AN ABSTRACT OF THE THESIS OF

Daniel Edwin Lockett IV for the degree of Master of Science in Marine Resource Management presented on March 27, 2012.
Title: A Bayesian Approach to Habitat Suitability Prediction

Abstract approved:

Chris Goldfinger Sarah K. Henkel

For the west coast of North America, from northern California to southern Washington, a habitat suitability prediction framework was developed to support wave energy device siting. Concern that wave energy devices may impact the seafloor and benthos has renewed research interest in the distribution of marine benthic invertebrates and factors influencing their distribution. A Bayesian belief network approach was employed for learning species-habitat associations for *Rhabdus rectius*, a tusk-shaped marine infaunal Mollusk. Environmental variables describing surficial geology and water depth were found to be most influential to the distribution of *R. rectius*. Water property variables, such as temperature and salinity, were less influential as distribution predictors. Species-habitat associations were used to predict habitat suitability probabilities for *R. rectius,* which were then mapped over an area of interest along the south-central Oregon coast. Habitat suitability prediction models tested well against data withheld for cross-validation supporting our conclusion that Bayesian learning extracts useful information available in very small, incomplete data sets and identifies which variables drive habitat suitability for *R. rectius*. Additionally, Bayesian belief networks are easily updated with new information, quantitative or qualitative, which provides a flexible mechanism for multiple scenario analyses. The prediction framework presented here is a practical tool informing marine spatial planning assessment through visualization of habitat suitability.

A Bayesian Approach to Habitat Suitability Prediction

by

Daniel Edwin Locket IV

A THESIS

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Master of Science

Presented March 27, 2012

Commencement June 2012

Master of Science thesis of Daniel Edwin Lockett IV presented on March 27, 2012

APPROVED:

Co-Major Professor, representing Marine Resource Management

Co-Major Professor, representing Marine Resource Management

Dean of the College of Earth, Ocean, and Atmospheric Sciences

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Daniel Edwin Lockett IV, Author

ACKNOWLEDGEMENTS

My experiences here at Oregon State University have been nothing short of positive and meaningful. I feel so fortunate to have been given such a great opportunity to succeed and I am most grateful to the following people:

Robert Allan - thank you for recognizing my potential. Your guidance along the way helped keep things in perspective.

Chris Goldfinger - thank you for enabling my success here by providing all the resources I could possibly need, especially stable funding, a comfortable workspace (except sea days), diverse experiences (including seas days), and the bird's-eye-view of the Willamette Valley. Cheers!

Sarah Henkel and Alix Gitelman, thank you for stepping up and providing guidance along the way. You devoted significant time in support of this project and my professional development – I couldn't have done it without you.

Lab partners - thank you for all your support, especially the laughs during the boat times. Special thanks to Chris Romsos and Morgan Erhardt for your countless hours of assistance. Your careful attention to detail and keen technical expertise improved the quality and efficiency of this research considerably, plus I learned a few things.

Family and friends, you were the support I needed through this long haul of life's ups and downs. Two years ago, when I thought nothing in life could be harder than it was, you were there to lift me up and carry me to the bright side.

Christina Rose Nese, my wonderful life-partner, you are the most sincere and cheerful spirit I will ever meet. I am so lucky you found me.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX FIGURES

In loving memory of Mary Jo Lockett

A Bayesian Approach to Habitat Suitability Prediction

Introduction

Wave Energy Development and the Benthic Environment

The wave climate, coastal infrastructure, and electrical power demand along the west coast of North America provide great potential for wave energy development in the region (Bedard et al., 2005). However, little is known about general direct or indirect impacts of wave energy developments. Impacts to migratory mammals and birds and fisheries habitat are of primary concern, and while these impacts are highly uncertain there is very little information about how wave energy developments impact benthic communities specifically.

Many processes of wave energy development potentially affect benthic communities through myriad stressors that vary in severity, intensity, and duration (Boehlert and Gill, 2010). Perhaps the most obvious stressors are alterations to the seafloor (Gill, 2005; Pelc, 2002), abrasion or mutilation of organisms (Abelson and Denny, 1997), and the artificial reef effect (Gill, 2005; Inger et al., 2009; Langhamer and Wilhelmsson, 2009), all of which may occur during various phases of development, maintenance and decommissioning.

Less obvious potential stressors are removal of kinetic energy or altering flow conditions (Commission, C. E., C. Ocean, and P. Council; 2008; Millar et al., 2007; Shields et al., 2011), acoustic effects and vibration, and electromagnetic effects, which are largely unknown (Boehlert and Gill, 2010). Altering flow conditions may cause benthic sediment scouring, interference with flow-dependent sessile, sedimentary or filter feeder species (Shields et al., 2011), and alter recruit settlement. Crab and lobster larvae have been shown to be attracted to reef noise (Montgomery et al., 2006) where acoustic effects generated by wave energy converters may interfere. Finally, electromagnetic effects may be detectable by species that have electroreception sensory capabilities (Gill, 2005). Electromagnetic fields associated with wave devices may attract, deter or injure aquatic animals (Cada et al., 2009)

Environmental effects of wave energy developments are highly uncertain, and research is needed to develop an understanding of potential impacts (Boehlert and Gill, 2010; Gill, 2005). The paucity of benthic impact estimates stems largely from the lack of seafloor mapping and benthic community surveys in areas of potential wave energy development, especially in the Pacific Northwest (Boehlert et al., 2008). Although considerable effort has been made to map the region's ocean floor (Goldfinger et al., 2012), most of the mapping has been in very nearshore waters that are less suitable for wave energy development. On a regional scale, in waters considered ripe for development, seafloor mapping and benthic surveys are significantly lacking.

Why worry about marine sedimentary biodiversity?

Marine ecosystems are incredibly diverse and the vast majority of species are invertebrates residing in (infauna) and on (epifauna) the sediments (Snelgrove, 1999). Of these invertebrates, most are polychaetes, crustaceans, mollusks and nematodes. Additionally, the seafloor sediments contain microbiotic species consisting of bacteria and protists which are poorly known. Snelgrove, (1997) estimates that less than 1% of worldwide marine species in sedimentary habitats are presently known. This is partly due to limitations of conventional seafloor survey methods (Wright and Heyman, 2008), logistics and effort (Snelgrove, 1999).

Although marine sedimentary ecosystems are poorly understood, there are significant ecological “products” benefiting life on earth (Snelgrove, 1999). The loss of marine sedimentary biodiversity would have cascading deleterious effects worldwide. Some “products” and services provided include: the reduction of atmospheric carbon dioxide through global carbon and geochemical cycling (Kristensen et al., 1992); a direct food source or prey for fisheries species humans consume (Reynolds, 2002); pollutant uptake and bioaccumulation (Farrington, 1991); and water clarity through filtration (Scheffer, 1999).

Naturally, humans have a tendency to protect the things we understand and therefore love (think of polar bears and whales). It is difficult to imagine Earth without the iconic polar bear, thus it is easier to be proactive in support of their conservation. It is also natural that humans tend to be less precautionary with regard to processes we do not fully understand and therefore find more difficult to love (think of worms and marine bacteria). This is true especially when the economic benefits of their protection are muted by the much more apparent benefits of alternatives to their protection (wave energy development). Conservation of marine sedimentary biodiversity and wave energy development are not necessarily mutually exclusive activities. As such, this study hopes to contribute to both understanding marine sedimentary diversity and distribution and the procession of a precautionary approach to management.

Species distribution modeling and predictive mapping

Species distribution models (*SDMs*) describe empirical relationships between species distributions and environmental variables thought to influence the ability of the environment to support a species (Franklin, 2009). *SDM*s are also referred to as habitat suitability models when describing the suitability of a habitat to support a species (Franklin, 2009). When used to predict in a geographical space, they have been called “predictive habitat distribution models” (Guisan and Zimmerman, 2000). In this study, we use the term *habitat suitability* as described by Franklin (2009). We also equate the terms presence/absence and true/false with habitat suitability probability (*HSP*).

Why model species distribution?

Advancements in remote sensing and geographic information systems and science (*GIS*) have accelerated interest in, and capacity to create, meaningful *SDM*s. We create *SDM*s to make sense of complex species-environment interactions and test hypotheses about species range characteristics. Additionally, governmental and non-governmental organizations are often charged with biological resources assessment and spatial planning decisions on regional scales. *SDM* products, such as habitat suitability maps, have been used to support various spatial planning and resource management decisions. Furthermore, there are usually strong financial incentives for using existing data to make predictions where surveying species is cost-prohibitive (Rushton et al., 2004), especially in the marine environment.

This document details research nested within a greater effort to survey the seafloor and benthos in areas of potential wave energy development (Boehlert et al., 2011). Utilizing species abundance and environmental data collected during the benthic survey, we developed a *predictive habitat suitability modeling framework* to support assessment of regional spatial planning alternatives, particularly within the context of marine renewable energy siting. We develop probabilistic Bayesian belief networks utilizing data mining techniques to define species-environment relationships warranted by the data.

Probability theory, uncertainty and graphical networks

Why bother with uncertainty?

Reasoning about any realistic domain requires leaving many facts unknown, unsaid, or crudely summarized (Pearl, 1988). There will always be exceptions to rules we develop in order to make sense of complex problems we wish to understand. We cannot afford to enumerate the exceptions and unambiguously define the conditions of our rules, so we summarize them (Pearl, 1988). Summarization is a “compromise between safety and speed of movement...in the minefield of judgment and belief” where we reason with exceptions (Pearl, 1988). Given inherent uncertainty in marine environmental data, especially of the benthos, probabilistic graphical networks are appropriate for gleaning the limited information within these data sets.

Probabilistic networks

Bayesian belief networks are graphical representations of relationships defined by our reasoning about the world. Networks encode fundamental, qualitative relationships of direct dependency as nodes in a graph connected by arcs or links (Pearl, 1988). Many patterns of human reasoning can only be explained by our inclination to follow pathways laid out by such networks (Pearl, 1988).

Belief networks play a central role in probability theory, where the aim is to provide a coherent account of how beliefs should change given partial or uncertain information (Pearl, 1988). Humans are bad at estimating quantities and therefore prefer to describe the world qualitatively. We therefore find probabilities useful because they are the “numerical summarization of uncertainty” (Pearl, 1988). The precise magnitude of our belief is less important than the specific structure of reasoning (i.e. the context and assumptions of the belief and the source of information that would change our belief) (Pearl, 1988). Thus, we are able to combine rough estimates in the manner we combine exact quantities thereby making the most of available information while constraining damage caused by imprecise estimates. Further, working with probabilities in a coherent model of reality prevents inconsistent conclusions which helps troubleshoot our inferences (Pearl, 1988).

Bayesian belief networks (*BBN*s), also called belief nets, causal networks (Duda et al., 2001) or influence diagrams (Pearl, 1988), are graphical representations of mathematical models where each variable is presented as a node (Uusitalo, 2007) that can take on two or more possible values. *BBN*s are directed acyclic graphs (Duda et al., 2001) where links between nodes describe the dependence or causal influences between variables. A link from node *A* to node *B* usually indicates that *A* causes *B*, that *A* partially causes *B* or predisposes *B*, that *B* is an imperfect observation of *A*, that *A* and *B* are functionally related, or that *A* and *B* are statistically correlated (Norsys Netica®). The strength of influence *A* on *B* is expressed by conditional probabilities (Pearl, 1988).

Conditional probability

McCarthy (2007) contains a very concise explanation of conditional probability theory and the following is modified from that explanation. Conditional probability theory can be expressed as follows, with *A* and *B* being outcomes that can be switched arbitrarily:

$P(A \text{ and } B) = P(B) \times P(A|B)$.

In English, this equation states that the probability of events *A* and *B* occurring together is equal to the probability of event *A* occurring “given the truth or occurrence” of event *B*, multiplied by the probability of event *B*.

For example, let’s say we are interested in the conditional probability that worms are present given an observation that mud is present. Let’s assume that worms are present in mud with a probability of 0.20 and in a particular study area, mud is present with a probability of 0.70. We can rearrange the expression above as follows,

$$P(worm|mud) = P(worm\ and\ mud) / P(mud) = 0.20 / 0.70 = 0.29,$$

and calculate the probability that worm will be observed in the study area.

Bayes rule is based on conditional probability (McCarthy, 2007; Pearl, 1988) and can be written in this example, as follows:

$$P(worm|mud) = P(worm) \times P(mud|worm) / P(mud).$$

Duda et al., (2001) express this informally in English by saying that

posterior probability = prior probability x likelihood / evidence

The *posterior* probability can be thought of as the ‘answer’ to our question of worm occurrence and therefore habitat suitability. The *prior* probability is a measure of our prior knowledge of worm occurrences. In other words, a *prior* distribution is a measure of how likely we are to observe worms before worms are actually observed. Priors are best defined by existing observational data sets. Although, priors can be defined by experts and, as a last resort, they can be set uniformly essentially expressing complete uncertainty. The method for determining priors has been hotly debated. The *likelihood* is the probability of observing mud given worms are present. The *evidence* is the probability of observing mud independent of worms.

Bayesian inference

Bayesian methods provide formalism, under the rules of probability theory, for reasoning about partial beliefs under conditions of uncertainty (Pearl, 1988). In other words, Bayesian methods and probability theory allow us to make inferences about the distribution of species and habitat using partial or uncertain information.

Inference allows for the possibility of answering questions such as “what is the chance that this ocean space is suitable habitat given a change in sediment composition?” or “how much can we expect habitat suitability to change given a rise in average temperature?” Because of the graphical nature and immediate propagation of results, the *BBN* approach allows visualization of ‘what-if’ scenarios with little effort. Examining a model’s response to perturbations is the beginning of exploring and expanding the domain knowledge base.

Methods

Data

Regional Sampling

Regional data were collected at six sites spanning the states of Washington, Oregon, and California along the west coast of North America (Figure 1). The southernmost site in the state of Oregon was arbitrarily chosen as a case study site for mapping surficial geologic habitat and predicting habitat suitability (Figure 2). Species density and environmental variable data were collected in each of the six sampling areas (Table 1). For a detailed description of the sampling methods for the benthic survey, see Boehlert et al. (2011) and references therein. Briefly, a total of 184 cores were taken using a 0.1 m^2 box corer and a CTD cast was conducted. Sampling stations for box coring were randomly selected using a Generalized Random Tessellation Stratified survey design. Upon landing the corer on the deck, a subsample of sediment was retained for later analyses. The remaining sample was sieved on 1 mm mesh. The invertebrates within the sieved sample were collected and preserved in formalin. Upon return to the lab, invertebrate specimens were identified and enumerated.

Additionally, shipek grab samples were collected such that unique habitats, as identified by multibeam acoustic data, were adequately sampled. Sediment grain size analyses were performed and used in conjunction with box corer sediment data to ensure that surficial geologic habitat classification was of the highest quality. Invertebrate data was not collected from Shipek samples. For more information on multibeam acoustic survey methods and surficial geologic habitat classification see Appendix A and Goldfinger et al. (2012).

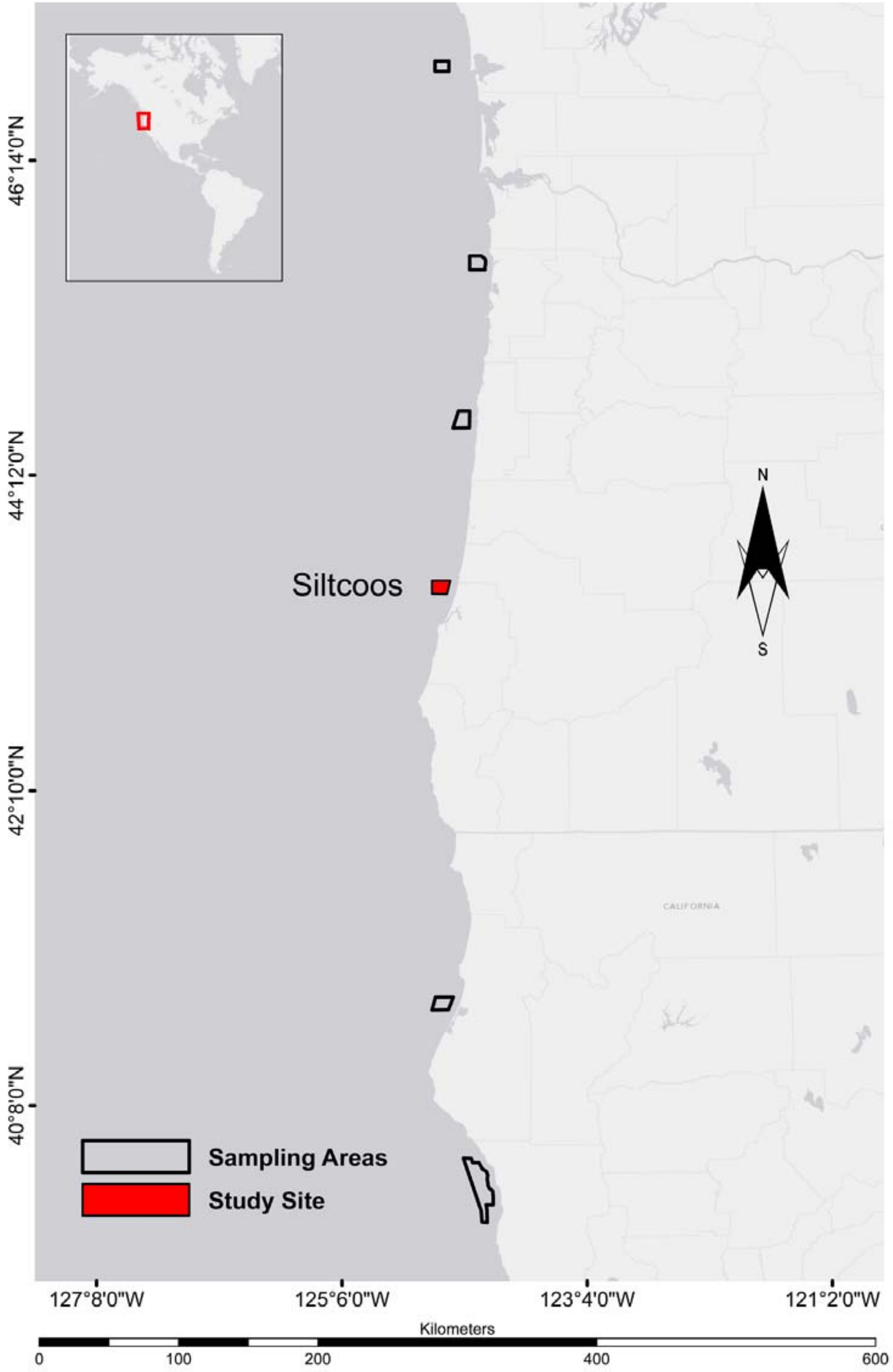


Figure 1. Sampling areas (N=6) over the region. Samples collected were used to create the species preference knowledge base. The Siltcoos, Oregon site was used as a case study for creating surficial geologic habitat and habitat suitability prediction maps.

