Towards a rigorous methodology for verification and accuracy assessment of qualitative wildlife habitat models.

Guidelines for qualitative habitat models, such as habitat suitability index models, emphasize the importance of verification and validation exercises as part of the model development and evaluation process. However, guidelines for effective verification procedures are lacking. Several factors associated with verification, including random versus targeted sampling, sample size requirements, analysis scale, field methods, personnel qualifications, and methods of comparing model predictions to field data can dramatically affect study design and interpretation of results. In this paper I discuss key issues associated with each of these factors and provide methods and guidance to assist others in the development of effective verification projects. I focus on verification using field ratings by species experts but these guidelines can be adapted to other types of verification data, such as index or sign data.

Key words: habitat model, verification, accuracy assessment, expert opinion, northern goshawk

Introduction

Qualitative wildlife habitat habitat models are commonly used to aid in management and conservation actions, including species reintroductions (e.g. Olsson and Rogers 2009), inventory of rare species (Cameron and Neily 2008), habitat risk assessment (Grech and Marsh 2008), species invasions (Williams et al. 2008), and land management (Marcot et al. 2001). Probably the most common type of qualitative model is the Habitat Suitability Index (HSI) model (United States Fish and Wildlife Service 1981). Simple look-up tables, which assign a suitability rating to specific combination of environmental variable conditions, such as forest type and seral stage, are another example (Resource Inventory Standards Committee 1999). Bayesian belief networks (BBNs) are also emerging as a popular tool for a suite of ecological modeling situations, including habitat modeling (Marcot et al. 2006). These types of habitat models are all considered to be qualitative because some aspect of the model is driven by expert judgement. In the case of HSIs and BBNs, expert judgement is often applied to three specific aspects of the model: 1) selection of environmental variables to include in the model, 2) development of suitability ratings curves for each environmental variables, and 3) combination of those variable ratings via a mathematical equation, or set of beliefs, to produce an overall model rating.

Guidelines for qualitative wildlife habitat models, notably HSI models, emphasize the importance of model testing via verification and validation exercises (e.g. Brooks 1997; Roloff and Kernohan 1999). And, by naïve practioners, often interchangeably. The exact definition of, or differences between, verification and validation are often unclear and sometimes the terms are used interchangeably. I define verification as a test of model performance using independent samples of indirect sign (e.g. scat for a foraging model) or field ratings by a species expert. I define validation as a test of model performance using independent samples of actual density estimates of, or frequency of use by, the species of interest. Although preferred, validation exercises may not be possible within the timeframes or budgets of planning or management initiatives, or due to low occurrence or detectability of the species of interest, and verification is the only feasible type of model testing available. Also, if validation data are anticipated to be available, one may want to forego a qualitative model and use the empirical data to derive a quantitative model, such as a resource selection function (Manly et al. 2002).

Although guidance for validation exercises have been developed (Roloff and Kernohan 1999), guidelines for developing and implementing effective verification projects are lacking. Several factors associated with verification, including random versus targeted sampling, sample size requirements, analysis scale, field methods, personnel qualifications, and analyses for comparing model predictions to field data can dramatically affect study design and interpretation of results. In this paper I discuss key issues associated with each of these aspects and provide methods and recommendations to assist others in the development of effective verification projects. I focus on verification using field ratings by species experts but these guidelines can be adapted to other types of verification data, such as index or sign data. Using the approach outlined here a typical verification project could be completed in approximately one month, including study design and sample plan development, field sampling, and data analysis and reporting.

Goals of Verification

Generally, there are two primary goals associated with verification exercises. The first is to provide a quantitative estimate of model performance. This involves comparing

model predictions to an independent sample of field data, such as suitability ratings assessed by a species expert. The second general goal of verification is to collect environmental data at each sample site to help evaluate and refine the model. More specifically, the environmental data can be used to 1) quantify errors and biases of GIS layers used in the model, 2) verify assumed relationships between GIS variables, used in the model as proxies, and primary variables or conditions in the field, and 3) provide a basis for adjusting variable rating curves in the model.

Accuracy Assessment versus Model Evaluation

Model performance can be assessed from two perspectives depending on whether the primary goal is 1) to assess the overall accuracy of the model predictions across a specific study area, which I define in this paper as accuracy assessment, or 2) to assess how well the model performs with respect to specific combinations of environmental conditions, which I define as model evaluation. To be clear with terminology through the remainder of this document, I explicitly use the terms accuracy assessment and model evaluation when discussing factors where implications differ between the two types of model assessment perspectives and I use the term verification when discussing factors that relate to both perspectives.

