686
Coastal	0.24502	0.873552
Log Average Stream width*Coastal	-2.25131	0.000637
Log upstream Acreage* Coastal	0.74879	0.002195
Log Discharge Rate *Coastal	0.41120	0.020139
Log Stream Gradient* Log upstream Acreage	0.17326	0.001748
Log Stream Gradient *Coastal	0.47049	0.010822

Based on the AIC, the Piedmont parameter was not needed in this model, indicating that there was not a significant difference between the Highland and Piedmont models. This logistic regression model was used to estimate the probability of observing mussels in sites sampled on and after 2014 (prediction set). Of the 349 sites in this prediction set, 136 sites had a predicted probability less than 0.03 and 213 had a predicted probability of greater than 0.03. Of the 136 sites that had a predicted probability less than 0.03, none had mussels present. Of the 213 sites that had a predicted probability greater than 0.03, only 30 had mussels present.

Table 2: Contingency Table of Training Model Using 0.03 as Threshold

		Mussels Present?		
		Yes	No	Totals
Model prediction	Predicted Less than 0.03	0	136	136
	Predicted Greater than 0.03	30	183	213
	Total	30	319	349

The threshold of "0.1 probability of observing a mussel" was also evaluated. The following table summarizes the results:

Table 3: Contingency Table of Training Model Using 0.1 as Threshold

		Mussels Present?		
		Yes	No	Totals
Model prediction	Predicted Less than 0.1	0	196	196
	Predicted Greater than 0.1	30	123	153
	Total	30	319	349

As the table indicates, no mussels were observed in the any of the 196 sites that had less than a 0.1 probability of observing mussels. Based on these results, the method of building the logistic regression model using these predictors seems reasonable if we use the model for the purpose of ensuring mussels are not going to be observed and using 0.03 as our threshold.

Full Model Development

In developing the full logistic regression model, the entire dataset was used. Only sampling events that reported all predictors were included in model development. A total of 1309 sampling events reported all five predictors.

The model with the best AIC included all predictors including the Piedmont and Coastal dummy variables (suggesting that there is a statistically significant difference between the Piedmont and Highland models when the sample size is larger). The following table summarizes the model coefficients:

Table 4: Model Coefficients Derived Using the Full Data Set

All Eco-regions With Interactions (all data)		
Coefficients	**Estimate**	**p-value**
Intercept	-13.76449	<0.00001
Log Stream Gradient	-1.33395	0.000918
Log Discharge (CFS)	0.54472	0.184373
Log Average Stream width	1.47645	0.00277
Log upstream Acreage	0.98029	<0.00001
Coastal	6.05889	<0.00001
Piedmont	5.78991	0.00345
Log Discharge (CFS)*Log upstream Acreage	-0.11569	0.036793

Log Average Stream width* Coastal	-1.73775	0.000809
Log Discharge (CFS)* Coastal	0.37623	0.043085
Piedmont* Log upstream Acreage	-0.64963	0.002434
Log Stream Gradient* Log upstream Acreage	0.14286	0.002369
Log Stream Gradient* Piedmont	-0.45549	0.013474
Log Discharge (CFS)*Log Average Steam width	0.21296	0.083798

The percentage of sampling events with predicted probability less than 0.03 for all eco-regions are summarized below. The tables below show that 33.5 percent of sample sites have a low predicted probability of observing a mussel based on the model with interaction terms. The proportion of sites is smaller in the Coastal Plain and higher in the Piedmont and Highlands. Furthermore, the number of actual mussel observations in sites with predicted probability of observing a mussel is proportionally low.

Table 5: Model Output by Ecoregion

	All eco-regions	Coastal Plain	Piedmont	Highlands
Total <0.03	438	29	208	201
Total samples	1309	586	388	335
Percentage < 0.03	33.5%	5%	53.6%	60%
Total actual mussel observations in sites with Equation 1, 2 or 3 output <0.03	4	0	2	2

The Maryland Department of Natural Resources was contacted regarding the four sites that were classified as having mussel observations, but which had calculated values below 0.03, and it was determined that these sites were outliers and do not provide mussel habitat. These four sites are described below:

- An unnamed tributary to Liberty Reservoir was sampled in 2011, and a half shell of a triangle floater (*Alasmidonta undulata*) was found. The sampling event is listed in the MBSS database as LIBE-191-X-2011. This particular species is tolerant of impoundment conditions, and was found

less than 100 meters from Liberty Reservoir. The shell fragment could have been placed there by a mammalian predator that had taken it from the reservoir.