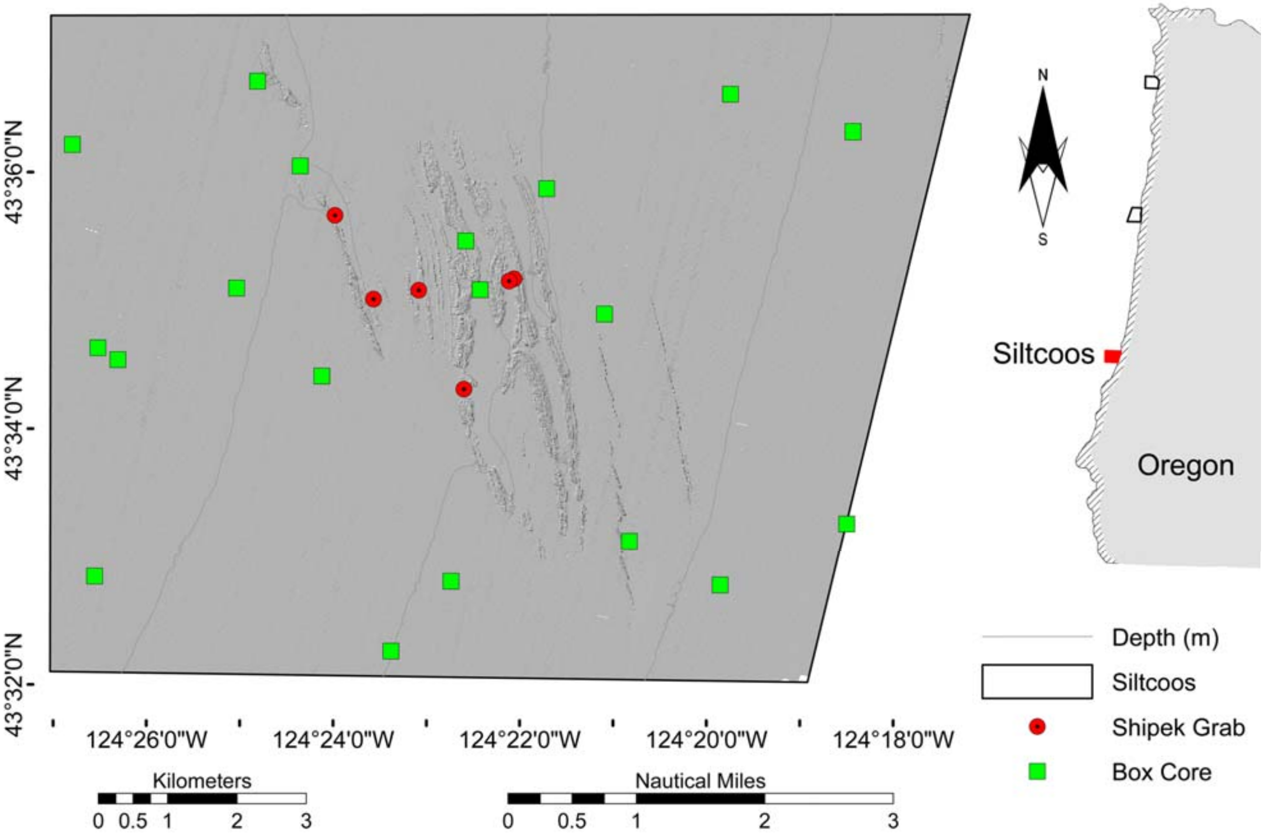
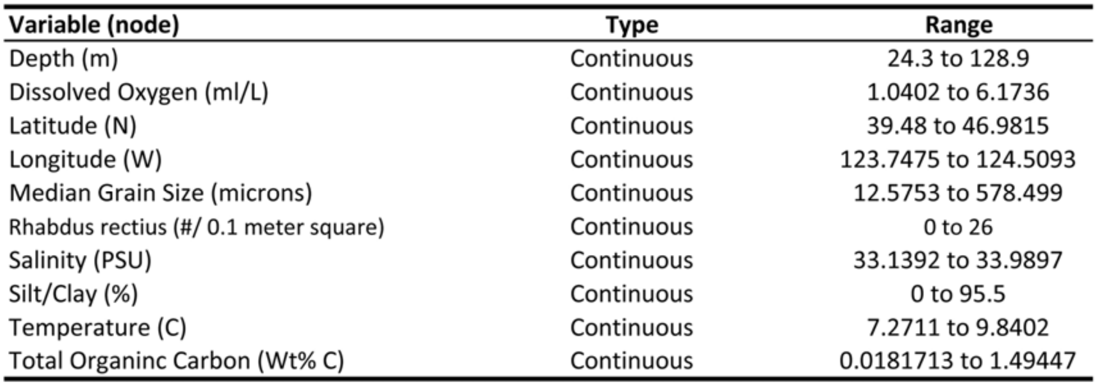


Figure 2. We used the southernmost site in the state of Oregon as a case study (115 km^2). The map shows sample stations within the case study site. Box core samples were processed for sediment grain size distribution and species density (1/10 m^2), CTD casts were taken at each box core station for water characteristics and used in model construction. Shipek grab samples were processed for sediment grain size distribution. All samples were used to create surficial geology habitat maps. The background image is a hill-shade relief raster (1m cell) with 10m depth contours (100m to 120m).



Variable (node)	Type	Range
Depth (m)	Continuous	24.3 to 128.9
Dissolved Oxygen (ml/L)	Continuous	1.0402 to 6.1736
Latitude (N)	Continuous	39.48 to 46.9815
Longitude (W)	Continuous	123.7475 to 124.5093
Median Grain Size (microns)	Continuous	12.5753 to 578.499
Rhabdus rectius (#/ 0.1 meter square)	Continuous	0 to 26
Salinity (PSU)	Continuous	33.1392 to 33.9897
Silt/Clay (%)	Continuous	0 to 95.5
Temperature (C)	Continuous	7.2711 to 9.8402
Total Organinc Carbon (Wt% C)	Continuous	0.0181713 to 1.49447

Table 1. Environmental variables used in the modeling process, the units of measurement, the data type, and observed domain range. *Rhabdus rectius* is the species of interest. Individuals were counted per box core sample.

Organism of interest

Rhabdus rectius was identified as one of the most prevalent mollusk species observed and selected for use as the target in the case study (Figure 3). *Rhabdus rectius* is a marine mollusk in the class Scaphopoda, which are identified by curved, open-ended tusks up to 6 cm in length (Reynolds, 2002). Their tusks taper from the wider head end to the narrower end where their respiratory currents exit. Scaphopods are carnivorous infaunal organisms that burrow headfirst into the sediment where they survive primarily on foraminiferans and other microorganisms (Reynolds, 2002). The global biogeographic patterns of Scaphopod diversity have been studied only preliminarily but, in general, they are found worldwide in marine sediments as deep as 6000 m with diversity decreasing with depth and toward polar regions (Reynolds, 2002).

Scaphopods seem to be an important component of their ecosystems. Ciliates and bacteria are associated with specific cell surfaces of *Rhabdus rectius* (Reynolds, 2002), in an assumed commensal relationship. Somewhat larger scaphopods are mutualistically associated with cnidarians such as anemones, corals and barnacles (Reynolds, 2002). Parasitism has been observed in *Rhabdus rectius* where flatworms replace the gonad (personal communication referenced in Reynolds, 2002). Additionally, scaphopods are preyed upon by ratfish and discarded shells provide refuge for a variety of spinunculans and hermit crabs (personal communication referenced in Reynolds, 2002).

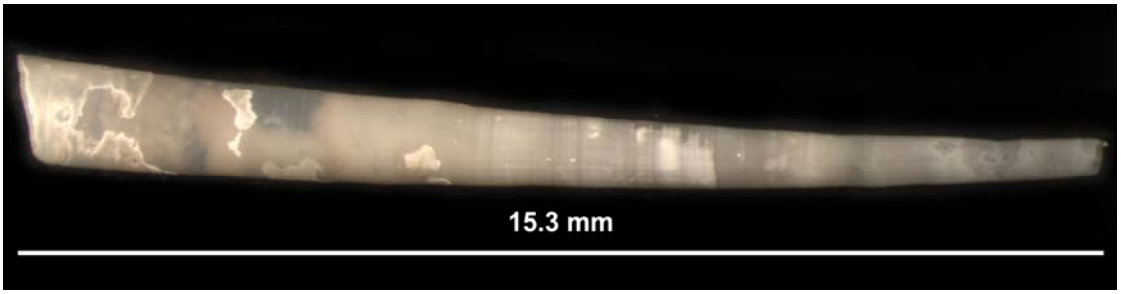


Figure 3. A *Rhabdus rectius* specimen featuring the tusk-like shell.

Model Construction

Conceptual framework

Creating an influence diagram is one of the first steps in the creation of a Bayesian Belief Network of the predictive modeling framework (Marcot et al., 2006). The influence diagram begins to address questions about which factors influence the target response variable, in this case, occurrence of benthic organisms. Guisan and Zimmerman (2000), stress that "an underlying *conceptual framework*" is crucial to "the formulation of an ecological model." Expert opinion and supplemental literature synthesis are commonly used to develop a conceptual diagram of boxes and arrows that illustrates the known or theoretical relationships between the target variable and predictor variables (Alameddine et al., 2011).

It is generally believed that distribution patterns of marine benthic fauna are determined largely by temperature, salinity, depth, surface productivity, and sediment dynamics over broad scales and by biological interactions, sediment geochemistry, and near-bed flow processes at finer scales (Snelgrove, 1999). However, the synergies between marine benthic invertebrates and the factors that influence their distribution are poorly understood. Initially, we assume that each variable exerts some influence on the target species and that each predictor variable is independent of the rest (i.e. there are no synergistic interactions between variables). However unlikely, this is a practical starting point for a modeling exercise such as this, where our first aim is to develop a framework of methods for predicting regional habitat suitability. Furthermore, including these synergies may be unnecessary for adequate predictions. As a result, we do not consider ecological interactions (such as food availability). Rather, we only consider physical environmental variables such as sediment and water column properties as factors influencing regional habitat suitability (Figure 4). The next step in the modeling framework is to convert the influence diagram into a belief network.

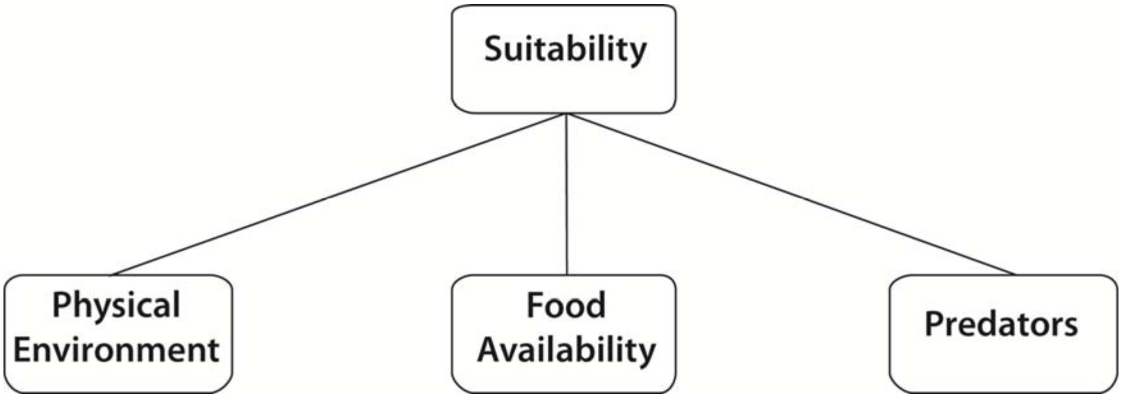


Figure 4. An influence diagram is one of the first steps in creating a BBN. On a regional scale, extremes in physical environmental factors, such as depth and temperature, are thought to limit species range and therefore the potential suitability of a given habitat to support species.

Basic model construction

Bayesian Belief Networks (*BBN*s) are directed, acyclic graphical representations of complex systems. Basic model construction follows a two-step process, each step briefly introduced here and described in detail below. The first step is model structure development and the second is model parameterization. Model structure is a graphical representation of the relationship between the target variable (response variable) and its environmental predictor variables. Model parameterization requires quantification of the relationships defined by the model structure. Parameterization (i.e. the population of conditional probability tables (*CPT*s)) occurs after accepting a model structure.

In the model structure, variables are represented by *nodes*, and the relationships between variables are defined by *links* (sometimes called arcs). Creating the *BBN* model structure can be done manually, automatically or using a hybridization of manual and automated methods. We use an automatic structure-generating technique in the Netica® software package. As noted in the previous section, we initially consider only direct correlation between predictor variables and the target variable (see naïve Bayesian networks below).

Data Partitioning

Before model structure development can begin, the data set was prepared for Netica® as a spreadsheet of appropriately named predictor variables, assigned special characters for missing data points, and rounded numerals. The data set was then partitioned to set the stage for later assessing prediction success.

Robust measures of prediction success make use of independent data (i.e. data not used to develop the prediction model) (Fielding and Bell, 1997). We used *K*-fold partitioning to independently cross-validate models. *K*-fold data partitioning methods are described well in Aguilera et al. (2010). The simplest method randomizes and partitions the original data set into training and testing subsets (*K*=2). Huberty (1994) provides a 75/25 percent ‘rule of thumb’ for calculating the training to testing ratio in presence/absence models. A more robust validation method is to divide the initial data set into *K* subsets (*K*>2) with each subset acting as the testing data once while the remaining sets are combined to act as the training subset. We randomly divided the original data (*N*=184) into four equal subsets (*K*=4). Each of the four subsets (*N*=46) acted as a testing partition while the remaining three subsets were combined (*N*=138) to act as the training set. Subsequently, four-fold cross-validation created four models to be used during model assessment.

To be clear, data partitioning is a model assessment technique only. None of the training/testing models should be used for prediction purposes (Fielding and Bell, 1997). The final prediction model was constructed with the entire data set to include all possible environments and facilitate model comparison, whereas the training models only consider a subset of the possible environments. The prepared and partitioned data set was imported to Netica® to begin model structuring.

Model structure

Variables and their domains: discretizing predictor variables

The first step of model structuring is discretization. Discretization, also referred to as binning, is the process of separating continuous variables, such as water depth and temperature, into manageable bins that cover the continuous variable space. Each discrete bin is a model state (i.e, a possible state of the world we are modeling). Netica® software requires discretization of continuous variables, which always results in a loss of statistical power (Uusitalo, 2007). Finding the optimal discretization method for continuous variables is an area of active research and remains one of the more challenging tasks associated with developing *BBN*s with continuous variables

(Almeddine et al., 2011). Thus, discretization optimization is beyond the purview of this study.

Unsupervised and supervised are the two basic discretization methods. First consider unsupervised discretization, where bins are created automatically by either an equal-interval or equal-frequency method. The equal-interval method divides the continuous data range, as defined by the data set, into equal bin sizes or widths. The equal-frequency method divides the data range such that equal data samples (or as close to equal as possible) fall within each state regardless of bin width.

Supervised discretization is the alternative to automatic binning. Supervised discretization occurs by manually placing bin cutoffs at expected environmental thresholds, standards or technological constraints. Assigning thresholds in this manner is critical to making a meaningful and useful model (Almeddine et al., 2011; Marcot et al., 2006). Marcot et al. (2006) develop a theoretical model where every aspect of the *BBN* is manually created in a supervised fashion. This is not the goal of the work presented here where we created a *BBN* automatically with supporting data.

Both discretization methods are sensitive to the number of continuous values in each bin. For example, data range outliers may cause sparsely populated bins in the equal-interval method (Almeddine et al., 2011; Liu, et al., 2002). Alternatively, many occurrences of a particular value may result in occurrences of that value in more than one bin (Liu et al., 2002). If possible, Lui et al. (2002) recommend performing data outlier detection and discretizing continuous variables so as to maintain a minimum of six cases per bin (Liu et al., 2002). Cursory examination revealed no obvious outliers nor were there any bins with fewer than six cases.