Depending on which of these model assessment perspectives is the primary objective, a fundamentally different sampling scheme should be used, and different estimates of model performance will result. Where the primary goal of model verification is accuracy assessment, the sampling scheme should select a sample ~~of habitat types~~ that is representative of the proportion of model predictions across the study area. Where the primary goal is model evaluation, the sampling design may be stratified, or arbitrarily assigned, to obtain a certain proportion of samples across a range of combinations of environmental conditions, irrespective of their occurrence in the study area.

To illustrate the potential differences of these two perspectives, consider a 2-class habitat model where coniferous=suitable, deciduous=unsuitable and where the model predicted 20% the study area was suitable and 80% was unsuitable. Assume true model accuracy is 50% for suitable and 100% for unsuitable. If accuracy assessment was the

primary objective and a proportional sample was drawn for 50 field samples, the number of samples in suitable and unsuitable areas would be approximately 10 and 40, respectively, and the corresponding model performance score would be 10/50*0.5+40/50*1.0=90%. If model evaluation was the primary objective, a reasonable sample approach may be to have half of the samples in each habitat class and the model performance score would be 25/50*0.5+25/50*1.0=75%. In this simple example the model evaluation sample could be adjusted to provide an overall accuracy assessment using a simple weighted average of model predictions. However, with a real model that contained several continuous and/or multi-class categorical environmental variables it may be difficult or impossible to adjust the overall accuracy estimate using weighted averages.

Sampling Design

The sampling schemes normally associated with accuracy assessment and model evaluation objectives are random and targeted, respectively. By targeted, or purposive, sampling I mean sampling that examines specific combinations of underlying environmental data. Many combinations of environmental data conditions occur infrequently and are unlikely to be sampled under a random sampling design. Traditionally, field calibration and verification of habitat models has used targeted sampling to ensure that a broad range of base data conditions are examined. This can often result in a strongly biased sample with respect to the proportional occurrence of those conditions across the project area. Rare conditions tend to be over sampled relative to their proportional occurrence and common conditions tend to be under sampled (Congalton and Green 2009). A fundamental requirement of obtaining a statistically unbiased estimate of overall model accuracy is that a random sample of model outputs is taken. Sampling random locations within stratified environmental conditions does not address this bias unless the sample is proportional to the extent of those conditions. Where project objectives include both accuracy assessment and model evaluation the appropriate approach is to draw a random sample to meet the accuracy assessment objectives, and supplement that with targeted sampling to meet model evaluation objectives.

In a large study area, or where access is poor, a completely random sample could result in exorbitant travel costs among widely dispersed sites. A clustered, random design may be employed to reduce travel costs among sample units, however, care must be taken to ensure the clustering parameters do not bias either the random aspect or the representativeness of the sample. If sample units are not far enough apart

Often a clustered-random sample is desirable for logistic reasons and is acceptable so long as the clustering does not bias the representativeness of the sample. For example, in a large study area, or where access is poor, a completely random sample could result in exorbitant travel costs among widely dispersed sites. Choosing one random point first, as a cluster centre, then selecting additional random points within a specified distance, can result in much more efficient field sampling and still maintain the random sample design required to assess the overall accuracy of a model.

Defining Project Scope

Ideally, the entire study area should be available for accuracy assessment/model evaluation. In some cases the extent may need to be reduced for logistic reasons, such as access (e.g. only areas within 5 km of a road may be considered for verification sampling). In those cases the area of inference of the verification becomes limited and that limitation should be explicitly quantified and stated (i.e. sampled versus available proportions of key strata within the study area).

In other cases, where a large part of the study area is unsuitable habitat that is classified with high certainty, that part of the project area should generally be excluded so that verification focuses on potentially suitable habitat. For example, consider a nesting habitat model for a species of songbird inhabiting forest remnants in a landscape consisting of 90% cropland, and where it is known with high certainty that the species does not nest in cropland. If a random sample was drawn and 90% of the samples were in cropland, the accuracy score of the model would likely be high, but largely uninformative. Assuming the habitat model was developed primarily to assess relative nesting habitat quality with respect to forest remnant conditions, the forest remnants should be the focus of the assessment and cropland should be excluded.