- Creeper (*Strophitus undulatus)* was observed in Poplar Lick Run in 2000 and 2001 at the sampling events listed as SAVA-202-C-2000 and SAVA-202-C-2001. This site was immediately downstream of a mill dam which is stocked with salmonids. The mussel species present has been previously documented to use salmonids as a host and be introduced into low gradient patches of cold water streams that are stocked. Freshwater mussels are otherwise absent from this watershed.
- Another triangle floater (*Alasmidonta undulata*) was observed in the very lower reach of Tiber Run right before the confluence with the Patapsco River. There have been confirmed observations of triangular floater in the Patapsco River. The sampling event was classified as PATL-207-R-2000 in the MBSS database. Only one shell was observed, and over five years of subsequent monitoring at the site, no evidence of mussels was found in the smaller stream. The shell fragment could have been placed there by a mammalian predator that had taken it from the Patapsco River.

Conclusion

The logistic regression model with the coefficients listed in Table 4 will be used as a basis for determining if a stream is likely to have freshwater mussel habitat. The use of this model (with interaction terms) provided accurate predictions of the absence of freshwater mussels when the model output is below 0.03.

It is important to note that the model cannot be used to predict the presence of freshwater mussels. If the model is extrapolated to higher probability outputs, the model does a poor job of predicting mussel observations. This is likely due to other factors limiting mussel distributions that are not incorporated into the model.

The use of the threshold of 0.03 to determine if a 75-meter stream transect does not have freshwater mussel habitat will ensure that freshwater mussels in Maryland streams are protected while providing flexibility to the regulated community.

The following tables provide the model prediction accuracy for each ecoregion at the model outputs of 0.1 and 0.03. Ultimately, the model output threshold of 0.03 was selected for use by the Department.

Table 6: Contingency Tables of Model Output by Ecoregion and Threshold

		Mussels Present?		
		Yes	No	
Model prediction (all eco-regions)	Predicted Less than 0.1	13	596	609
	Predicted Greater than 0.1	338	362	700
	Total	351	958	1309

		Mussels Present?		
		Yes	No	
Model prediction (Coastal Plain)	Predicted Less than 0.1	2	81	83
	Predicted Greater than 0.1	271	232	503
	Total	273	313	586

		Mussels Present?		
		Yes	No	
Model prediction (Piedmont)	Predicted Less than 0.1	4	269	273
	Predicted Greater than 0.1	41	74	115
	Total	45	343	388

		Mussels Present?		
		Yes	No	

Model prediction (Highlands)	Predicted Less than 0.1	7	246	253
	Predicted Greater than 0.1	26	56	82
	Total	33	302	335

Model summary using 0.03 as threshold

		Mussels Present?		
		Yes	No	
Model prediction (all eco-regions)	Predicted Less than 0.03	4	425	429
	Predicted Greater than 0. 03	348	533	881
	Total	352	958	1310

		Mussels Present?		
		Yes	No	
Model prediction (Coastal Plain)	Predicted Less than 0. 03	0	29	29
	Predicted Greater than 0. 03	273	284	557
	Total	273	313	586

		Mussels Present?		
		Yes	No	
Model prediction	Predicted Less than 0. 03	2	206	208

(Piedmont)	Predicted Greater than 0. 03	43	137	180
	Total	45	343	388

		Mussels Present?		
		Yes	No	
Model prediction (Highlands)	Predicted Less than 0.03	2	199	201
	Predicted Greater than 0.03	31	103	134
	Total	33	302	335

References

Ashton, M. J. (2010). *Freshwater mussel records collected by the Maryland Department of Natural Resources' Monitoring and Non-tidal Assessment Division (1995-2009): Investigating environmental conditions and potential host fish of select species*. Maryland Department of Natural Resources, Resource Assessment Service, Monitoring and Non-Tidal Assessment Division.