Netica® defaults suggest three to five bins per variable with twenty percent rounding. Fewer bins results in higher model accuracy but less precision and there is little effect on threshold values above twenty percent rounding. All variables (nodes) are continuous in this study and therefore were discretized to three states with zero percent rounding. This ensures that approximately equal numbers of samples fall within each node state. Again, the entire data set (every partition) is used during discretization, not individually partitioned data sets, which would result in incomparable model boundaries and wild model variance.

The next step in model structuring is defining the causal linkages. The links between variables are essential for reasoning plausibly about a domain (Pearl, 1988). This process is also completed manually or automatically.

Causal interpretation of structure

Explaining the directed links between variables can be a major challenge faced when using Bayesian networks (Sebastiani and Perls, 2008). Directed links from effects to causes elicit comments that the arcs should be “the other way around” regardless that they only represent a convenient factorization of the joint probability of the network variables (Sebastiani and Perls, 2008). Netica® suggests that to classify, predict or diagnose a target variable with the best accuracy, it is best to capture its relation to as many predictor variables as possible with many links leaving the target variable. This model structure is called a naïve Bayesian network (*NB*) (see naïve Bayesian networks below). Netica® further posits that while the relationships have been captured, they are only considered in isolation and any synergies between variables have been ignored. This research employs the *NB* structure simply as a starting point for the development of the modeling framework.

Naïve Bayesian networks

Figure 5 illustrates that the naïve Bayesian network is a simple network linking the target variable to all the environmental variables (Duda et al., 2001). *NB* networks strongly assume conditional independence between environmental variables given the state of the target variable. Conditional independence assumptions ignore synergistic relationships between variables, such as species interactions or how water temperature varies with latitude and water depth.

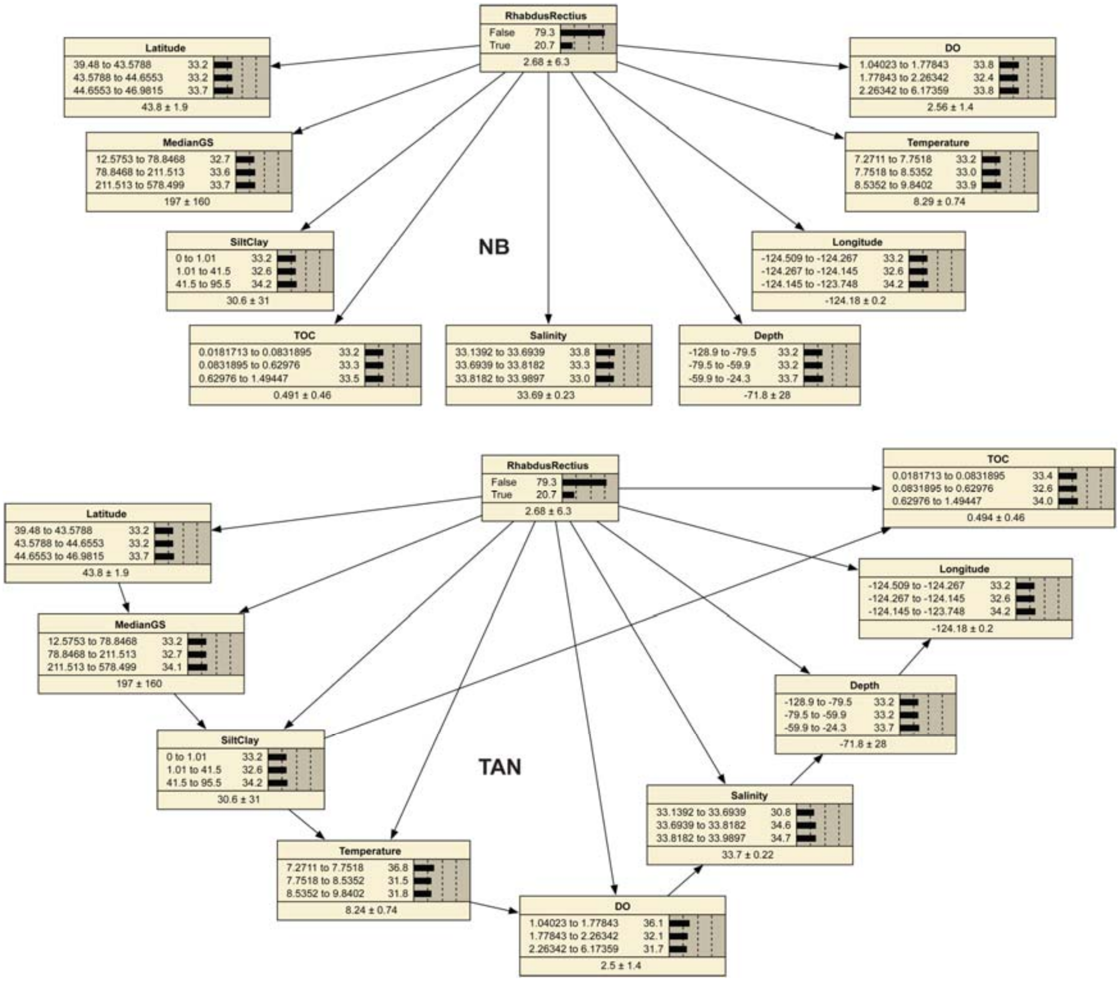


Figure 5. NB and TAN belief network models for *Rhabdus rectius*. Each node represents an environmental variable of the original data set. Continuous environmental variables were discretized to three states. The target node, *Rhabdus rectius*, was discretized to true/false states.

We performed sensitivity analyses (refer to sensitivity analysis section) to determine the most influential variables. We reversed the links of the strongest two predictor variables so that they enter the target variable. In essence, this considers the synergies between the most influential variables of the target node, and for the rest of the variables, considers them, but not all the synergies between them.

Tree-augmented naïve Bayesian networks (TAN).

Structure learning techniques attempt to remove the strong assumptions of independence in *NB* models by finding correlations among attributes that are warranted by the data (Friedman et al., 1997) and improve modeling accuracy (Aquilera et al., 2010). One such technique is the tree-augmented naïve Bayes network (*TAN*) structure (Figure 5). Netica® uses its sensitivity to findings function to automatically induce the *TAN* structure (personal communication). The target variable has no parents and each environmental attribute has as parents the target variable and at most one other environmental attribute (Friedman et al., 1997).

Model induction is an area of active research, an especially appealing method when little is known about the dependencies between variables in a complex system. However, structure learning algorithms have been shown to miss dependencies (Almeddine et al., 2011). Given that structured learning algorithms are not perfect and environmental data are inherently noisy, domain expert supervision of the structured learning process is more appropriate (Almeddine et al., 2011; Marcot et al., 2006).

TAN models were discretized in exactly the same manner as *NB* models to facilitate comparison of modeling techniques and include all possible environments. Model training and validation techniques are described in the following section.

Eliciting model parameters

Parameterization

Counting learning is the simplest true Bayesian parameter learning algorithm offered in the Netica® software package. Parameter learning proceeds from a state of ignorance at each node (i.e all *CPT*s have uniform probabilities) unless there are data that support prior information. The counting algorithm only updates the beliefs of nodes with cases that supply values for itself and its parents, thus counting learning is not recommended when there are latent variables or missing data. We do not intend to describe latent variables and our data set contains very few missing values therefore counting learning is acceptable. However, the original data set is very small with less than 200 individual cases and any cases skipped during the learning process would further reduce the data available for learning.

Importantly, Netica® algorithms respect expert defined model structure and parameters before learning commences. It is therefore possible to predefine relationships between variables and their parameters, then update the parameters as new data become available (see future work section below). This applies to all of Netica's® learning capabilities.

Handling Missing Data

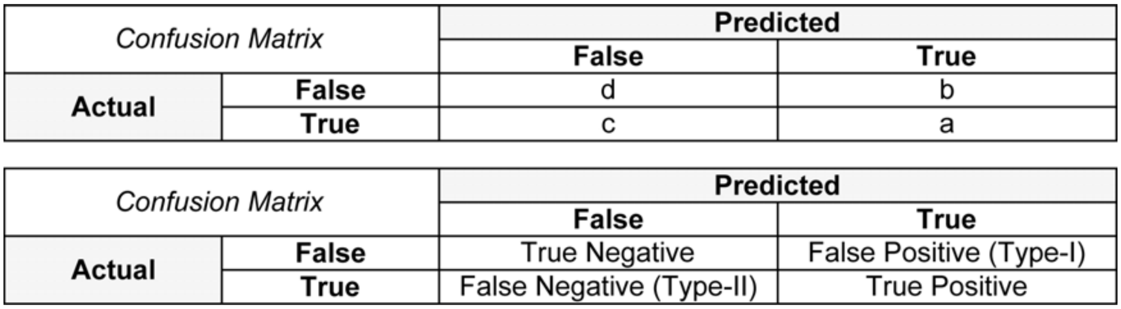
To account for missing data, we adopted the two-step, expectation-maximization (*EM*) learning algorithm, which begins when the expectation (*E*) step first computes the expected value for missing data. The maximization (*M*) step then maximizes a particular function for net parameters given the original data set and expected values.

Model validation and performance assessment

Each training model was validated with the complimentary testing subset described in the data partitioning section. We tested model prediction error through the computation of confusion matrices. We tested the target variable’s sensitivity to findings at predictor nodes with the Netica® software.

Confusion matrices

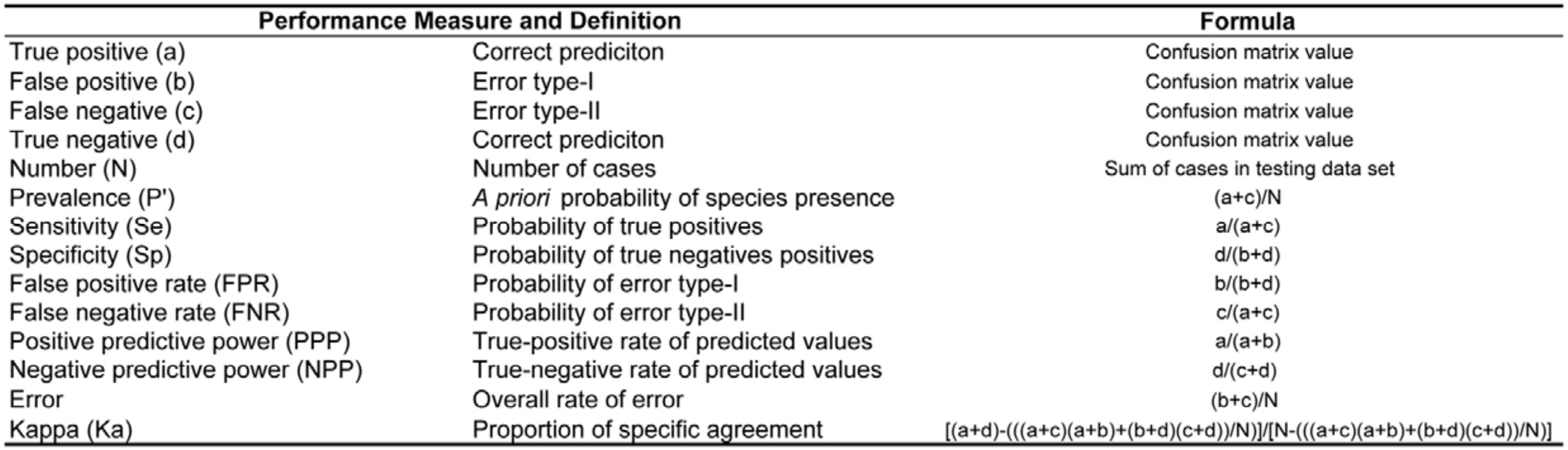
The performance of a presence/absence model is normally summarized in a confusion matrix. We compared observations of species presence/absence to model predictions of presence/absence and translate them to a confusion matrix of true/false outcomes. Species absence indicates a *false habitat suitability* whereas species presence indicates *true habitat suitability* (Table 2). Confusion matrix values were calculated following methods detailed in Fielding and Bell (1997) and used as initial performance measures (Table 3).



Confusion Matrix		**Predicted**	
		False	**True**
Actual	**False**	d	b
	True	c	a

Confusion Matrix		**Predicted**	
		False	**True**
Actual	**False**	True Negative	False Positive (Type-I)
	True	False Negative (Type-II)	True Positive

Table 2. Theoretical confusion matrix for two-state output models.



Performance Measure and Definition		Formula
True positive (a)	Correct prediciton	Confusion matrix value
False positive (b)	Error type-I	Confusion matrix value
False negative (c)	Error type-II	Confusion matrix value
True negative (d)	Correct prediciton	Confusion matrix value
Number (N)	Number of cases	Sum of cases in testing data set
Prevalence (P')	*A priori* probability of species presence	(a+c)/N
Sensitivity (Se)	Probability of true positives	a/(a+c)
Specificity (Sp)	Probability of true negatives positives	d/(b+d)
False positive rate (FPR)	Probability of error type-I	b/(b+d)
False negative rate (FNR)	Probability of error type-II	c/(a+c)
Positive predictive power (PPP)	True-positive rate of predicted values	a/(a+b)
Negative predictive power (NPP)	True-negative rate of predicted values	d/(c+d)
Error	Overall rate of error	(b+c)/N
Kappa (Ka)	Proportion of specific agreement	[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/N)]/[N-(((a+c)(a+b)+(b+d)(c+d))/N)]

Table 3. Performance measures calculated according to Fielding and Bell (1997).

Presence/absence models are normally judged by the number of prediction errors they create. In testing presence/absence models there are four possible outcomes, of which two types are erroneous: *false-positives* and *false-negatives* sometimes referred to as error type-I (commission) or error type-II (omission), respectively. A *false-positive* error occurs when the model incorrectly predicts presence when the observation actually was absence. Incorrectly predicting absence when the observation actually was presence is called a *false-negative* error. If both prediction and observation are true it is called a *true-positive*. The last possible outcome is a correct prediction of a false observation, which is called a *true-negative*. The overall rate of error is the primary performance assessment metric in *SDM*s but other metrics can improve understanding of the value of its predictions.

Sensitivity (*Se*) is the probability that the algorithm correctly classified the case as true (true-positive rate of actual values). Specificity (*Sp*) is the probability that the algorithm correctly predicted the case as negative (true-negative rate of actual values). The false-positive rate (*FPR*) is the rate of false-positive errors and the false-negative rate (*FNR*) is the rate of false-negative errors. Positive predictive power (*PPP*) is the probability that a case is true when it is predicted true (true-positive rate of predicted values). Negative predictive power (*NPP*) is the probability that a case is false when it is predicted false (true-positive of predicted values). Prevalence (*P'*) is the number of presences divided by the number of cases. In other words, *P'* is the *a priori* information for species occurrence as determined by the original data set, and as such, shall not change between models. Kappa (*Ka*) uses all of the values in the confusion matrix and is a good single metric for assessing classifier improvement over chance (Fielding and Bell, 1997; Forbes, 1995). The *Ka* statistic measures the proportion of all possible cases of presence or absence that are predicted correctly by the model (Manel et al., 2001). *Ka* statistic values range from [0-1] with the following quality of agreement: poor

Ka < 0.4; good 0.4 < *Ka* < 0.75 and excellent *Ka* > 0.75 (Fielding and Bell, 1997). Landis and Koch (1977) offer an alternative definition for quality of agreement: slight to fair *Ka* < 0.4; moderate 0.4 < *Ka* < 0.6; substantial 0.6 < *Ka* < 0.8 and almost perfect *Ka* > 0.8.

Error rate, predictive power measures and *Ka* give some indication of model performance but should be used with caution because they are influenced by *P'*, especially error rate. For example, it is possible to achieve a 95% correct classification rate when *P'* is 5% by classifying all cases as negative (Fielding and Bell, 1997). Therefore, a model's predictive value should not hinge on error rate alone. Error rate, *PPP*, and *Ka* decrease with decreasing *P',* while *NPP* increases. *Ka* is less affected by low prevalence but may fail when one class far exceeds the other (Fielding and Bell, 1997).

Sensitivity Analysis

Sensitivity analyses were conducted using a custom report generated with Netica® then ranked by mean percent variance reduction. Variance reduction is a measure of the sensitivity of the target variable to a finding at an environmental variable. We used the sensitivity analyses to justify alterations to link directionality in the *NB* prediction model. Regardless of sensitivity, the link between *MedianGS* was necessarily reversed to prevent violation of the acyclicity of *BBN*s.

Influence of temperature

For the *NB* model only, a separate sensitivity test was performed using hypothetical inputs for temperature. This was essentially a 'what if' scenario where the goal was to test the effects of temperature on model outputs by running the model two times; once with temperatures always in the highest state, then again with the temperatures always in the lowest state, while all other inputs remain unchanged. The posterior habitat suitability probabilities were mapped in the GIS for each temperature run and a ‘difference’ raster was created by using map algebra functions in the GIS.

Receiver Operating Characteristic Curve

In Figure 6 are examples of receiver operating characteristic (*ROC*) curves, which are threshold-independent measures of model performance (Deleo, 1993). The *ROC* curve is obtained by plotting all sensitivity values (true-positive fraction) against (1-specificity) values (false-positive fraction) at each available cutoff. High performing models will have a curve that approaches the top left corner. The *AUC* is a single measure of overall accuracy that is not dependent upon a particular threshold. Area under the function (*AUC*) was calculated using the trapezoid method (Table 4).