Analysis Scale and Sample Unit Design

Historically, habitat models were often applied using one type of input data – a polygon-based vegetation map. With that type of map a polygon was an obvious sample unit and field assessment was relatively straightforward – single or multiple plots or transects could be established within polygons.

Now, with widely accessible GIS, powerful computers, and a suite of digital environmental data, often available in both vector and raster formats (e.g. vegetation, soils, elevation, slope, aspect, and soil moisture), even a simple overlay of available input data can in a very complex base map. If spatial variables, such as distance from edge, fragmentation metrics, or neighbourhood analysis, are included the complexity of map overlays increases even more, to the point where the functional map unit with unique environmental conditions essentially becomes an individual, or small number of, pixels. This creates a problem for field verification because spatial accuracy can become a substantial issue at the 25-100 m size of pixels commonly used in habitat models.

Spatial accuracy issues result from at least three sources. 1) Base habitat map layers have limited spatial accuracy associated with them. The spatial accuracy of vector-based map data (line and polygon data such as roads, streams, and vegetation maps) is typically in the order of 25-50 m for 1:20,000 scale products and 50-100 m for 1:50,000 scale products. Raster-based data, such as digital elevation models (DEMs) often have specified accuracy of 25 m at 1:20,000 scale and 50 m at 1:50,000 scale. 2) Conversion of vector data to raster data results in spatial accuracy loss up to half the pixel size. For example, if the boundary of a vegetation map polygon falls near the middle of 100 m pixel, when it is converted to a raster file the entire pixel will be classified either as the reference polygon or its neighbour, according to a majority rule, effectively shifting the edge by approximately 50 m. 3) GPS errors typically range from 5-20 m, although differential correction can virtually eliminate these errors.

In addition to spatial accuracy issues, heterogeneity associated with thematic map accuracy can become an issue when sampling at a finer resolution than the map unit. For example, values of attributes associated with a polygon in a vegetation map are based on an average value or majority rule across the polygon. If only a portion of the original polygon is surveyed, because overlays of multiple maps result in subdivision of the

vegetation map polygon, the attribute value for that portion of the subdivided polygon could be inaccurate, even though the attribute value was accurate for the original polygon. For example, if polygon has an accurate canopy closure value of 60%, and that polygon is subdivided into four smaller polygons the true canopy closure in each of the new polygons could be 40%, 60%, 70% and 70%.

The approach I recommend for dealing with this spatial accuracy issue is similar to a "small area" sampling design developed by Moon et al (2005) for assessing accuracy of ecosystem maps. This method uses a sample unit that is several times larger than both the resolution of the map units (typically a pixel) and the estimated spatial accuracy of the least accurate input data layer. Several subsamples are then taken within the sample unit, and field ratings and model predictions from the subsamples are compared aspatially within the sample unit (i.e. without maintaining subsample level comparison). For example, consider a model with a rating scheme of suitable (1) and unsuitable (0) habitat, and a sample unit with four subsamples having field and model ratings of 0/1, 1/1, 0/0 and 1/0. In this case the sample unit level accuracy would be 100% because both model and field ratings contain two 0's and two 1's. At the subsample level the accuracy would be 50% because two out of four of the subsamples have ratings that correspond. The small area sampling approach assumes that both the field and model subsamples occur within the sample unit but, due to potential spatial accuracy errors, not necessarily at the exact subsample location that was targeted.

Another important consideration for selecting sample unit size is that it should be smaller than the size of typical management units (e.g. forest cutblocks). This is because interpretations related to the sample unit size cannot be confidently applied to a smaller resolution. For example, a sample unit size of 10 ha may be appropriate for assessing habitat model accuracy relative to cutblocks that average 20 ha in size, but not to within-cutblock forest retention patches that average 2 ha. I view small area sampling as a surrogate for the traditional stand-scale polygon sampling approach, but in cases where polygons are not available, and generally recommend sample unit sizes of 5-30 ha.

Circular or square sample unit shapes are preferred to account for the spatial accuracy problem. A variety of systematic or random sampling designs may be used to establish subsamples within the sample units. One efficient design is to use a circular

sample unit with subsamples along a triangular transect within the sample unit (Figure 1). This design provides reasonably good coverage of the sample unit and it is efficient to sample because the surveyor completes the transect near where they started. Using a clustered random sampling design with a 5 km radius bounding sample units within the cluster, and a circular 10 ha sample unit surveyed using a 900 m triangular transect with subsample plots every 100 m, two or three sample units can be surveyed in a day by one field crew.

