Evaluating effects of rescaling and weighting data on habitat suitability modeling

Ying Xue [a,b], Lisha Guan [b], Kisei Tanaka [b], Zengguang Li [a,b], Yong Chen [b], Yiping Ren [a,*]

[a] Fisheries College, Ocean University of China, Qingdao 266003, China
[b] School of Marine Sciences, University of Maine, Orono, ME 04469, USA

ABSTRACT

Abundance index (AI), used to establish the suitability index (SI), provides critical information in habitat suitability index (HSI) modeling. The distributions of AIs derived from fisheries-independent surveys tend to be right skewed because of heterogenous distributions of fishes. The existence of large AI values and failure to consider it might result in underestimation of HSI values for most sampling areas. We compared the performance of HSI models based on original AIs (without any transformation) versus rescaled AIs (i.e., log-transformed AIs) using American lobster (*Homarus americanus*) along the coast of Gulf of Maine as an example. Impacts of weighting environmental variables on HSI modeling based on boosted regression tree (BRT) were also evaluated. Both cross-validation and predicted habitat suitability maps suggested that the weighted HSI model based on log-scaled AI data tended to yield a more reliable prediction of optimal habitats for American lobster. The unweighted HSI model based on the original AI data, however, tended to underestimate optimal habitats and overestimate suboptimal habitats. We recommend using log-transformed AIs and determining the weights of different environmental variables based on the BRT method in HSI modeling, especially when AI data are highly skewed.

© 2016 Published by Elsevier B.V.

1. Introduction

Suitable habitats represent favorable environmental conditions for species and thus provide space-time envelopes of optimal conditions (Petitgas et al., 2014). The abundance of species could be reduced, possibly to extinction when the suitable habitat shrinks or disappears (Morrison et al., 2006). Identification of essential habitats has been recognized as an effective way to identify priority areas for conservation and establish area-based management plans (Yi et al., 2016), which are key components for ecosystem-based fisheries management (Vayghan et al., 2013; Tanaka and Chen, 2016).

Habitat suitability index (HSI) models, developed in the early 1980s, have become one of the most important tools in identifying the suitable habitat of organisms (e.g., Chen et al., 2010; Chang et al., 2013; Yu et al., 2016). An empirical HSI model is based on a suit of suitability indices (SIs) as a function of one or more key environmental variables to indicate the habitat suitability of target species (Brown et al., 2000; Tian et al., 2009; Chen et al., 2010). All of SIs

are then combined to develop a composite HSI score with its values ranging from 0 to 1 to represent "poor" to "good" habitat quality (Brooks, 1997). The outputs of HSI modeling in combination with geographic information systems (GIS) can provide a flexible cost-effective decision support tool for natural resource management, ecological impact assessments and ecosystem restoration (Brooks, 1997; Burgman et al., 2005).

In HSI modeling, the abundance index (AI) is generally used to establish the suitability index (SI) for each habitat variable (Tian et al., 2009; Vayghan et al., 2013; Tanaka and Chen, 2016), and a positive linear relationship was assumed to exist between SIs and AIs (e.g., Chang et al., 2013; Vayghan et al., 2013). The areas with the highest AIs are designated as optimal habitats (SI = 1.0) and the areas with the lowest AIs as poor habitats (SI = 0) (Brown et al., 2000; Tian et al., 2009; Chang et al., 2013). However, AIs tend to have highly right skewed distribution as a result of large spatial heterogeneity of fish distributions (von Szalay and Somerton, 2005; Cope and Haltuch, 2012). The occurrence of large AI values might result in underestimation of SI values for most areas. For example, for a species with AIs ranging from 0 to 5000 (ind./tow), the SI values for sampling sites with AIs of 5000, 4000, 3000, and 1000 will be equal to 1.0, 0.8, 0.6 and 0.2, respectively. However, the habitat quality for the first three sampling sites with high AI values are similar actually, and the SI value (SI = 0.2) for the last sampling site

with the AI value of 1000 is clearly underestimated. In this case, the assumption that a simple linear relationship exists between the SIs and AIs may be misleading. This problem has received little attention and rarely been considered in HSI modeling. One approach to avoid this problem is to use rescaled AIs (e.g., log-transformed AIs) in SI models, which can reduce the impact of large AI values on habitat suitability modeling.

All environmental variables in HSI modeling are usually assumed to have the same weights due to a lack of information about the relative importance of these variables (e.g., Tian et al., 2009; Vayghan et al., 2013; Yu et al., 2016). This implies that different environmental variables have the same impacts on the AIs, which is obviously unrealistic. Thus, there is a need to develop an approach that can differentiate the role of each environmental variable in the construction of HSI models. Gong et al. (2012) demonstrated that different weighting schemes would result in different spatial distributions of suitable habitats. A few studies explored the impacts of assigning weights for different environmental variables in HSI models based on relative importance of each variable identified through literature review or expert knowledge (e.g., Gong et al., 2012; Zohmann et al., 2013; Yi et al., 2016). A structured approach is needed for assigning weights to different environmental variables in habitat modeling.