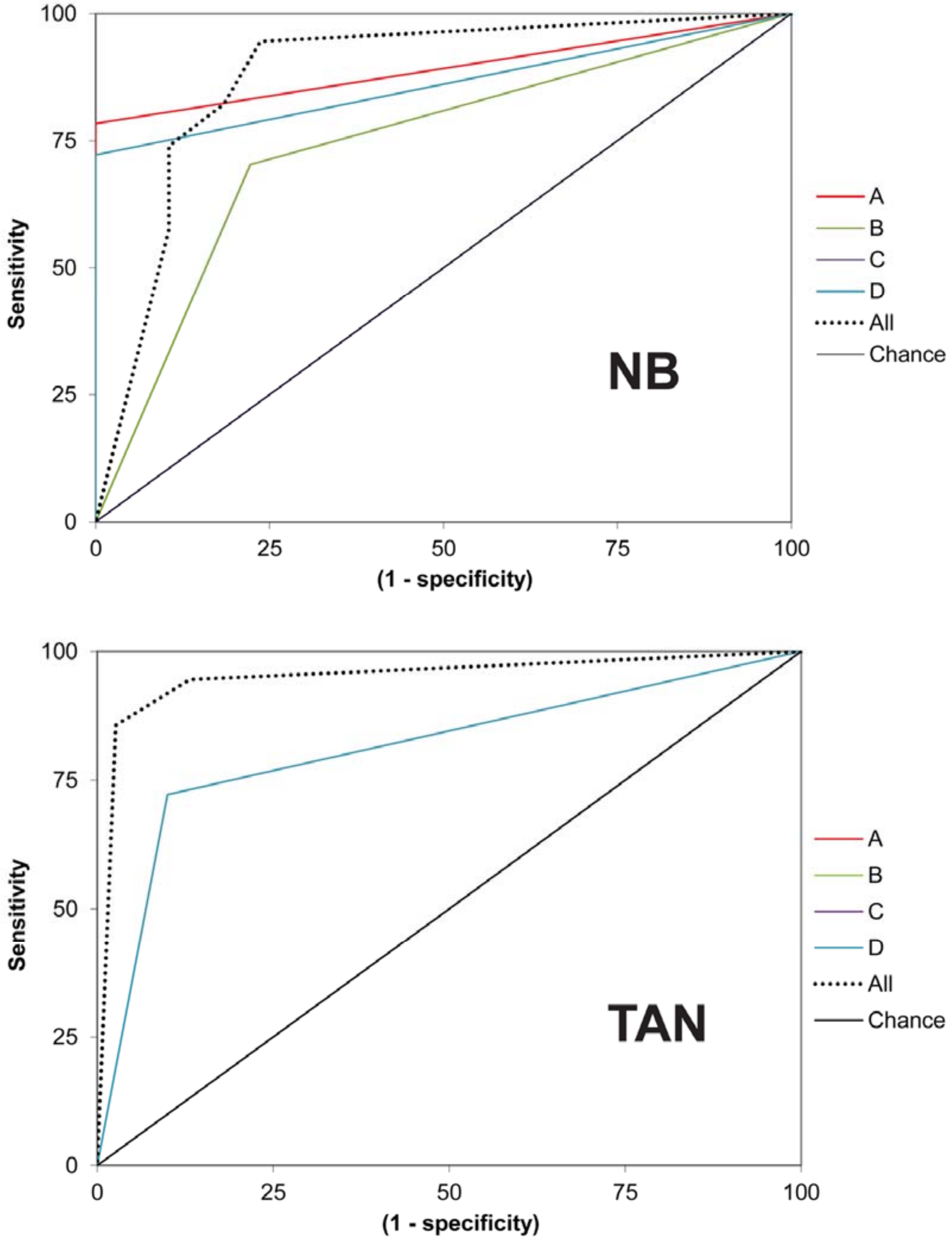
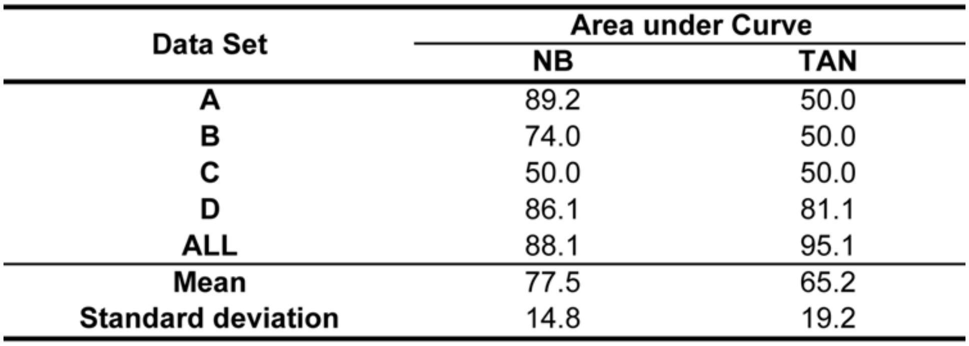


Figure 6. ROC curves for each training-testing NB and TAN model.



Data Set	Area under Curve	
	NB	TAN
A	89.2	50.0
B	74.0	50.0
C	50.0	50.0
D	86.1	81.1
ALL	88.1	95.1
Mean	77.5	65.2
Standard deviation	14.8	19.2

Table 4. Area under the curve (AUC) was calculated for each model using the trapezoid rule. Results were averaged to determine a plausible performance metric for the prediction model.

Mapping Predictions

Incorporating case files

The paucity of relevant data can make predicting habitat suitability challenging in the marine environment. *BBN*s allow us to make probabilistic inferences about environmental variables based on learned relationships warranted by observational data. Thus, it is possible to make accurate predictions with limited information.

We compiled in a case file of rows and columns (cases and variables) of available information for environmental predictor variables. (Appendix A). Each case file row represented a location with an associated water depth and surficial geologic habitat (*SGH*) class, a surrogate for median grain size (*MGS*) (the model node for SGH is called *Gridcode*). The case files were incorporated as positive findings in both *NB* and *TAN* model structures. Posterior probabilities were inferred for remaining variables, including the target variable. Posterior probabilities were appended to the original case file and imported into a geographic information system (*GIS*) for further analyses.

Habitat suitability raster analyses

Each model output is essentially a geographic coordinate grid of environmental variables with an additional posterior probability attribute attached to every location. Every output has the same spatial extent, which facilitates comparison between models when mapped in a *GIS*. Output grids were converted to eight-meter resolution rasters and compared.

Confidence raster analyses

During a model run, each node carries with it a measure of error in the form of one standard deviation around the expected value of that node, including the target variable. Therefore, for each instance in a case file, the calculated posterior probability for *Rhabdus rectius* carries with it and measure of confidence. This measure is a final accounting of error propagation throughout the network. A confidence raster for the TAN model was created for assessment of our confidence in each posterior probability value.

Results

Model sensitivity

Variance reduction

The first step of model assessment is to test the target sensitivity to findings at each node (Table 5). Results indicate the rank order of influence on the target did not change between models while percent variance reduction changes were negligible. *Longitude* is the strongest predictor variable responsible for, on average, a 31% ± 4.9% variance reduction in both models. Not surprisingly, sediment properties rank among the most influential variables. Each variable describing sediment characteristics is responsible for at least 20% reduction of the variance for a respective finding entered at that variable, whereas variables describing water column properties rank the lowest at <10.2% variance reduction with *temperature* very low at 1.5%.

NB

	A	B	C	D	Mean	Standard Deviation
RhabdusRectius	100	100	100	100	100	0
Longitude	29.0	28.1	39.4	27.4	31.0	4.9
TOC	19.1	30.0	34.4	12.7	24.1	8.6
SiltClay	19.0	27.7	30.5	17.4	23.7	5.6
MedianGS	17.4	24.3	25.6	15.5	20.7	4.3
Depth	14.5	19.0	26.2	9.0	17.2	6.3
Latitude	9.0	15.0	12.1	9.1	11.3	2.5
DO	8.5	12.3	8.9	10.4	10.0	1.5
Salinity	8.1	11.2	11.5	6.3	9.3	2.2
Temperature	0.7	4.2	2.5	3.4	2.7	1.3

TAN

	A	B	C	D	Mean	Standard Deviation
RhabdusRectius	100	100	100	100	100	0
Longitude	29	28.1	39.4	27.4	31.0	4.9
TOC	19.1	29.3	34.9	13.7	24.3	8.3
SiltClay	19	27.7	30.5	17.4	23.7	5.6
MedianGS	16.8	22.9	25.3	15.4	20.1	4.1
Depth	14.5	19	26.2	8.97	17.2	6.3
Latitude	8.95	15	12.1	9.05	11.3	2.5
DO	8.82	11.8	9.88	10.1	10.2	1.1
Salinity	6.3	9.03	9.72	4.48	7.4	2.1
Temperature	0.155	2.55	1.78	1.67	1.5	0.9

Table 5. Sensitivities were computed with the Netica® software sensitivity function using training-testing partitions. Values indicate the percent variance reduction at the target node as a result of a finding entered at a particular environmental variable node. Then the sensitivities were ranked by the mean value and used to identify which linkages to reverse in the NB model.

Influence of Temperature

The temperature sensitivity test revealed that by simply adding temperature as input, *HSP* values increased slightly (Table 6 and Figure 7). On the low end, increases ranged between approximately 0.03 and 0.08, and on the high end, increases ranged between approximately 0.01 and 0.02. Additionally, both temperature runs decreased the possible range of *HSP* values approximately 0.02 to 0.06. The low temperature model run had the greatest influence on overall *HSP*.

Habitat Suitability		***Temperature***	
Value	**No Temp**	**High**	**Low**
High	0.93	0.94	0.95
Low	0.37	0.40	0.45
Difference	0.56	0.54	0.50

Table 6. This table compares the ranges of posterior probabilities for the standard NB model (No Temp) with the addition of *Temperature* variable pegged at each extreme.

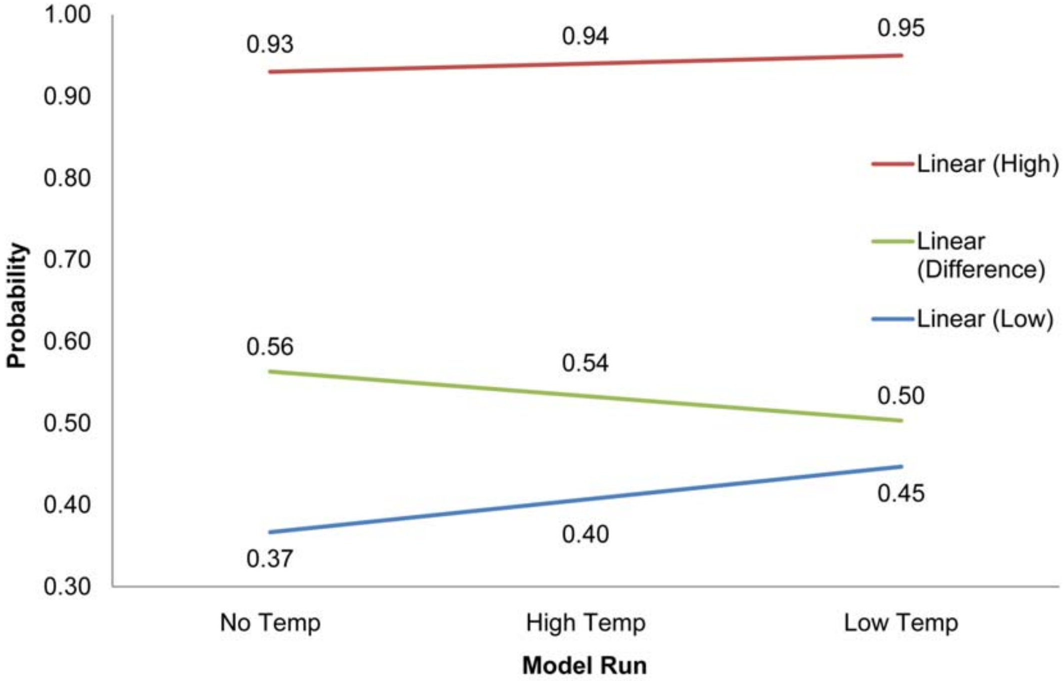


Figure 7. This figure demonstrates how the addition of temperature increased overall HSP while at the same time decreased the range of possible posterior probabilities.

Model construction

Figure 8 and Figure 9 both are three-state discretized models with the following modification. In each model, the link from *MedianGS* to the target was reversed to facilitate acceptance of the *Gridcode* variable (i.e. to facilitate the population of the *CPT*). For both models, sensitivity analyses identified *Longitude* and *TOC* as the two variables with the greatest influence on the distribution of *Rhabdus rectius*. In the *NB* model, the links from *Longitude* and *TOC* to the target were reversed to account for the possible synergistic relationships between them and the target variable. Doing so caused automatic insertion of links from *MedianGS* to *TOC* and *Longitude* and from *Longitude* to *TOC*. For the *TAN* model, reversing the links of the most influential variables was unnecessary. However, the link between *MedianGS* and *Latitude* was necessarily reversed to prevent cyclicity, which violates rules of *BBN* structure.

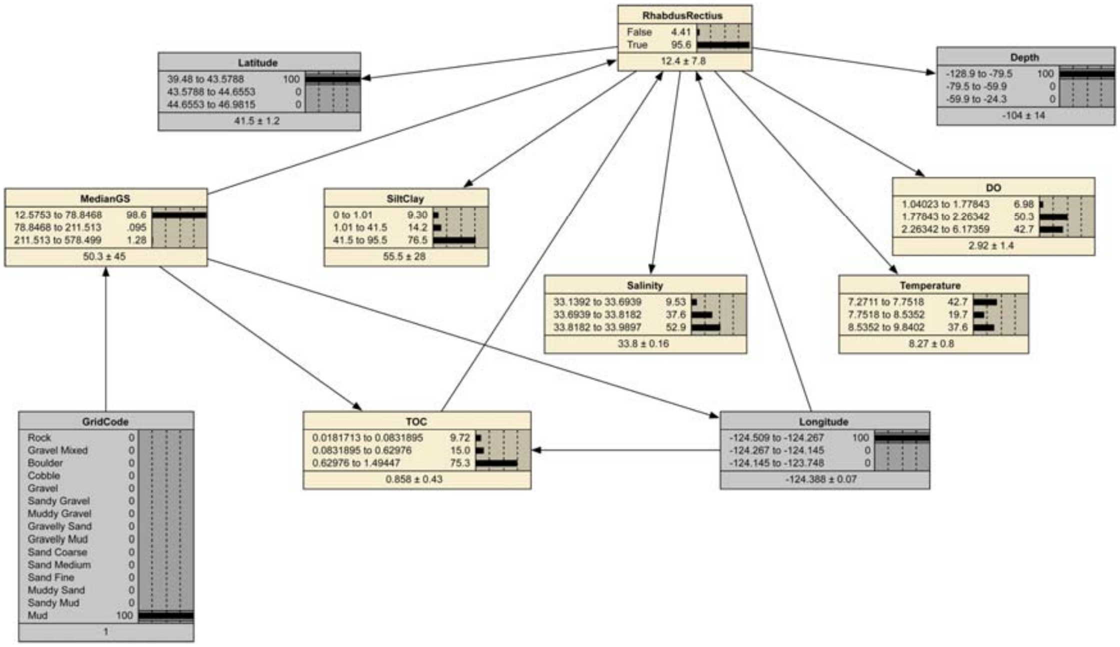


Figure 8. In this NB prediction model, gray nodes indicate a finding has been entered for a particular map cell in the study area. For every cell, the model output is a probability distribution for the species node, *Rhabdus rectius*. This particular set of findings suggests a very high probability of habitat suitability.

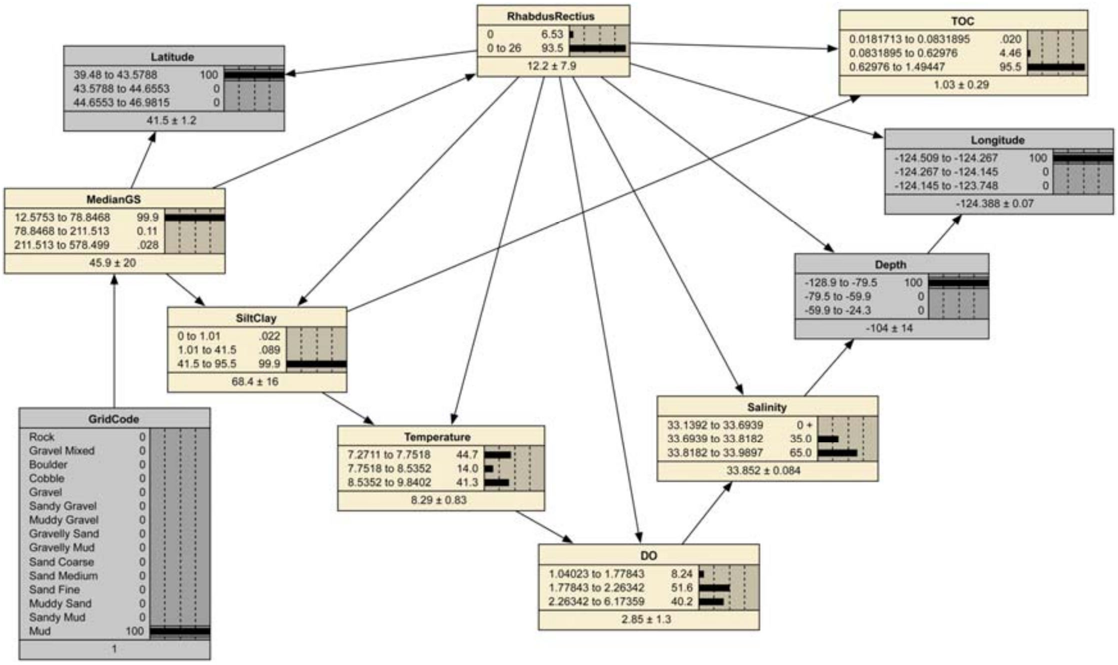


Figure 9. In this TAN prediction model, gray nodes indicate a finding has been entered for a particular map cell in the study area. For every cell, the model output is a probability distribution for the species node, *Rhabdus rectius*. This particular set of findings suggests a very high probability of habitat suitability.