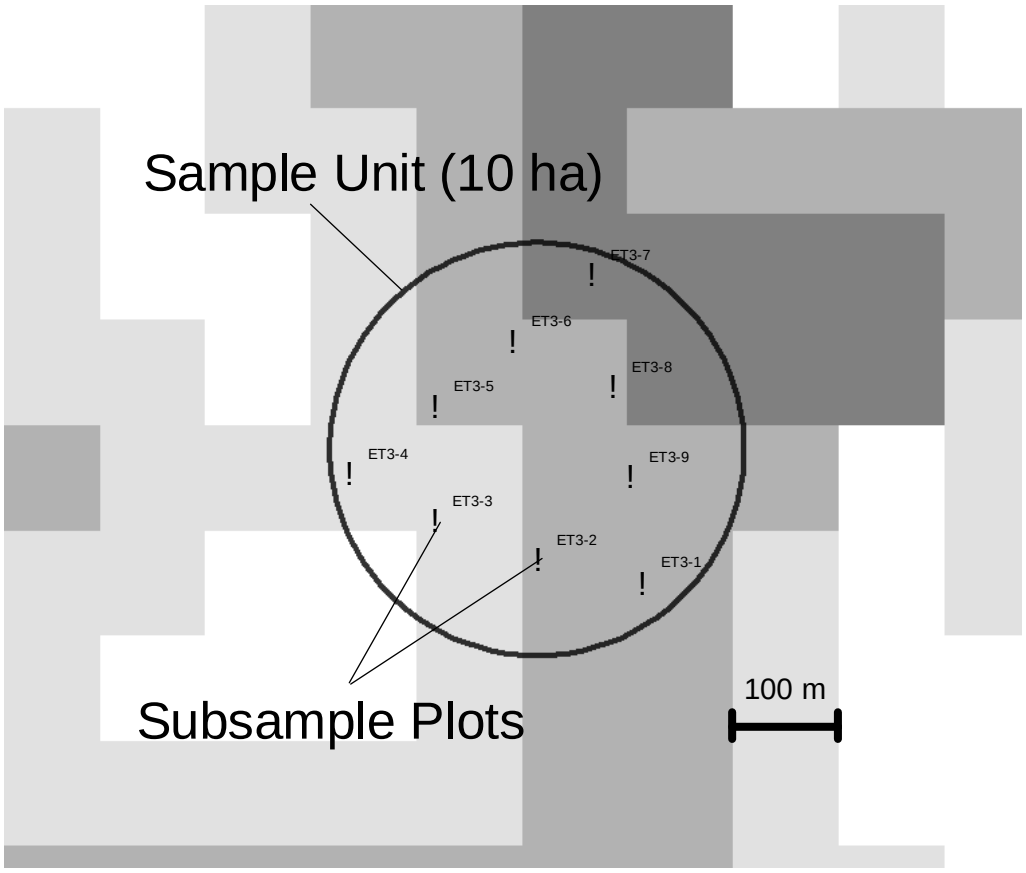

Figure 1. Sample unit design for the small area sampling approach. Shaded raster data is habitat suitability index model output classed into four categories. Heterogeneity of the habitat data, coupled with uncertainty of positional accuracy at the 100 m pixel scale, limits the use of sampling designs that target specific habitat types.

Setting a Priori Accuracy Targets

In most circumstances it is useful to establish an accuracy target that can be used as a benchmark to assess whether the model outputs are acceptable for specific management

or planning purposes or to determine whether model revisions are required. Acceptable accuracy levels should be established before field verification is conducted to avoid biasing that decision by the field results.

Setting accuracy targets is a subjective exercise and will depend on the intended uses of the model outputs and comfort levels of the biologists, mangers or planners using those outputs. The appropriateness of accuracy targets also varies depending whether continuous or categorical scoring methods are used, and for categorical schemes, how many categories there are. Accuracy scores tend to be higher using continuous schemes than categorical schemes. For categorical schemes, accuracy scores will be higher the fewer the categories there are. For accuracy assessments of one environmental variable mapped using remotely data, the longstanding, albeit arbitrary, standard has been 85% (Congalton and Green 2009). When multiple variables are considered, as is usually the case with habitat models, classification errors typically compound in a multiplicative manner (Congalton and Green 2009). As an example, a composite map of four environmental variables with thematic accuracy of 90% for each variable would be expected to have an overall accuracy of 0.9^4=66%. Unless a habitat model collapses some of the categories for each variable, this compound accuracy represents the highest potential accuracy the model could have. This factor alone emphasizes the importance of parsimonious model construction. One example of an accuracy target for a model using multiple variables is 65% for predictive ecosystem mapping in British Columbia (Meidinger 2003).