Boosted regression tree (BRT) models are built using an algorithm that combines the strengths from two statistical techniques: regression trees and boosting, which can result in strong predictive performance and good descriptions of modeled relationships (Elith et al., 2008). Regression trees automatically fit interactions between variables, can handle missing values, are insensitive to outliers, and can model almost any type of predictor variables (Compton et al., 2012). Boosting improves model performance by focusing on the observations that are hard to predict with a sequential model fitting process and by adding a probabilistic component to optimize predictive performance (Compton et al., 2012). The advantage of BRTs in ecological studies is that it can cope with non-linear relationships, correlated and interacting variables, and can interpret complex relationships between species and environment (Torres et al., 2015). In addition, BRTs are unaffected by multicollinearity, missing predictor values and outliers (Elith et al., 2008). The relative importance of each environmental variable can be determined by its contribution to the total variance explained by the BRT model. In this study, BRTs were used to determine the weights of different environmental variables in the HSI modeling.

Comparative studies are needed to evaluate the performance of different approaches in constructing HSI models before they can be used in identifying critical habitats (Chen et al., 2010). We use American lobster (*Homarus americanus*) in the coastal areas of Gulf of Maine as an example to evaluate the performance of different HSI models in quantifying suitable habitats based on data collected from the Maine–New Hampshire inshore bottom trawl survey from 2000 to 2013. Our specific objectives are to compare the performance of HSI models derived using original AIs (without any transformation) versus rescaled AIs (i.e., log-transformed AIs) and evaluate the effectiveness of BRT-based weighting approach. The approach derived in this study can improve the quality of HSI modeling, leading to better definitions of suitable habitats.

2. Materials and methods

2.1. Study area

The Maine-New Hampshire inshore trawl survey is a semi-annual fishery-independent trawl survey conducted by the Maine Department of Marine Resources (DMR) in spring and fall of each year since 2000. The survey includes coastal waters up to 12 miles
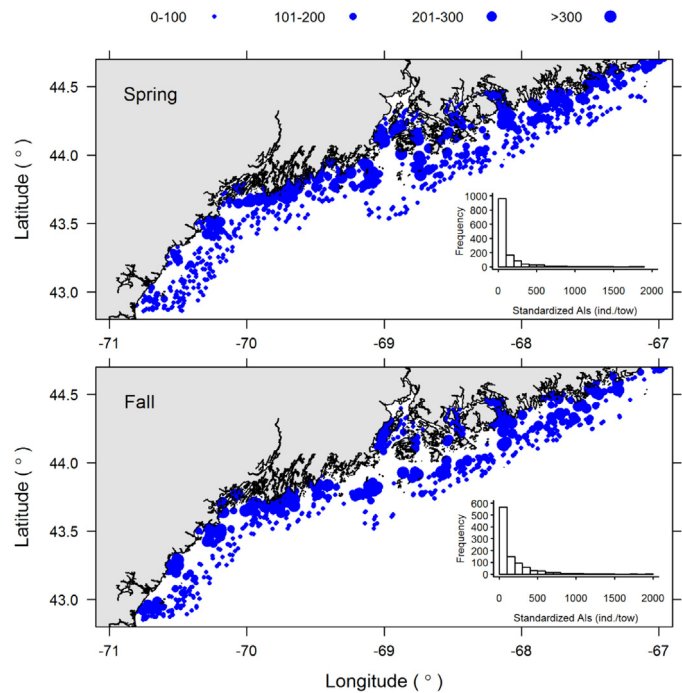

Fig. 1. Spatial distribution and frequency of standardized abundance indices (AIs) (ind./tow) for American lobster (*Homarus americanus*) based on spring and fall Maine–New Hampshire inshore bottom trawl surveys from 2000 to 2013.

offshore and follows a stratified random design by depth and regions (Sherman et al., 2005; Chen et al., 2006b). The number of stations per stratum is allocated in proportion to the area of each stratum, and a target of 115 stations is selected for sampling during each survey. The target tow duration is 20 min at 2.1–2.3 knots to cover an average swept area of ~1.48 km² per tow (Sherman et al., 2005; Tanaka and Chen, 2016). At each station, tow date, tow location (latitude and longitude), tow duration, environmental variables (e.g., bottom temperature, bottom salinity, and depth), and biological information of the catch are recorded (Sherman et al., 2005). Further details of the survey design and procedures can be found in Sherman et al. (2005, 2014). Fourteen years (2000–2013) survey data for spring (April–June) and fall (September–December) were used in this study (Fig. 1).

The standardized AI of American lobster was calculated for each station (Tanaka and Chen, 2016) as:

$$AI = \left(\frac{Count}{Tow\ duration} \right) \times 20 \qquad (1)$$

where count is the total number of lobsters caught per tow, and tow duration is towing time measured in minutes, which was standardized to 20 min at each sampling station.

The AIs standardized using Eq. (1) has a highly right skewed distribution (Fig. 1), with the mean, median and maximum AIs being 113.2, 34.7 and 1822.0 (ind./tow) in spring and being 156.9, 62.9 and 1924.2 (ind./tow) in fall, respectively. Log-transformations give strong transformation effect on distribution shape (von Szalay and Somerton, 2005) and are also necessary to normalize the data for parametric statistical analysis (Soniat and Brody, 1988). Hence, the natural logarithmic transformation was applied to reduce the skewness, with a value of 1 being added to the AIs (hereafter lnAI) to account for zero catch values (Howell and Kobayashi, 2006; Mugo et al., 2010) based on the preliminary assessment of alternate transformations.