Model validation and performance assessment

Confusion matrices

The results of the confusion matrix calculations are summarized (Table. 7). As expected, species prevalence (*P'*) did not differ between models. Mean error rates were quite low for both the *NB* model (0.14 ± 0.08) and *TAN* model (0.08 ± 0.03) with considerable overlap (Figure 10). Mean true-positive rates *(Se)* and variation decreased from 0.84 ± 0.15 in the *NB* model to 0.79 ± 0.12 in the *TAN* model. Mean true-negative rates *(Sp)* increased from 0.86 in the *NB* model to 0.95 in the *TAN* model, while the variance decreased from ± 0.09 to ± 0.01. *FPR* and *FNR* are inversely related to *Se* and *Sp* and as such, behaved accordingly. Ka statistics increased to 0.75 ± 0.10 in the *TAN* model from 0.63 ± 0.17 in the *NB* model.

NB Model

Species	Data Set	a	b	c	d	N	P'	Se	Sp	FPR	FNR	PPP	NPP	Error	Ka
Rhabdus rectius	A	8	4	1	33	46	0.20	0.89	0.89	0.11	0.11	0.67	0.97	0.11	0.69
	B	8	11	1	26	46	0.20	0.89	0.70	0.30	0.11	0.42	0.96	0.26	0.42
	C	6	3	4	33	46	0.22	0.60	0.92	0.08	0.40	0.67	0.89	0.15	0.54
	D	10	2	0	34	46	0.22	1.00	0.94	0.06	0.00	0.83	1.00	0.04	0.88
Mean Values		8	5	2	32	--	0.21	0.84	0.86	0.14	0.16	0.65	0.96	0.14	0.63
Standard Deviation		1	4	2	3	--	0.01	0.15	0.09	0.09	0.15	0.15	0.04	0.08	0.17

TAN Model

Species	Data Set	a	b	c	d	N	P'	Se	Sp	FPR	FNR	PPP	NPP	Error	Ka
Rhabdus rectius	A	8	1	1	36	46	0.20	0.89	0.97	0.03	0.11	0.89	0.97	0.04	0.86
	B	8	2	1	35	46	0.20	0.89	0.95	0.05	0.11	0.80	0.97	0.07	0.80
	C	6	2	4	34	46	0.22	0.60	0.94	0.06	0.40	0.75	0.89	0.13	0.59
	D	8	2	2	34	46	0.22	0.80	0.94	0.06	0.20	0.80	0.94	0.09	0.74
Mean Values		8	2	2	35	--	0.21	0.79	0.95	0.05	0.21	0.81	0.95	0.08	0.75
Standard Deviation		1	0	1	1	--	0.01	0.12	0.01	0.01	0.12	0.05	0.03	0.03	0.10

Table 7. The table shows values for each metric calculated with the confusion matrix.

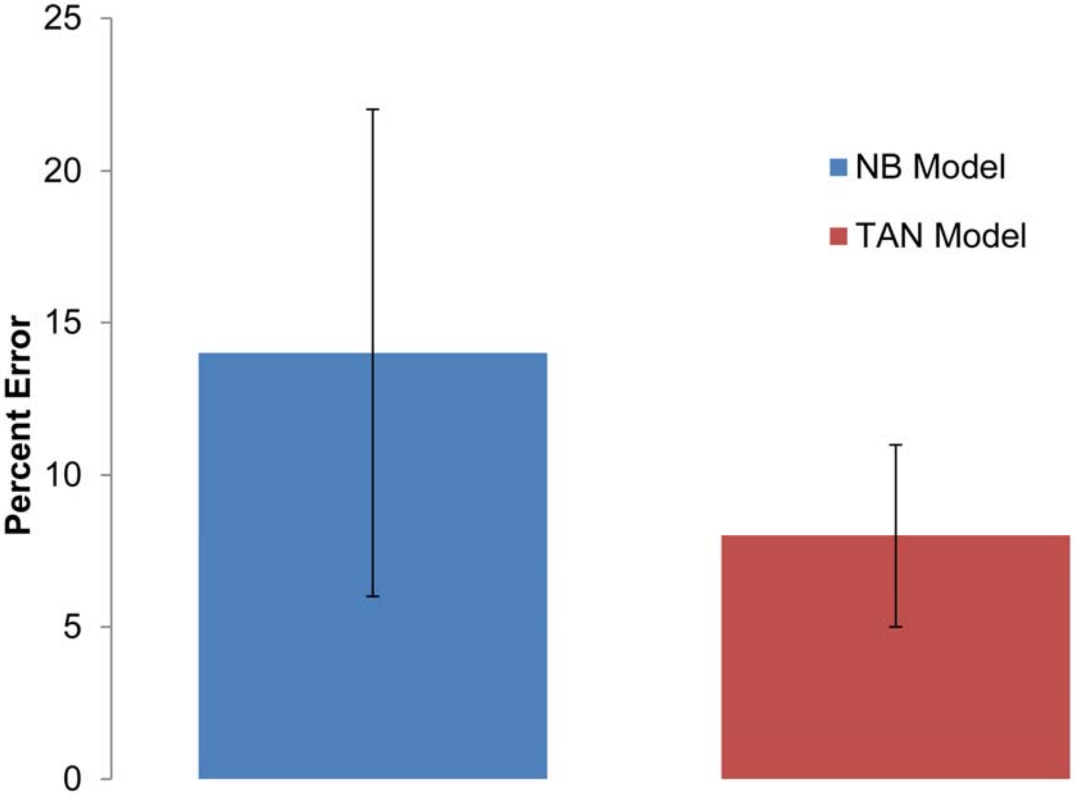


Figure 10. Mean error rates and standard deviations are compared. Mean error rate is 14% and standard deviation 8% for the NB model. Mean error rate is 8% with a standard deviation of 3% for the TAN model.

Receiver Operating Characteristic Curve

For both model types and each training-testing set, *ROC* curves were plotted (Figure 6) and the respective *AUC* calculated (Table 4). *NB* model *C* and *TAN* models *A*, *B* and *C* all follow the *Chance* diagonal with *AUC* values of 50.0. In both *ROC* plots, the *All* curve represents the prediction model, which used the entire data set for training instead of partitions. The *AUC* for the *NB All* data model produced was 88.1 while the *TAN All* data model produced an *AUC* of 95.1.

Spatial application of habitat suitability probability

NB versus TAN models

The *NB* model produced posterior probabilities for the target, *Rhabdus rectius*, which ranged from approximately 0.37 to 0.93 (Figure 11). The highest probabilities were observed in southern habitats of finer grain size while the lowest probabilities were found in association with rocks. There was a marked increase in *HSP* with decreasing latitude over the study area. In the south, the range of *HSP* was from approximately 0.83 to 0.93 (range = 0.10) and in the north that spread was approximately 0.37 to 0.83 (range = 0.46). Thus, surficial geology appears to have a greater influence on the posterior probabilities in the northern latitude state.

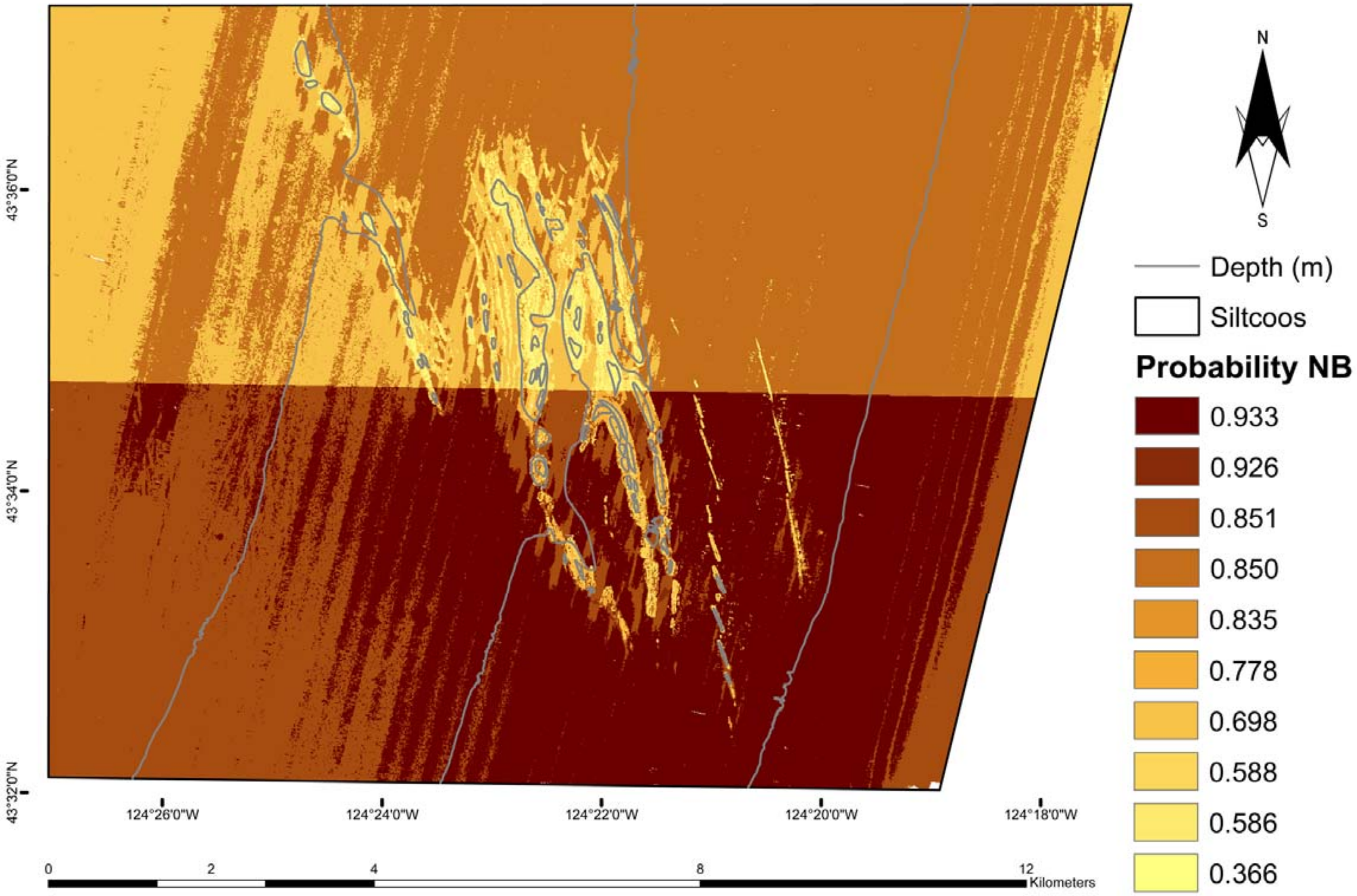


Figure 11. The habitat suitability map spatially depicts posterior probabilities for the NB model. Probabilities range from approximately 0.37 to 0.93 with darker colors representing higher probabilities.

The *TAN* model produced posterior probabilities for the target, *Rhabdus rectius*, which ranged from 0.00 to approximately 0.97 (Figure 12). Again, a latitudinal influence is marked by a horizontal delineation in the map center. In general, *HSP*s were very high except for areas known to contain rocks. The *TAN* model expressed a very slight decrease (approximately 0.03) in *HSP* toward the southern portion when the sediment grain size was finest, while with *HSP* increased slightly (approximately 0.06) toward the southern portion when sediment grain size was coarser. Otherwise, there was a similar pattern of weak influence by grain size in the southern portion of the study area as with the *NB* model.

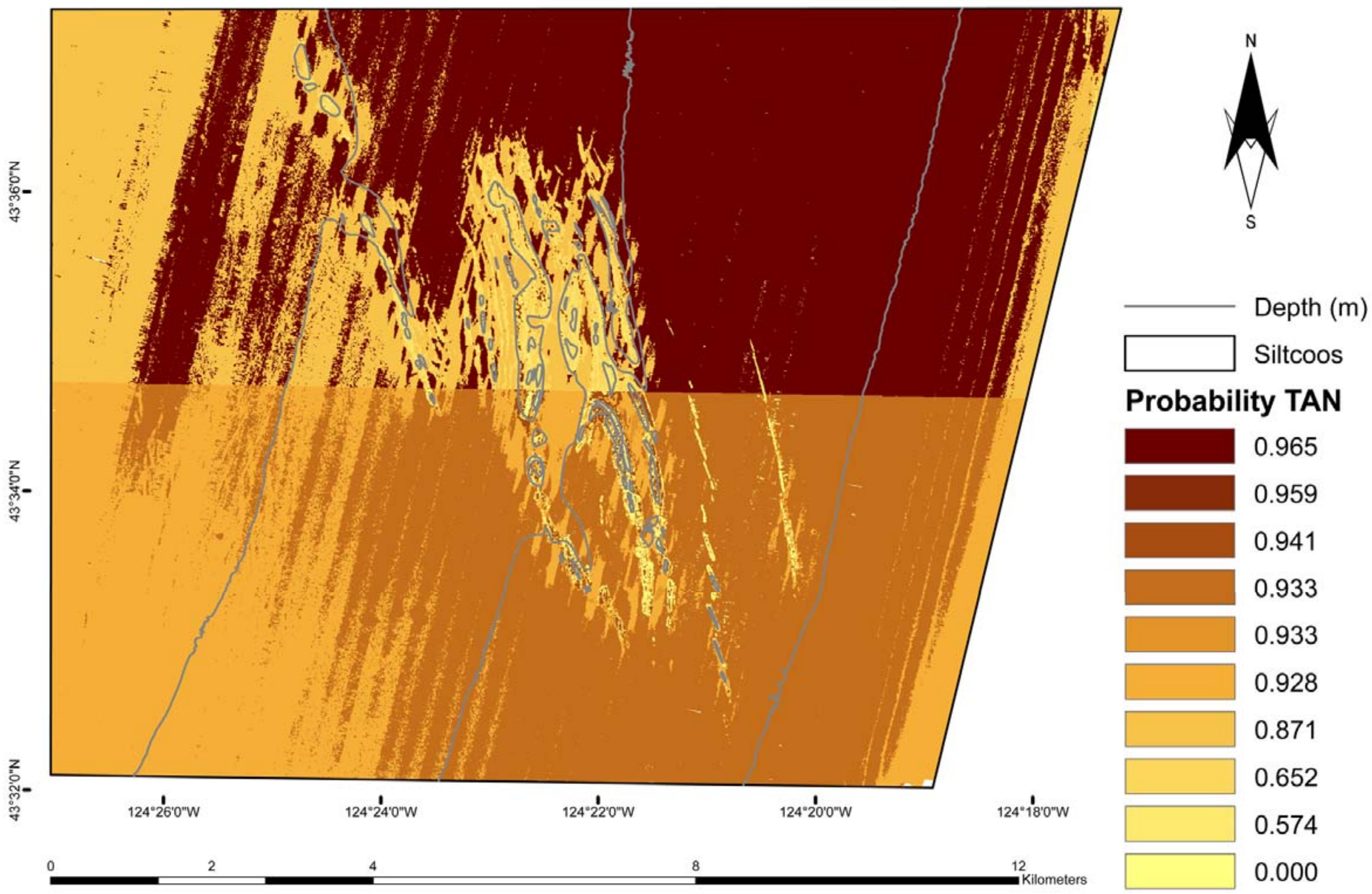


Figure 12. The habitat suitability map spatially depicts posterior probabilities for the TAN model. Probabilities range from approximately 0.00 to 0.97 with darker colors representing higher probabilities

NB model temperature sensitivity test

This *HSP difference* map allows a visualization of the location and magnitude of the temperature sensitivity test (Figure 13). Between the two runs, possible *HSP* values varied as little as 0.01 to as much as 0.05 with the greatest changes occurring in the northern portion of the study area. The smallest change occurred in the southern portion of the study area where sediments were finest while the greatest change occurred over rock substrates in the northern portion. This sensitivity analysis revealed the same north versus south pattern of sediment influence.