Ashton, M. J. (2012). How a Statewide Stream Survey Can Aid in Understanding Freshwater Mussel (Bivalvia: Unionidae) Ecology: Examples of Utility and Limitations from Maryland. *Freshwater Mollusk Biology and Conservation*, *15*(1), 1-11.

Baldigo, B. P., Riva-Murray, K., & Schuler, G. E. (2004). Effects of environmental and spatial features on mussel populations and communities in a North American river. *Walkerana*, *14*(31), 1-32.

Beaver, J. R., Crisman, T. L., & Brock, R. J. (1991). Grazing effects of an exotic bivalve (Corbicula fluminea) on hypereutrophic lake water. Lake and Reservoir Management, 7(1), 45-51.

Campbell, C. A., & Hilderbrand, R. H. (2017). Using maximum entropy to predict suitable habitat for the endangered dwarf wedgemussel in the Maryland Coastal Plain. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *27*(2), 462-475.

Cohen, R. R., Dresler, P. V., Phillips, E. J., & Cory, R. L. (1984). The effect of the Asiatic clam, Corbicula fluminea, on phytoplankton of the Potomac River, Maryland. Limnology and Oceanography, 29(1), 170-180.

Grabarkewicz, J. & W. Davis. 2008. An introduction to freshwater mussels as biological indicators. EPA-260-R-08-15. U.S. Environmental Protection Agency, Office of Environmental Information, Washington, DC. 140 pp.

Hakenkamp, C. C., Ribblett, S. G., Palmer, M. A., Swan, C. M., Reid, J. W., & Goodison, M. R. (2001). The impact of an introduced bivalve (Corbicula fluminea) on the benthos of a sandy stream. Freshwater Biology, 46(4), 491-501.

Hakenkamp, C. C., & Palmer, M. A. (1999). Introduced bivalves in freshwater ecosystems: the impact of Corbicula on organic matter dynamics in a sandy stream. Oecologia, 119(3), 445-451.

Kutner, M.H., C. J. Nachtsheim and J. Neter. (2004). Applied Linear Regression Models, Fourth Edition. McGraw-Hill Irwin. 701 pp.

Lauritsen, D. D. (1986). Assimilation of radiolabeled algae by Corbicula. American Malacological Bulletin. 1986.

Maryland Department of Natural Resources. 2019. Maryland biological stream survey round four sampling manual.

McMahon, R. F. (1983). Ecology of an invasion pest bivalve, Corbicula. The mollusca, 6, 505-561.

Mynsberge, A. R., Strager, M. P., Strager, J. M., & Mazik, P. M. (2009). Developing predictive models for freshwater mussels (Mollusca: Unionidae) in the Appalachians: Limitations and directions for future research. *Ecoscience*, *16*(3), 387-398.

Phelps, H. L. (1994). The Asiatic clam (Corbicula fluminea) invasion and system-level ecological change in the Potomac River estuary near Washington, DC. Estuaries, 17(3), 614-621.

Sepkoski Jr, J. J., & Rex, M. A. (1974). Distribution of freshwater mussels: coastal rivers as biogeographic islands. *Systematic Biology*, *23*(2), 165-188.

Strayer, D. L. (1993). Macrohabitats of freshwater mussels (Bivalvia: Unionacea) in streams of the northern Atlantic Slope. *Journal of the North American Benthological Society*, *12*(3), 236-246.

Strayer, D. L. (2008). *Freshwater mussel ecology: a multifactor approach to distribution and abundance* (Vol. 1). Univ of California Press.

Southerland, M. T., G. M. Rogers, M. J. Kline, R. P. Morgan, D. M. Boward, P. F. Kazyak, R. J. Klauda, and S. A. Stranko. 2005. New biological indicators to better assess the July 2011 STREAM COMMUNITY THRESHOLDS 1677 condition of Maryland streams. DNR-12-03-05-0100. Maryland Department of Natural Resources, Monitoring and Non-tidal Assessment Division, Annapolis, Maryland, USA

United States Environmental Protection Agency. 2013. Revised deletion process for the site-specific recalculation procedure for aquatic life criteria. EPA-823-R-13-001.