Sample Size Requirements

The number of samples required to meet a specified confidence interval for accuracy assessment can be estimated using the conventional sample size formula:

Sample size = (t-value)^2 x (standard deviation)^2 / (acceptable error)^2

Approximate sample sizes across a range of typical confidence levels, standard deviations, and acceptable errors are provided in Table 1. The t-value is based on the *a priori* confidence level at n-1 degrees of freedom, where the degrees of freedom are a best guess at the required sample size. For example, in the first cell in Table XX, using a 90% confidence level and estimated sample size of 30, $t_{0.10(2),30}$=1.697. Thankfully, the

degrees of freedom have relatively little influence on the sample size estimates. Again, for the first cell in Table 1, doubling the degrees of freedom to 60 changes the sample size estimate by less than one. The standard deviation is that of the anticipated sample, and again a best guess must be input into the equation. To account for uncertainty of the standard deviation it is often useful to consider a range of values, such as in Table 1. Acceptable error is the maximum difference that the sample mean can deviate from the true population mean before you call the difference significant. This is a choice that should be made as part of the *a priori* decision about the acceptable accuracy target will be. Typical values range from 5 to 10%, with smaller values being more conservative.

Table 1. Approximate sample size requirements for small area sampling accuracy assessment.

Confidence level	Sample Error	Sample Variance (SD)			
		0.150	0.175	0.200	0.225
0.9	0.05	26	35	46	59
0.9	0.07	13	18	24	30
0.9	0.09	8	11	14	18
0.8	0.05	16	22	28	36
0.8	0.07	8	11	14	18
0.8	0.09	5	7	9	11

To address the limitations of guessing at initial sample sizes and standard deviations to seed the sample size equation, above, Moon et al. (2005) recommended a staged field sample, where a major portion of the estimated sample size is surveyed, then the standard deviation of that sample is calculated and used to rerun the sample size estimate and determine how many additional samples are required to meet the desired confidence level. The remaining samples must then be randomly selected again across the entire project area. Although staged sampling is a preferred approach to ensure confidence levels are met, it has potentially serious logistical limitations of requiring a break during field work and, often, travel back to regions of the study area near where field sampling was already conducted. In large study areas, or areas with poor access requiring aerial transportation, this may be prohibitively expensive. As an alternative to staged sampling, it may be more cost effective to simply survey extra samples to ensure confidence levels are met.

In addition to meeting statistical requirements, consideration should also be given to ensure the sample size is adequate to provide a representative sample. Types of representation should include the range of model outputs, the underlying environmental conditions, and geographic representation across the study area. Generally, a minimum of 50 samples should be considered to meet basic objectives of representation (Congalton and Green 2009). Again, it is often useful to supplement the random sample with targeted samples to ensure certain habitat variable combinations are met for model evaluation (but those additional samples should not be included for accuracy scoring).

Sample size is normally determined for the entire study area and inferences from that sample cannot be extrapolated to portions of, or strata within, the study area, such as a certain biogeoclimatic zones, with statistical confidence. If accuracy scores are desired at a finer level than the study area, sample sizes need to be calculated for each stratum or subregion of interest. Because this can result in large sample sizes most accuracy assessments will, at least initially, be for the entire study area.

Personnel Qualifications and Calibration

I have already noted that using expert-opinion field ratings is generally a less desirable type of verification than using more objective data, such as an index of use of the species of interest, and it is important that experienced personnel and a rigorous rating procedure are used to minimize bias and variation associated with expert-opinion ratings.

The following should be considered minimum qualifications for personnel conducting the field assessments. 1) Field personnel should have at least two seasons of field-related experience with the species-habitat relationship of interest. No matter how familiar a biologist is with the literature, direct field experience is necessary to be able to accurately and reliably estimate habitat suitability. 2) Field personnel should be familiar with local habitat use patterns, as well as the broader patterns of habitat use of the species across its range as documented in the literature. I have seen cases where biologists with considerable experience working with a species in one location did a poor job of assessing habitat suitability in a different location because they were not familiar with the range of habitat use patterns the species exhibited across their range.