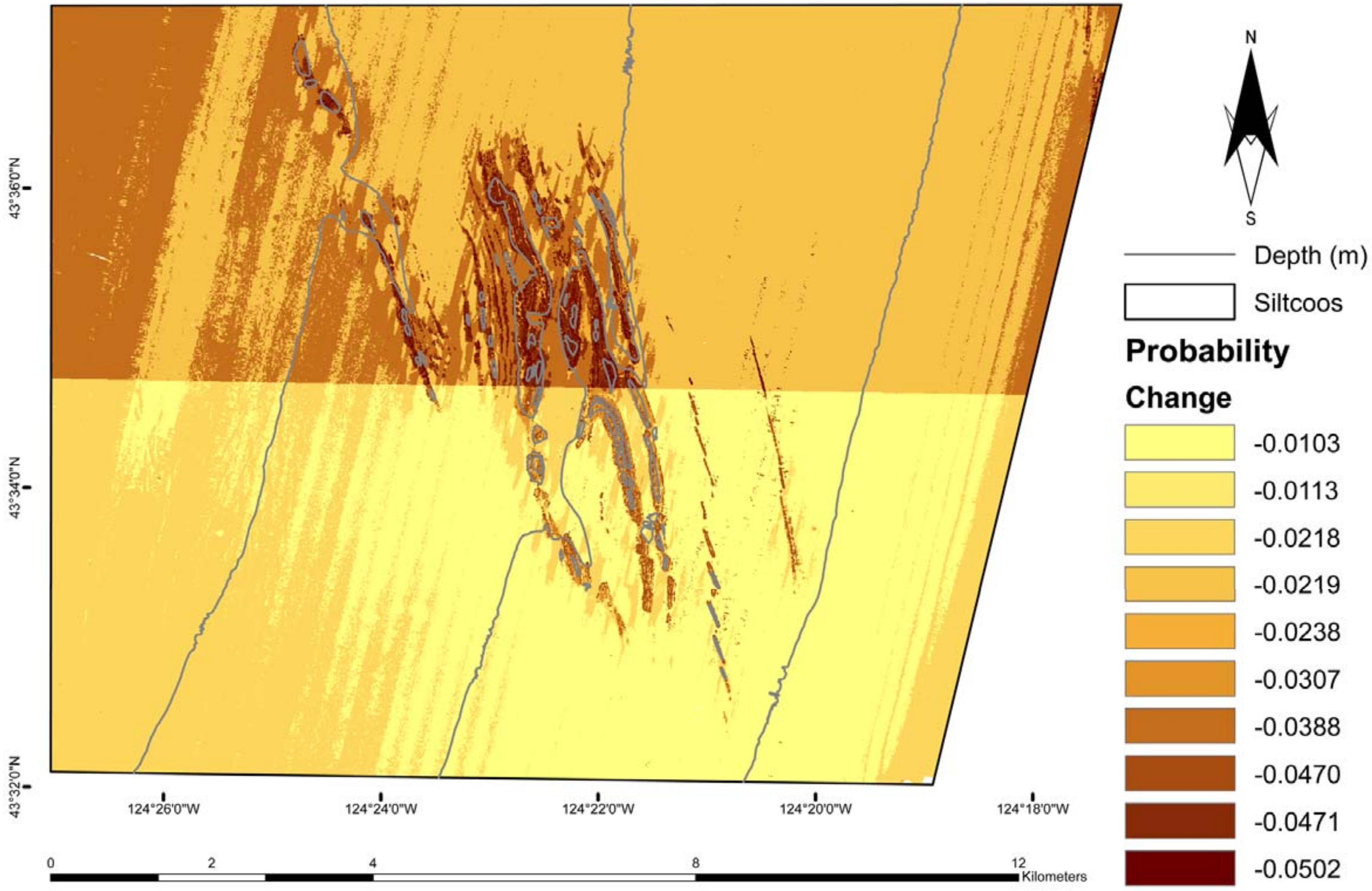


Figure 13. HSP map showing the difference between the NB temperature-high and temperature low runs.

Confidence raster for TAN model

Confidence values for the TAN model range from approximately ± 0.01 to ± 0.49 (Figure 14) A striking result is that the northern portion of the map contains both extremes in the range of confidence values. Predictions over sandy mud habitat to the north have high measures of confidence (± 0.96 ± 0.18) while probabilities associated with rocky habitat show almost complete uncertainty (0.57 ± 0.49). We predicted, with great certainty, a suitability of approximately 0.00 ± 0.01over rocky substrate in the southern portion. In general, we observed a comparable decrease in confidence as posterior probabilities approached uncertainty (0.50).

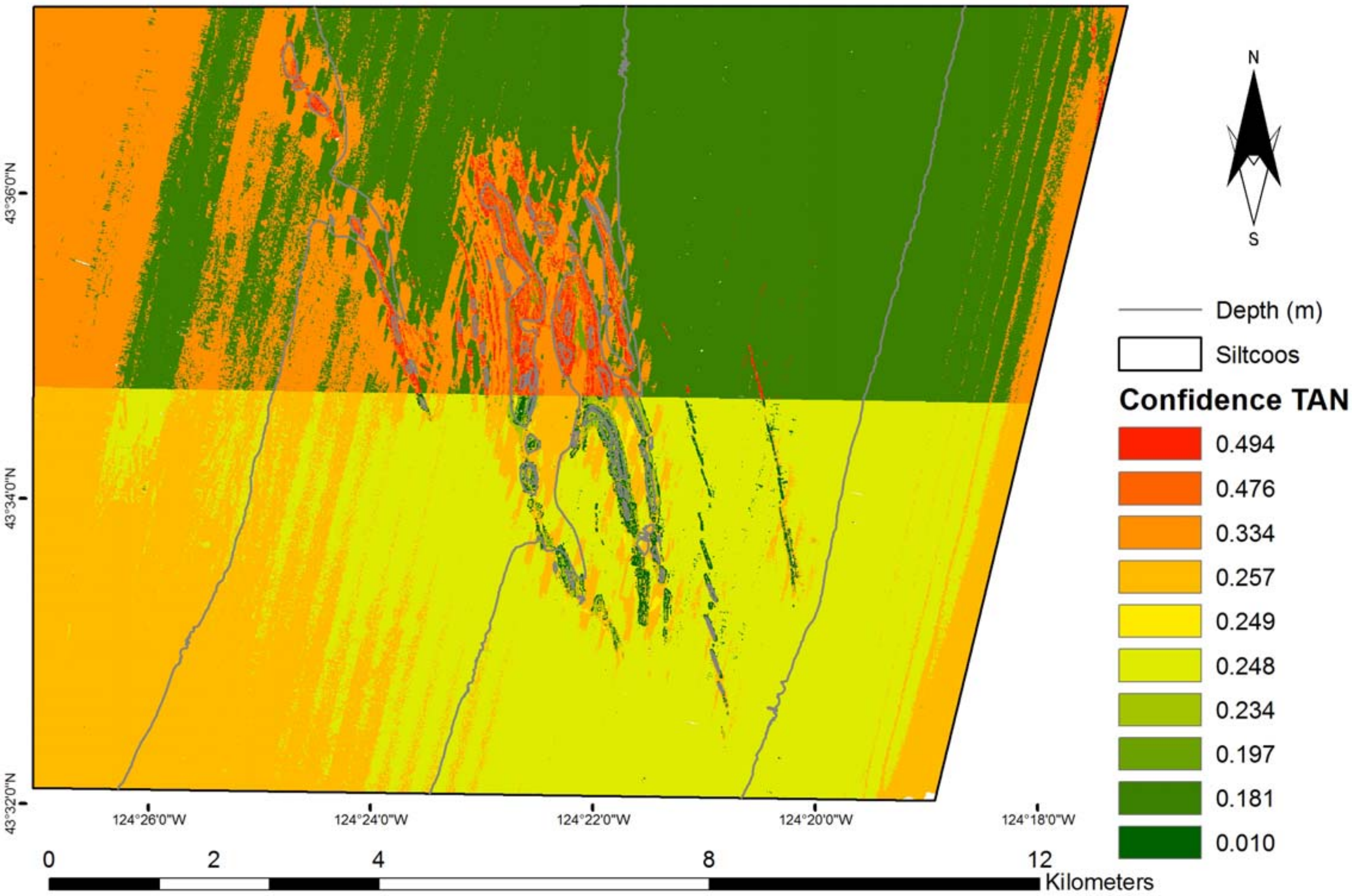


Figure 14. The confidence raster shows one standard deviation for each posterior probability value. Values can range from 0.00 (high certainty) to 0.50 (complete uncertainty).

Discussion

Modeling Approach

A shifting paradigm

Most species distribution modeling techniques employ a null hypothesis testing approach when considering the relationships between the target variable and the predictor variables (Rushton et al., 2004). Predictor variables are accepted and included into the model if they decrease the model variance by a suitable amount or if the regression coefficient is significantly different than zero (Rushton et al., 2004). This forces the modeler to make decisions to include or exclude predictor variables based on levels of arbitrary statistical significance (Rushton et al., 2004). Burnham and Anderson (2002) question the value of the null-hypothesis testing approach to distribution modeling because, in experiments, the scientist has control over the target and predictor variables and randomization of the treatment and controls. Complete control is difficult in field experiments especially in the marine environment. Furthermore, the ability to compare the influence of predictor variables on the overall distribution of suitable habitat provides an interesting approach toward understanding the processes linking habitat suitability.

The use of Bayesian methods in this study differs from the traditional modeling approach in that all predictor variables were included regardless of the magnitude of their influence on the target variable. This allows for a more complete representation of the abiotic factors influencing species distribution and therefore habitat suitability. Additionally, sensitivity findings analyses do provide interesting insight toward understanding which predictor variables may contribute to overall suitability. Bayesian belief network methods differ from direct interpolation of species data using geostatistical methods, such as krigging, which limit the investigation of environmental correlates thus preventing extrapolation over similar environments (Franklin, 2009). Additionally, the intuitive graphical interface of a *BBN* and rapid propagation of posterior probabilities facilitates exploring hypothetical scenarios of inference. Rapid inference is a very interesting strength of the *BBN* approach but the utility and value of any model depends on its intended use and the performance should be assessed accordingly.

Interpretation of HSP maps

When interpreting the results of this modeling process, it is essential to understand that our learning data set was collected over a regional scale but the *HSP* maps provided are on a local scale. The local scale *HSP* maps essentially represent ‘postage stamp’ within a regional prediction model. For this reason is difficult to immediately appreciate the predictive capacity of these models.

Ecological realism and the origin of prediction error

The best model is usually the one with the lowest error rate. When considering relative performance, we must consider both accuracy and cost. Fielding and Bell (1997) recommend exercising caution when making statements about model accuracy. The statement of model error should justify the choice of error measure. In other words, the intended use of model predictions will determine which performance metrics are most informative. Almost all errors are either algorithmic or biotic in nature (Fielding and Bell, 1997). Algorithmic errors are limitations imposed by the classification algorithm and data collection processes. Biotic errors force us to consider the ecological context of predictions.

Algorithmic errors

Performance

Data partitioning and testing with independent data is a robust method for model performance assessment. However, the data partitioning method can influence error rates. Dual partitioning (K=2) is the simplest method but the accuracy assessment is highly dependent on data contained within a single partition, especially with very small data sets. Our small data set, when partitioned, leaves just enough training and testing data to derive meaningful performance assessment results. Both confusion matrices exhibit the effects of partitioning, although much more noticeably for data sets *B* and *D* of the *NB* model (Table 7). Data set *B* has the highest overall prediction error rate (0.26) and lowest *Ka* statistic (0.42) of all models while data set *D* is the opposite with a 0.04 error rate and 0.88 *Ka* statistic. Small training partitions can reduce accuracy whereas larger testing partitions reduce the variance of error. Thus, there are trade-offs between developing an accurate classification rule and sufficiently assessing model performance.

The effects of partitioning can be seen again by examining the *ROC* curves and *AUC* values. Testing partitions contained insufficient data for Netica® to determine cutoff divisions, which explains the *Chance* results. While it is true that the *All* models likely overestimate model performance, it is revealing of the partitioning and data set size effects. Perhaps an average of models *A-D* including the *All* model will provide a closer estimate.

The true-positive (*Se*) and true-negative (*Sp*) rates for both models suggest the models are making accurate and balanced predictions, especially the *NB* model. *FPR* (0.14) and *FNR* (0.16) are similarly low and balanced in the *NB* model indicating non-discrimination between error types. The *TAN* model appears quite good at predicting species absence with a high true-negative rate of 0.95, which suggests that low *P'* may be affecting the model performance. For example, if this model's intended use was for conservation of *Rhabdus rectius*, we would require a reduced *FNR* relative to *FPR*. We could tolerate incorrectly predicting true habitat suitability whereas incorrectly predicting habitat as unsuitable would be more harmful.

The *Ka* statistic is less affected by low *P' .* For the *NB* model, a *Ka* of 0.63 suggests a quality of agreement ranging from good to substantial while the *TAN* model exhibits excellent or substantial quality of agreement (*Ka* = 0.75).

Revisiting the *ROC* curves and *AUC* values, we can make additional assessments of model performance that are less affected by *P'* and thresholds. Measures derived from confusion matrices assume that both error types are equivalent (Fielding and Bell, 1997). It is possible to alter the thresholds to mimic the costs associated with error types, if desired. The *ROC* values are obtained by applying a 0.5 threshold value to the continuous output variable that lies in the interval [0,1]. If the threshold is changed, the values of the confusion matrix will change (Fielding and Bell, 1997). Since these models assume 50-50 threshold values, information contained in the *ROC* curves should be relatively straightforward. The *AUC* value of 77.5 for the *NB* model means that 77.5% of the time a prediction is better than chance (Deleo, 1993).

Sampling design

Two examples illustrate the scenario where suitable habitats have been inadequately sampled. The most obvious example is the spatial limitation of our sampling design. This study adopted a sampling design targeting areas on the continental shelf considered suitable for wave device placement rather than adopting a sampling design that targeted a more realistic potential range of benthic invertebrates. For example, Scaphopods are known to exist to depths well beyond what is suitable for current technological constraints of anchored wave devices. Sampling should be to the scale appropriate for the ecological processes thought to determine species distribution (Rushton et al., 2004). *Rhabdus rectius* are very small organisms, and once recruited could not travel great distances on their own. It is unlikely that most infaunal species move great distances about the landscape, or move at rates that would cause difficulty in sampling, however they may form clusters that our sampling design may have missed. Cluster detection and analyses should be performed and/or accounted for in the belief network.

Collecting instantaneous measurements of environmental variables is a major source leading to less well-defined links between variables (Almeddine, et al., 2011) and a possible source of error. Marine environmental data is inherently variable, especially the ephemeral characteristics of the overlying water column. Marine sediments and water properties change on different time scales over the study area and this presents a challenge for developing model parameters. One solution is to perform robust multivariate outlier detection to identify and minimize the impact of outliers (Alameddine, et al., 2010). However, variables describing water properties had little influence on *Rhabdus rectius* yet those variables change more ephemerally than do sediment variables. Sediment variables do change but are relatively more stable, especially at most of the depths surveyed for this project. So, in this particular study, this may be a smaller source of error than in other studies.

Limitations of habitat classification

When using the *SGH* surrogate for MedianGS, the issue of *SGH* misclassification arises immediately. The issue is about the degree of certainty that the classified habitat is actually that habitat. A few examples illustrate the issue clearly.

The first example of *SGH* classification error is the simplest to consider. Sediment samples are processed for grain size distribution which is numeric value on a continuum. A grain size value does not hold as much value as a sediment type or class for most mapping purposes. At some point the grain size value has to be translated into classes by rules we define. For the purposes of habitat suitability prediction requiring grain size, consider making grain size distribution maps instead of cross-walking from grain size to a class then back to grain size. This circuitous technique certainly loses information along the way and propagates uncertainty. These maps could be updated periodically and perhaps used for other purposes. Another solution is to add the sediment classes to the original data set used for learning. This would not change the uncertainty accrued during the classification stage but would prevent the need for a surrogate node during modeling.

Another error begins during multi-beam data acquisition and processing. Sounding artifacts are detectable along the nadir beam tracks and areas where soundings are sparse. In essence, these artifacts introduce ‘hot’ soundings (representative of highly reflective substrate, such as rock) that are nearly impossible to remove. The ‘hot’ soundings propagate through the mapping process and are misclassified because they imitate more reflective substrate. Artifacts are somewhat smoothed out during GIS mapping but not entirely. The next example is due to the fact that rock mapping methods are semi-quantitative requiring a considerable amount of expert supervision (Appendix A). Some rocky areas may be over-represented and vice versa. The result is an imperfect account of accuracy.

Very few circumstances would permit a benthic infaunal organism, such as *Rhabdus rectius*, to inhabit rock substrates. One exception is when the rock substrate is buried in a layer of sediment that is otherwise suitable for the burrowing organism. Contemporary *SGH* mapping techniques cannot differentiate between rock substrates with and without a soft sediment surficial layer, at least not in a cost-effective way on a regional scale. Thus, modified rock substrates may be misclassified as purely rock. While these imperfections in the *SGH* mapping process are noted, the total area misclassified is minuscule relative to the prediction scale and has a negligible effect on regional habitat suitability prediction.

Biotic errors

Species niche concept

The Hutchinsonian species niche is defined as the environmental dimensions within which that species can survive and reproduce. Furthermore, the realized ecological niche is defined as a subset of the fundamental physiological niche, where biotic factors are excluded. Territory size is probably not fixed for a species but varies with the individual. In this study, we are less concerned with the variance of an individuals' range (very small scale for infauna) rather we are more interested in the utilized niche on a regional scale. We could consider the global range (fundamental niche) of the species, or Scaphopoda for that matter, but the result would likely indicate that our entire region is highly suitable. This provides little use for a regional spatial planning initiative. Rather, we accept that the region of the world is highly suitable but consider how habitat suitability varies within the region.

Biotic errors represent "unaccounted for ecologically relevant processes" (Fielding and Bell, 1997) such as unsaturated habitat, biotic interactions and latent variables influencing species occurrence. For example, a benthic infaunal species may not occupy all suitable habitats because the species has not yet reached equilibrium with the environment. This scenario is an example of unsaturated habitat.