Whenever possible, the person or team who developed the model should be involved in the field verification. The reason for this is that the verification exercise should be an assessment of how well a set of beliefs and assumptions, expressed as a model, reflect inferred habitat quality in the field using those same beliefs and assumptions.

When multiple observers are used in the field, calibration is required to minimize observer bias and reduce variation in field ratings. Three formal types of calibration exercises that should be conducted are: 1) review and discussion of the model assumptions and structure, local habitat use data, and broader habitat use patterns, 2) discussion and definition of typical suitability ratings found across the range of environmental variable conditions (essentially discussion of criteria in Appendix 1), and 3) calibrations surveys by all personnel at the same plots to rationalize the implementation of criteria from 1 and 2 in the field, and to calibrate to similar field ratings.

Reducing Model to Variables Assessable in the Field

Many habitat models include spatial variables, such as distance to edge, or larger scale variables, such as patch size, that field personnel may not be able to perceive and assess effectively in the field. Generally, when this occurs the field ratings should only be compared to a reduced version of the model that excludes any variables personnel are not able to assess in the field.

In some cases the appropriateness of excluding certain variables will not be immediately clear and requires thoughtful review of the rationale for including the variable in the model and the type of information used to parameterize the ratings for that variable. For example, consider distance to edge, which I used in a nesting habitat model for Northern Goshawks (Mahon et al. 2008). The rationale for using this variable was that goshawks avoided locating nests near edges and I was able to parameterize a rating curve using 62 nest sites showing strong avoidance 0-100 m from an edge and moderate avoidance 100-200 m from an edge. Edge was excluded from the model outputs compared to field verification ratings against for two reasons. First, was that the observed use pattern did not correspond to obvious stand structure that could be assessed

in the field. Possibly the use pattern was behaviourally driven, such as to reduce predation risk or inter-specific nest competition that might be higher near edges. Second, it was difficult for field personnel to perceive edges greater than 30 m away due to thick forest vegetation. In other circumstances a model may include distance from edge to account for a more direct habitat condition, such as reduced canopy cover due to windthrow. Limitations to perceiving distances to edges would not be an issue because the primary condition of interest is change in canopy closure. In that case it would be appropriate to include distance from edge in the model for comparison to field results.

Field Sampling

Rating Habitat Quality

A key aspect of successfully using expert opinion ratings to verify a habitat model is to formalize and standardize the assumptions and criteria used to define ratings along a specified scale. This exercise is critical to facilitating consistent ratings by the same individuals as well as among individuals. Usually the foundation of this knowledge can be transferred from the rationale used to develop the habitat model. In many cases, however, the variables used in a model may be surrogates of the primary variables of interest. For example, amount of shrub cover may be a primary variable affecting habitat quality for a certain species of bird. If shrub cover is not available in a GIS database, however, a surrogate, such as canopy closure, may be used in the habitat model. Criteria for field ratings should be based on the variables most directly relating to habitat quality. As part of this exercise biologists also need to identify relationships among habitat attributes in contributing to overall habitat quality. For example, do certain variables or conditions act in compensatory or non-compensatory ways in contributing to overall suitability? Key ratings assumptions should be formally documented such as the example in Appendix 1. Again, the purpose of this exercise is to formalize and standardize rating assumptions among field personnel. It should not be interpreted as a cookbook that non-specialized personnel could use to derive a rating.

Two types of rating schemes can be used in the field: 1) ordinal ratings (e.g. nil, low, moderate, high) or 2) continuous numerical ratings (e.g. 0 – 1). Generally, the rating scheme used in the field should match the type of outputs from the habitat model.

Ordinal ratings are usually associated with simple decision-tree or look-up table type models (e.g. ecosystem by seral stage table). Continuous numerical ratings usually result from any model where suitability ratings for multiple variables are combined via a mathematical equation (e.g. habitat suitability index models).

Ordinal ratings are, perhaps, the most intuitive scheme. In British Columbia, standards have been developed for qualitative, ordinal wildlife habitat ratings applied to ecological mapping projects (Resource Inventory Standards Committee 1999). Many aspects of those guidelines are applicable to habitat model verification and biologists undertaking a verification project would benefit from reviewing those standards. The number of habitat classes used in the field will generally be the same as the number of classes derived from the model, however, occasionally it may be useful to subdivide a class in the field to support model refinement. One limitation with using ordinal ratings in the field is that estimated suitability often falls near a class break. Field personnel must ultimately select the rating class that seems most appropriate, however, indicating the end of the class the suitability falls in may also help with model evaluation and refinement.