Organisms may be in their current positions because of past events rather than current events (Fielding and Bell, 1997). For example, the continental shelf submarine environment has experienced significant geological change since the last glacial maximum. An advancing coastline leaves in its wake a shallow sea of altered surficial geology, increased seafloor surface area and exposure to dynamic currents. These historical events influence the distribution of potential habitat and dispersal of benthic organisms.

Benthic species are not only influenced by the physical seafloor characteristics but by the water column properties and movement above them. The whim of ocean currents subject many benthic organisms to an uncertain fate during early life ontogeny. Larval dispersal and recruitment success raises the issue of the role of population dynamics. Some habitats may be sources of larvae for dispersal whereas other habitats may be sinks for dispersed larval organisms. Although areas of the ocean may appear to be suitable because we observed species there, they may in fact be unsuitable sink habitats where metapopulations are slowly shrinking.

Furthermore, species interactions such as predation by ratfish, noted earlier, could further contort model performance and prediction interpretations. A species will not have unconstrained habitat selection and therefore exist within an ecological ‘bubble’ influencing their preferences. Failure to incorporate these considerations into prediction models can influence the value of such models. While some of these issues would be difficult to incorporate within a predictive model, the models that do are more ecologically meaningful (Rushton et al., 2004) and perhaps more useful.

Spatial context of errors

Previous model performance metrics do not account for spatial context of errors. Two different models can create identical errors but produce different prediction maps. Spatial autocorrelation is an issue rarely accounted for in predictive model and any efforts to address these issues should benefit the understanding of the predictive utility.

Future action

Model modifications and validation

Independent testing

The simplest way to assess any model is to compare model predictions with observed data (Rushton et al., 2004). At each available opportunity, effort should be made to go to a new area within the model bounds and repeat the sampling methods. Ideally, the area should be mapped with multi-beam sonar. Using this combined data, create surficial geologic habitat maps. For *BBN* testing purposes, make predictions at the new site with the existing model(s) and compare with the organism occurrences found there.

Update the knowledge base

Once the test has been completed, the new data that was gathered should be used in an exercise of updating the knowledge base. Rather than repeating the modeling process with the new data set appended to the old, one should incorporate the new cases into the existing model to update the conditional probability tables. Of course, one could rebuild the models from scratch if warranted. For example, if the new data is outside of the existing model bounds, the modeling process should be repeated. One should then compare model parameters, performance measures and spatial predictions.

Avoid proximal variables

Examination of the TAN HSP and confidence rasters (Figure 12 and Figure 14) begs the question, “What is it about latitude that causes the large difference in posterior probabilities and confidence for rocky substrates?” The answer begins with our definition of rocky substrate. Unlike the all other SGH classes, we have defined rocky substrate as known with complete certainty (i.e we are absolutely certain our classification of rocks is accurate). However, as noted in the *Limitations of habitat classification* section and Appendix A, use of a discrete SGH classifier as a surrogate for a continuous grain size variable introduces a different form of uncertainty.

Since the SGH variable was not used in the learning process, decisions had to be made regarding its placement in the model structure and parameterization (see Appendix A). The largest grain size state in the *MedianGS* node is 211.5 µm (fine sand) to 578.5 µm (coarse sand). We know that rock does not belong in this *MedianGS* state but, in an effort to facilitate initial model development, we did not create an additional state for rocky substrate. Thus, the model views rocky substrate as sand, which increases the probability of observing *Rhabdus rectius*, even though we have never sampled rocky substrate. This is why we see posterior probabilities of 0.57 ± 0.49 over rocky substrate in the northern latitude state. Conversely, we observe probabilities 0.00 ± 0.01 over rocky substrates in the southern latitude state because *Rhabdus rectius* was never observed in grain sizes greater than 211.5 µm south of 43.57 N.

There are several solutions to this kind of problem but the best solution for the foreseeable future is to create a grain size distribution map for the region. This would eliminate the need for the SGH proxy thus improving accuracy and confidence and greatly simplifying the model.

Hypothesis generation

Sensitivity

A great strength of *BBN* modeling is the ability to incorporate disparate data types and make inferences about uncertain or unknown relationships between variables. One could collect or synthesize empirical data, expert opinion and/or theory, build it into the model and make inferences. For example, one of these models above could be modified to include a *Predation* node whose influence on *Rhabdus rectius* is theoretically negative. The magnitude of this negativity is questionable, but playing out 'what-if' scenarios allows sensitivity exploration. Sensitivity testing can be thought of as an assessment of the influence an environmental variable has on the target species. This can be a very useful tool for exploring other relationships as well. The *NB* model temperature sensitivity test was an attempt to illustrate this approach.

Temperature seemed to be an appropriate candidate for an additional uncertainty test where we could simulate a 'warming' trend, in the context of climate change. The sensitivities to changes in temperature were not huge yet they were noticeable and may reveal more if mapped over the regional scale. In hindsight, we should have considered using a different variable or introducing another, such as a theoretical biotic interaction, because not only did the 'built-in' Netica® sensitivity testing reveal *Temperature* had little effect on the target, but it also exhibited a bimodal response curve. This probably further disguised the sensitivity. Nevertheless, studying the sensitivities and responses to perturbation in model findings can perhaps shed light on old problems and inspire new questions.

Probabilistic Inference

Although we are immediately concerned with the species variable output, we may be interested in the results at other nodes. An immediately intuitive way to investigate inference is to force the target variable to a maximum and examine how the predictor variable curves behave (Figure 15). This exercise essentially reveals the organism's preferences for the predictor variables. Examining the response curves for cases where *Rhabdus rectius* was present (True) shows unimodal distributions for all predictor variables. One exception, which is true for both model types, is the bimodal distribution for the temperature variable.

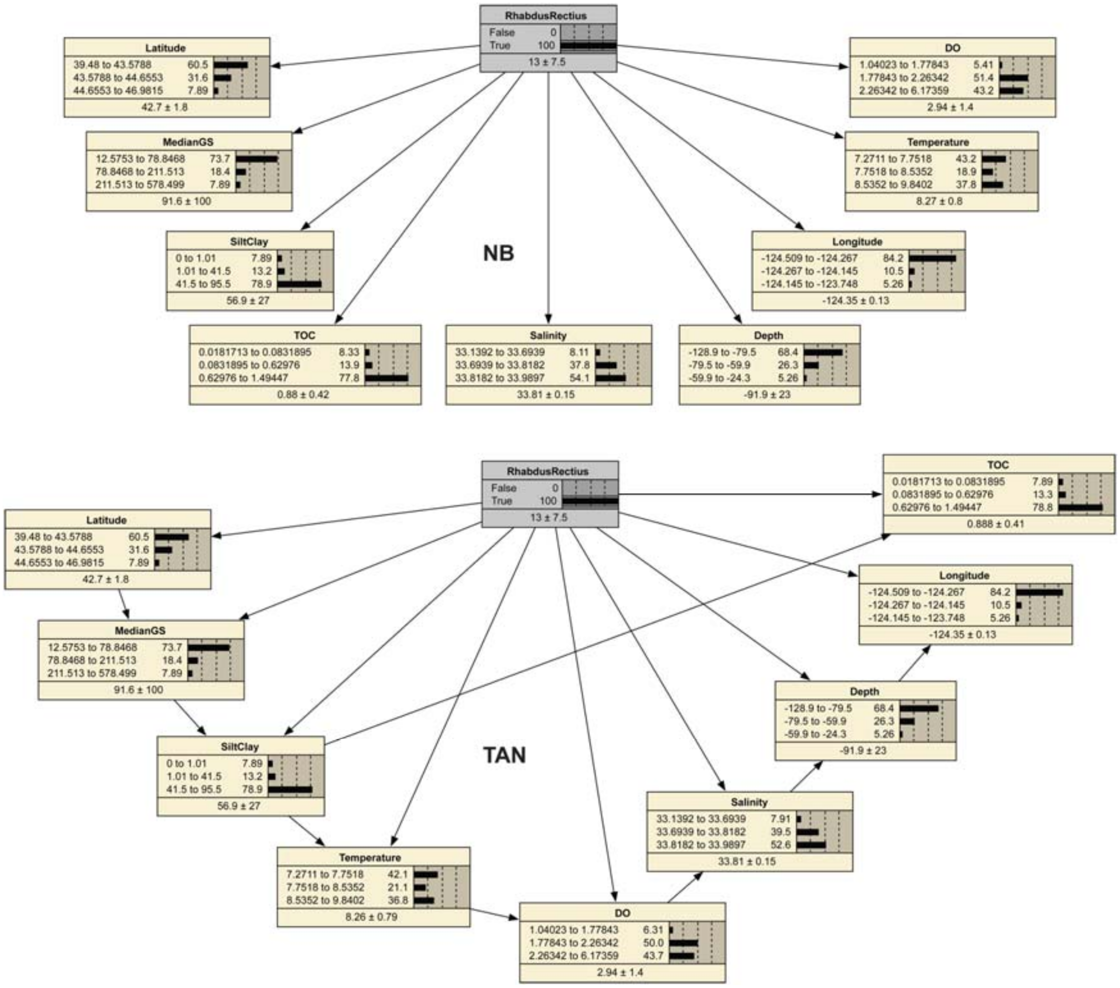


Figure 15. Both models were forced to show the prior distributions for cases where the target species, *Rhabdus rectius*, was present. The resulting curves can be thought of as the target's preferences for the environmental variables or the abiotic conditions necessary for target occurrence.

Model integration with developing marine spatial planning tools

This study serves as stepping stone toward the development of a holistic marine spatial planning decision-support tool called Bayesian Analysis for Spatial Siting (*BASS*). At the time of this writing, the status of *BASS* is nascent, however it holds promise to be one of the most unique approaches to management of marine resources. The methods described here can be used in isolation and may add value when used in conjunction with tools such as

BASS, which incorporates and combines many models like the ones presented here with stakeholder input. This approach of merging science with qualitative, values-based information is a novel approach to making decisions. Furthermore, the decision-making power of *BASS* should increase as individual models are created, updated and refined.

Bibliography

Abelson, A., and M. Denny (1997), Settlement of Marine Organisms in Flow, *Annual Review of Ecology and Systematics*, *28*(1), 317-339, doi:10.1146/annurev.ecolsys.28.1.317. [online] Available from: http://arjournals.annualreviews.org/doi/abs/10.1146%2Fannurev.ecolsys.28.1.317

Aguilera, P. A., A. Fernández, F. Reche, and R. Rumí (2010), Hybrid Bayesian network classifiers: Application to species distribution models, *Environmental Modelling Software*, *25*(12), 1630-1639, doi:10.1016/j.envsoft.2010.04.016. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S1364815210001222

Alameddine, I., Y. Cha, and K. H. Reckhow (2011), An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics, *Environmental Modelling Software*, *26*(2), 163-172, doi:10.1016/j.envsoft.2010.08.007. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S1364815210002355

Anderson, D. R., and K. P. Burnham (2002), Avoiding pitfalls when using information-theoretic methods, The Journal of Wildlife Management, 66(3), 912–918. [online] Available from: http://www.jstor.org/stable/3803155

Bedard, R., G. Hagerman, M. Previsic, O. Siddiqui, R. Thresher, and B. Ram (2005), Final Summary Report Project Definition Study Offshore Wave Power Feasibility Demonstration Project, *Energy*.

Boehlert, G. W., and A. B. Gill (2010), Environmental and ecological effects of ocean renewable energy development: A current sythesis, *Oceanography*, *23*(2), 68-81.

Boehlert, G. W., S. Henkel, and C. Goldfinger (2011), Survey of Benthic Communities near Potential Renewable Energy Sites Offshore the Pacific Northwest. Annual Report for Year 1: June 1, 2010 – June 30, 2011.

Cada, G., J. Ahlgrimm, M. Bahleda, T. Bigford, S. D. Stavrakas, D. Hall, R. Moursund, and M. Sale (2007), Potential Impacts of Hydrokinetic and Wave Energy Conversion Technologies on Aquatic Environments, *Fisheries (Bethesda)*, *32*(4), 174-181, doi:10.1577/1548-8446(2007)32[174:PIOHAW]2.0.CO;2. [online] Available from:

http://afsjournals.org/doi/abs/10.1577/1548-8446(2007)32%5B174%3APIOHAW%5D2.0.CO%3B2

Commission, C. E., C. Ocean, and P. Council (2008), Developing wave energy in coastal california: potential socio-economic and environmental effects, *Energy*, (November).

Deleo, J. M. (1993), Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty, in *Uncertainty Modelling and Analysis Proceedings on the Second International Symposiumon*, pp. 318-325, IEEE Computer Society Press.

Duda, R. O., P. E. Hart, and D. G. Stork (2001), *Pattern Classification*, edited by R. O. Duda, P. E. Hart, and D. G. Stork, Wiley. [online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.1318&amp;rep=rep1&amp;type=pdf

Farrington, J. W. (1991), Biogeochemical processes governing exposure and uptake of organic pollutant compounds in aquatic organisms., *Environmental Health Perspectives*, *90*, 75-84. [online] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1519506&tool=pmcentrez&rendertype=abstract

Fielding, A. H., and J. F. Bell (1997), A review of methods for the assessment of prediction errors in conservation presence/absence models, *Environmental Conservation*, *24*(1), 38-49, doi:10.1017/S0376892997000088. [online] Available from: http://www.journals.cambridge.org/abstract_S0376892997000088

Forbes, A. D. (1995), Classification-algorithm evaluation: five performance measures based on confusion matrices., *Journal Of Clinical Monitoring*, *11*(3), 189-206. [online] Available from: http://www.springerlink.com/content/j50806x155833798/

Franklin, J. (2009), *Mapping Species Distributions : Spatial Inference & Prediction*, Cambridge University Press.

Friedman, N., D. Geiger, and M. Goldszmidt (1997), Bayesian Network Classifiers, edited by G. Provan, P. Langley, and P. Smyth, *Machine Learning*, *29*(1), 131-163, doi:10.1016/j.patcog.2004.05.012. [online] Available from: http://www.springerlink.com/index/PG72774318173365.pdf

Gill, A. B. (2005), Offshore renewable energy: ecological implications of generating electricity in the coastal zone, *Journal of Applied Ecology*, *42*(4), 605-615, doi:10.1111/j.1365-2664.2005.01060.x. [online] Available from: http://doi.wiley.com/10.1111/j.1365-2664.2005.01060.x

Guisan, A., and N. E. Zimmermann (2000), Predictive habitat distribution models in ecology, *Ecological Modelling*, *135*(2-3), 147-186, doi:10.1016/S0304-3800(00)00354-9. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0304380000003549

Goldfinger, C. et al. (2012), Oregon State waters mapping program final report. *In preparation.*

Huberty, C. J. (1994), *Applied Discriminant Analysis*, Wiley. [online] Available from: http://www.amazon.com/dp/0471468150

Inger, R. et al. (2009), Marine renewable energy: potential benefits to biodiversity? An urgent call for research, *Journal of Applied Ecology*, *46*(6), 1145-1153, doi:10.1111/j.1365-2664.2009.01697.x. [online] Available from: http://doi.wiley.com/10.1111/j.1365-2664.2009.01697.x

Kristensen, E., T. K. Andersen, and T. H. Blackburn (1992), Effects of macrofauna and temperature on degradation of macroalgal detritus: the fate of organic carbon, *Limnology and Oceonography*, *37*, 1404-1419.

Landis, J. R., and G. G. Koch (1977), The measurement of observer agreement for categorical data, *Biometrics*, *33*(1), 159-174, doi:10.2307/2529310. [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/843571

Langhamer, O., D. Wilhelmsson, and J. Engström (2009), Artificial reef effect and fouling impacts on offshore wave power foundations and buoys – a pilot study, *Estuarine, Coastal and Shelf Science*, *82*(3), 426-432, doi:10.1016/j.ecss.2009.02.009. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0272771409000626

Liu, H., F. Hussain, C. L. I. M. Tan, and M. Dash (2002), Discretization: An Enabling Technique, *Data Mining and Knowledge Discovery*, *6*(4), 393-423, doi:10.1023/A:1016304305535. [online] Available from: http://www.springerlink.com/index/TUXY32PW4LG6832M.pdf

Manel, S., H. C. Williams, and S. J. Ormerod (2001), Evaluating presence-absence models in ecology: the need to account for prevalence, *Journal of Applied Ecology*, *38*(5), 921-931, doi:10.1046/j.1365-

2664.2001.00647.x. [online] Available from: http://doi.wiley.com/10.1046/j.1365-2664.2001.00647.x

Marcot, B. G., J. D. Steventon, G. D. Sutherland, and R. K. McCann (2006), Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation, *Canadian Journal of Forest Research*, *36*(12), 3063-3074, doi:10.1139/X06-135. [online] Available from: http://article.pubs.nrc-cnrc.gc.ca/ppv/RPViewDoc?issn=1208-6037&volume=36&issue=12&startPage=3063&ab=y

McCarthy, M. A. (2007), *Bayesian Methods for Ecology*, edited by C. U. Press, Cambridge University Press. [online] Available from: http://books.google.com/books?hl=en&amp;lr=&amp;id=WpeZyTc6U94C&amp;oi=fnd&amp;pg=PA1&amp;dq=Bayesian+methods+for+ecology&amp;ots=4Aqe-IMlZk&amp;sig=nFSizA4ol2sFvpDmeI4AnCgDeZM

Millar, D. L., H. C. M. Smith, and D. E. Reeve (2007), Modelling analysis of the sensitivity of shoreline change to a wave farm, *Ocean Engineering*, *34*(5-6), 884-901, doi:10.1016/j.oceaneng.2005.12.014. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0029801806001272

Montgomery, J. C., A. Jeffs, S. D. Simpson, M. Meekan, and C. Tindle (2006), Sound as an Orientation Cue for the Pelagic Larvae of Reef Fishes and Decapod Crustaceans, *Advances in Marine Biology*, *51*(06), 143-96, doi:10.1016/S0065-2881(06)51003-X. [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16905427

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann. [online] Available from: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp;path=ASIN/1558604790

Pelc, R. (2002), Renewable energy from the ocean, *Marine Policy*, *26*(6), 471-479, doi:10.1016/S0308-597X(02)00045-3. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0308597X02000453

Reynolds, P. D. (2002), The Scaphopoda., *Advances in Marine Biology*, *42*, 137-236. [online] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12094723

Rushton, S. P., S. J. Ormerod, and G. Kerby (2004), New paradigms for modelling species distributions?, *Journal of Applied Ecology*, *41*(2), 193-200, doi:10.1111/j.0021-8901.2004.00903.x. [online] Available from: http://doi.wiley.com/10.1111/j.0021-8901.2004.00903.x (Accessed 9 October 2011)

Scheffer, M. (1999), The effect of aquatic vegetation on turbidity; how important are the filter feeders?, *Hydrobiologia*, *408*, 307-316.

Sebastiani, P., and T. Perls (2008), *Bayesian networks: a practical guide to applications*, edited by O. Pourret, P. Naim, and B. Marcot, John Wiley and Sons.

Shields, M. A., D. K. Woolf, E. P. M. Grist, S. A. Kerr, A. C. Jackson, R. E. Harris, M. C. Bell, R. Beharie, A. Want, and E. Osalusi (2011), Marine renewable energy: The ecological implications of altering the hydrodynamics of the marine environment, *Ocean Coastal Management*, *54*(1), 2-9, doi:10.1016/j.ocecoaman.2010.10.036. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0964569110001924

Snelgrove, P. V. R. (1997), The importance of marine sediment biodiversity in ecosystem processes, *Ambio*, *26*(8), 578-583. [online] Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0031474294&partnerID=40&md5=a3e64e65a0015da3aa4b7a7db19bd24c

Snelgrove, P. V. R. (1999), Getting to the Bottom of Marine Biodiversity : Sedimentary Habitats, *BioScience*, *49*(2), 129-138, doi:10.2307/1313538. [online] Available from: http://www.jstor.org/stable/1313538?origin=crossref

Uusitalo, L. (2007), Advantages and challenges of Bayesian networks in environmental modelling, *Ecological Modelling*, *203*(3-4), 312-318, doi:10.1016/j.ecolmodel.2006.11.033. [online] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0304380006006089

Wright, D., and W. Heyman (2008), Introduction to the Special Issue: Marine and Coastal GIS for Geomorphology, Habitat Mapping, and Marine Reserves, *Marine Geodesy*, *31*(4), 223-230, doi:10.1080/01490410802466306. [online] Available from: http://www.informaworld.com/openurl?genre=article&doi=10.1080/01490410802466306&magic=crossref.

APPENDIX

Appendix A. The classification and incorporation of grain size values

Surficial geologic habitat classification

The Oregon state waters mapping effort made significant progress toward improving the quantity and quality of environmental data available for predictive mapping, specifically through collection of high-resolution multi-beam sonar data and classification of surficial geologic habitat (*SGH*). State waters *SGH* mapping techniques combined high-resolution multi-beam sonar information with grain size analyses to characterize surficial geology. The result was a conversion from observed numerical grain size values to a classified surficial geologic habitat (*SGH*). We employed these methods similarly with the following modifications.

The *SGH* map created for this study is an eight-meter resolution raster (Figure 1). We used median rather than mean grain size values for our maximum likelihood habitat classification. Every raster cell received a positive integer 'gridcode' representative of the particular *SGH* class found there. The gridcode value of zero was reserved for the rock habitat class. Bathymetry and backscatter rasters were examined to roughly determine where rocks exist. A rugostiy raster was created then compared with bathymetry and backscatter and reclassified to account for rocks. Rocks were reclassified as zeros while everything else received a value of one. This new raster was multiplied by the original *SGH* raster resulting in a new *SGH* raster including rocks.

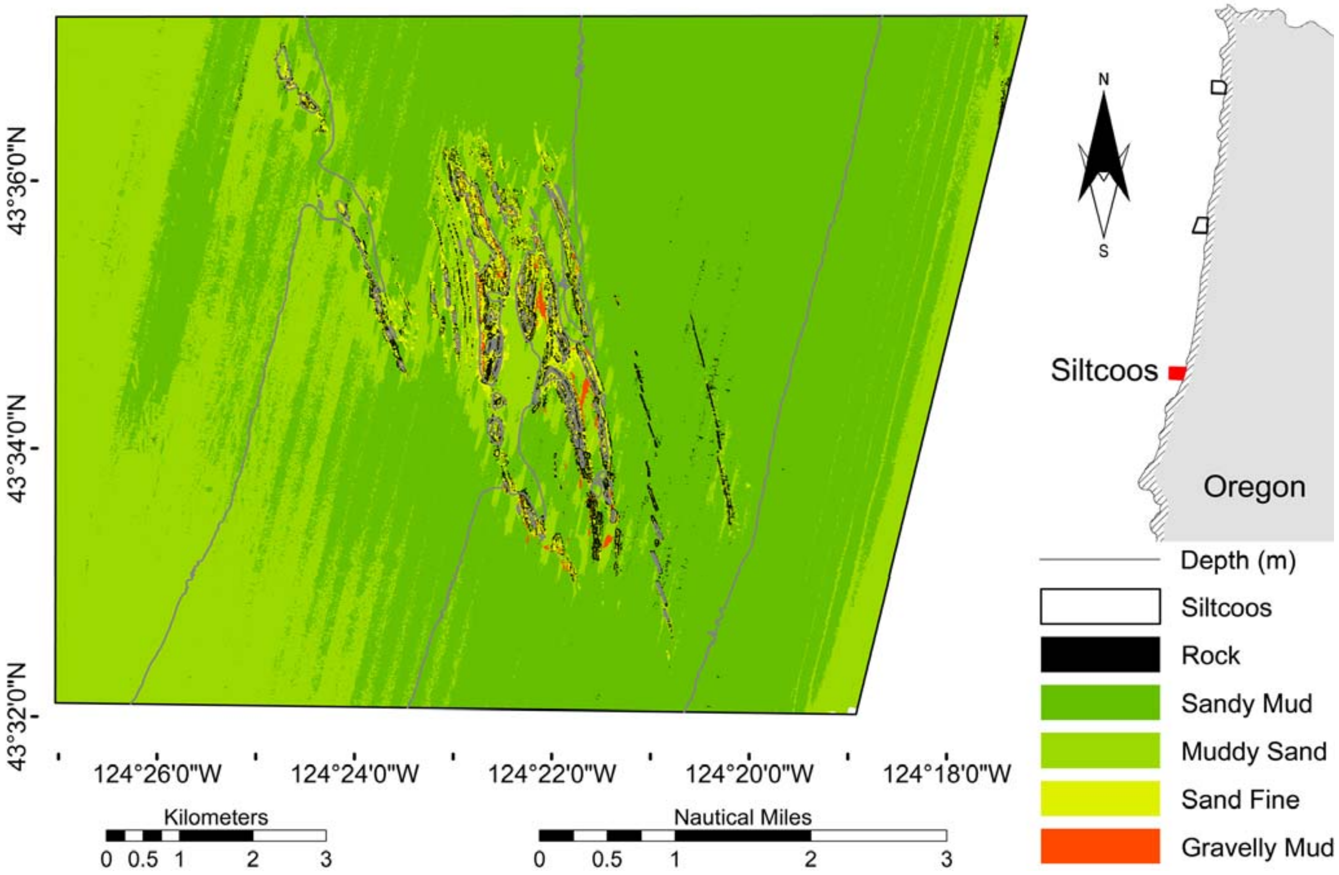


Figure 1. The map shows surficial geologic habitat (SGH) types found in the study area. Box core samples were processed for environmental variables, species density (1/10 m2), and used in model construction. Shipek grab samples were processed for sediment grain size distribution. All samples were used to create the SGH map.

Habitat as grain size surrogate

SGH data is available regionally, although the quality varies. Ideally, *SGH* is determined through rigorous seafloor sampling and grain size analyses because it will produce the most detailed and accurate maps. Unfortunately, grain size analyses are patchy on a regional scale. In general, sediment properties are thought to play a major role in the distribution of benthic infauna. Therefore, it was necessary to find an acceptable surrogate for median grain size (MGS) if we are to later make useful predictions on a regional scale.

MGS is determined by analyzing the grain size distribution of a seafloor sediment sample. Since grain size information is used to most accurately classify *SGH*, it was a reasonable solution to let *SGH* serve as a surrogate for MGS.

Gridcode node

The surrogate node, *Gridcode*, represents all possible *SGH*s identified in the Oregon state waters mapping project (Figure 2). This node was introduced simply to allow the model to accept *SGH* as an input. Linking the two nodes together required populating a conditional probability table (*CPT*) for the child node. Recall, the *CPT* is a numerical table quantifying the unique possible combinations between linked variables. Populating the *CPT* required identifying within which *MedianGS* state a given *SGH* finding would fall when entered into the model. *MedianGS* and *SGH* both are representations of the grain size continuum; however the bin widths are determined differently.

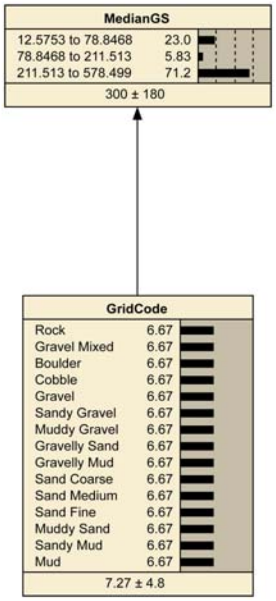


Figure 2. The Gridcode node represents the surficial geologic habitat (SGH) classes possible to date as defined by the Oregon state waters mapping project. SGH is a surrogate for MedianGS.

Gridcode classes are defined by how grain size characteristics align with the habitat classification scheme. For a thorough description of *SGH* characterization see the Oregon state waters mapping report (Goldfinger, et al., in-press). Recall that discretization divided the MGS node states into bins with approximately equal cases. These learned states did not align with sediment class breaks along the grain size continuum so the overlap between MGS and *SGH* needed to be determined.

For a given *SGH* finding entered into the model, within which MGS state(s) will it fall? For a sample to earn an *SGH* class of ‘Gravel’, 90 percent of the grain size distribution has to contain particles greater than 2000 µm. By definition, the remaining ten percent of sample has an equal chance of falling within *SGH* classes smaller than 2000 µm. The remaining ten percent was divided between classes accordingly (Figure 3). The percentages were summed and entered into the *CPT* (Table 1).

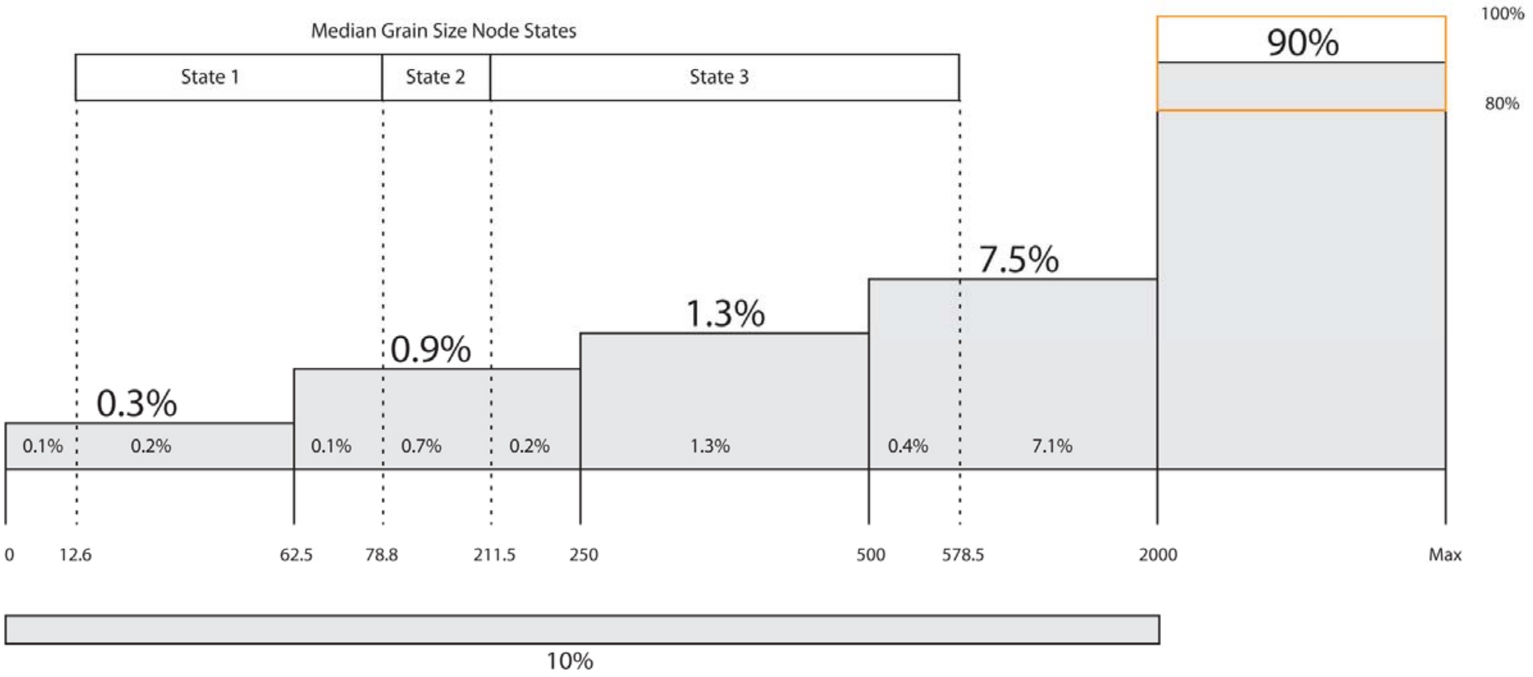
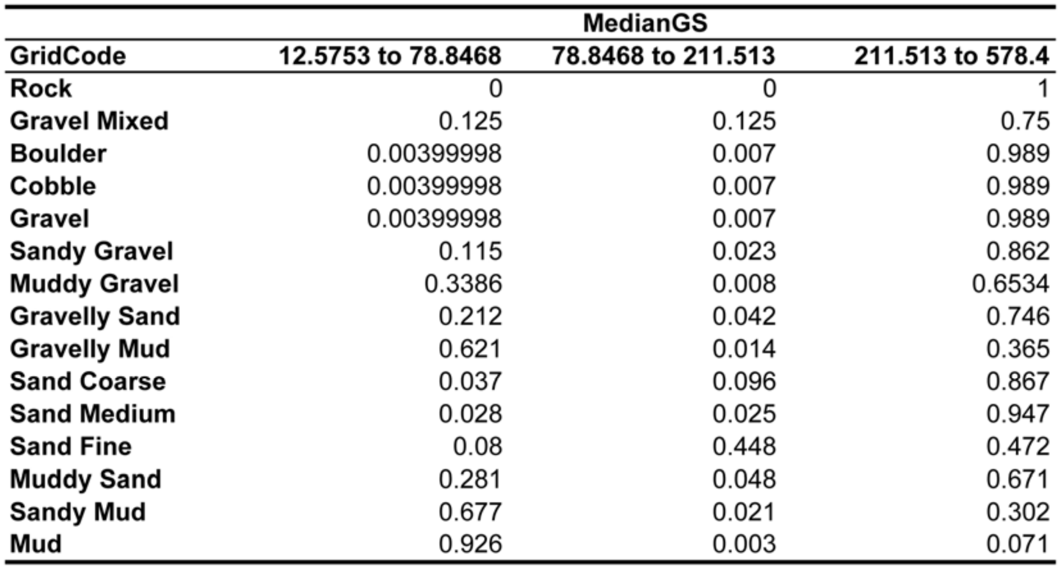


Figure 3. The grain size continuum showing the gravel SGH classification scheme with MedianGS node state widths overlay. Percentages within MedianGS node states were summed to determine the probability of the gravel falling within a given MedianGS state when gravel is entered as a Gridcode node finding. This diagram is not to scale.



	MedianGS		
GridCode	12.5753 to 78.8468	78.8468 to 211.513	211.513 to 578.4
Rock	0	0	1
Gravel Mixed	0.125	0.125	0.75
Boulder	0.00399998	0.007	0.989
Cobble	0.00399998	0.007	0.989
Gravel	0.00399998	0.007	0.989
Sandy Gravel	0.115	0.023	0.862
Muddy Gravel	0.3386	0.008	0.6534
Gravelly Sand	0.212	0.042	0.746
Gravelly Mud	0.621	0.014	0.365
Sand Coarse	0.037	0.096	0.867
Sand Medium	0.028	0.025	0.947
Sand Fine	0.08	0.448	0.472
Muddy Sand	0.281	0.048	0.671
Sandy Mud	0.677	0.021	0.302
Mud	0.926	0.003	0.071

Table 1. The table shows the possible unique combinations for both nodes. Probabilities are calculated by summing the bin percentages described in Figure 3.