A continuous rating scheme may seem more daunting to use initially, but has the advantage of not being constrained by arbitrary boundaries of categorical bins. The approach I use when deciding on a rating is to first decide on the appropriate quartile (e.g. 75-100), then decide on a more precise rating based on whether suitability is on the low, middle or high end of that quartile. Measurement precision should also be specified as part of the study design. Generally, I have found that five percent increments are a satisfactory increment to use. The exception to this is when ratings may subsequently be categorized subsequent purposes. When that may occur I add a rule that ratings cannot occur on class breaks (e.g. 25, 50, 75 for quartiles), and should be reduced or increased by one value (e.g. 74 or 76), as appropriate, to avoid the class break. When in doubt I recommend using finer, rather than broader, increments to provide the maximum flexibility for accurately estimating suitability along a continuous scale. In doing so, however, it is important to note that finer measurement precision does not imply finer precision or accuracy in a statistical sense.

It is a fundamental assumption that field estimates are, on average, accurate, but it is also important to recognize there is unknown, but not insignificant, variance associated with field ratings. That variance results from a number of sources including observer bias, random individual variation, and imperfect habitat perception. While we might be able to quantify components of that variance (e.g. observer bias) through study design and extra field work, it is probably impossible to quantify all of the variation associated with field ratings.

Another important aspect of field ratings is that the assessment must be 'blind'. That is field personnel should not know the map predictions for the area they are assessing. Several studies have shown that bias results if observers know what the predicted classifications are ( ).

Measuring Environmental Variables

There are normally three objectives of measuring environmental variables in the field: 1) to verify assumed relationships between GIS variables used in the model as surrogates and the primary variables or conditions of interest in the field, 2) to quantify errors and biases in the GIS data, and 3) to develop or strengthen relationships between suitability and environmental conditions that provide a basis for revising variable rating curves in the model. The relative importance of these objectives will influence the number and type of environmental variables examined and the level of detail at which they are assessed. Under time or budget constrained projects I generally recommend an extensive versus intensive sampling approach that favours surveying more sample units, as opposed to surveying fewer sample units and measuring environmental attributes more precisely. For example, a visual estimate of tree species composition and one measurement of stand height may be adequate to assess the general accuracy of forest cover data, and could be conducted in a fraction of the time it takes to conduct a fixed radius tree mensuration plot.

Removing Samples with Incongruent Field and GIS Conditions

In some cases field samples may occur on sites with recent anthropogenic or natural disturbance events (e.g. logging, road building, wildlife, windthrow, landslides)

that have not been updated in the base GIS data used in the habitat model. Generally, using these samples is inappropriate for accuracy assessment or model evaluation because the field and model ratings are based on incongruous underlying habitat data and the samples should be culled prior to analysis.

A related issue is when environmental conditions at a plot differ between the field measurements and GIS data as a result of thematic mapping errors or positional accuracy errors. For accuracy assessment it is important to include these samples in the analysis because the accuracy of underlying data is a key factor affecting overall accuracy of model predictions. For model evaluation, it is useful to run the analysis both with and without samples with disparate environmental conditions. To evaluate true model performance (i.e. rating curves for individual variables and the way ratings from multiple variables are combined), only samples with agreement in the environmental conditions should be used. However, the difference in model accuracy between the full and reduced samples provides an indication of how robust the model is to underlying environmental data errors.

Accuracy Scoring Approaches

Accuracy, by definition, is how close an estimate is to a true value. For assessing the accuracy of habitat models we assume that field ratings are true values and that model outputs are estimates. Different scoring methods are required for ordinal and continuous rating schemes. Simple scoring procedures are provided below for each method. These methods focus on quantifying the degree of between field ratings and model predictions (i.e. effect size) in a form that can be compared to an *a priori* accuracy target.

As previously mentioned, accuracy scores can be derived at the resolution of subsample, sample unit, and study area by maintaining or dissolving spatial dependencies of the model to field rating comparison at each resolution. For example, at the subsample unit resolution model predictions and field observations are compared for the same subsample. At the sample unit resolution model predictions and field observations for the subsamples are compared aspatially. Again, the purpose of this is to address issues of spatial accuracy at the subsample level. Although we have poor confidence that underlying data corresponds at the subsample level, we have much higher confidence at

the sample unit resolution. Another way to put it is that we are confident that both the field and model subsamples occur within the sample unit, but we are not confident of exactly where they are within it. At the study area resolution all spatial dependencies are dissolved. Although I recommend that the primary focus should be on the sample unit scores, it is often informative to present accuracy scores at all three resolutions.

Scoring for Ordinal Rating Schemes

Accuracy scoring procedures for categorical or ordinal mapping products have been well developed in the geomatic disciplines (Congalton and Green 2009) and these methods can be applied to assessing categorical wildlife habitat models (Fielding and Bell 1997). It is important to note that this approach considers accuracy as the proportion of times the model predictions correspond to the field observations, rather than the usual statistical definition of accuracy, which is how close an estimate is to the true value.

The standard method for summarizing and scoring map or model predictions against field observations is using a confusion matrix. A confusion matrix is a contingency table where field observations and model predictions are listed as column and row headings, respectively, and the numbers of samples corresponding to each combination are tallied in the cells (Table 2). Overall accuracy is simply the proportion of corresponding observations (sum of the diagonal cells) relative to the number of samples. A confusion matrix is also useful for assessing bias in model predictions. The proportion of observations in cells above the diagonal cells represent false positives (i.e where the model is overestimating habitat quality) and the proportion of observations in cells below the diagonal represent false negatives (i.e. where the model is underestimating habitat quality). See Fielding and Bell (1997) and Congalton and Green (2009) for a comprehensive description of several additional measures of accuracy and errors that can be estimated from a confusion matrix. The Kappa statistic and 'fuzzy' accuracy assessment are two somewhat more sophisticated analysis techniques that are commonly used for categorical accuracy assessment (Congalton and Green 2009). The Kappa statistic is a corrected version of the simple accuracy statistic, above, which takes to into account the likelihood of chance agreement within cells (Congalton et al. 1983). Fuzzy accuracy assessment gives full or partial score to map predictions that are within a certain number of classes (normally one) of the field data (Gopal and Woodcock 1994).

Table 2. An example of a confusion matrix used to estimate the accuracy of categorical model data (after Congalton and Green 2009).

		Field Data				
		Nil	Low	Moderate	High	Total
Model Data	Nil	65	4	22	24	115
	Low	6	81	5	8	100
	Moderate	0	11	85	19	115
	High	4	7	3	90	104
	Total	75	103	115	141	434

Overall Accuracy =	(65+81+85+90) / 434 = 74%
False Positives =	(4+22+24+5+8+19) / 434 = 19%
False negatives =	(6+0+4+11+7+3) / 434 = 7%

One of the requirements for developing a confusion matrix is that paired samples of field observations and model predictions for the same sites are available. Using the small area sampling approach outlined above, the original pairs of the model and field data at the subsample plots are not maintained and a normal confusion matrix cannot be developed. However, the overall accuracy can still be calculated based on the number of corresponding ratings. Examples of accuracy scoring at the subsample and sample unit levels for the same data are shown in Tables 5 and 6. The overall accuracy score for the model is estimated as the mean of all the sample unit scores, and confidence intervals can be calculated using the mean and variance from the sample unit scores.

Two approaches can be used to approximate model errors for the sample unit. One is to create a confusion matrix using all of the subsample level data (where pairing is maintained) and assume the ratio of false positives and false negatives is similar for the sample unit level data. The second approach is to independently order the field observations and model predictions within each sample unit and use the new rank order pairs to build a confusion matrix. Using the small area sampling approach, the overall accuracy of the model is estimated by the mean of the sample unit scores and the variance of the scores can be used to derive confidence intervals.

Table 3. Example accuracy scoring using a four class rating scheme at the subsample level, which maintains explicit plot-level comparisons of model predictions and field ratings.

SampleID	Field Rating	Model Rating	Field Class	Model Class	Subsample Score
T1-1	0.90	1.00	H	H	1
T1-2	0.90	1.00	H	H	1
T1-3	0.00	1.00	N	H	0
T1-4	1.00	0.00	H	N	0
T1-5	0.55	0.45	M	L	0
T1-6	0.70	0.675	M	M	1
T1-7	0.45	0.675	L	M	0
T1-8	0.30	0.45	L	L	1
T1-9	0.45	1.00	L	H	0
				Total	4/9=44%

Table 4. Example accuracy scoring summary for the same data from Table 5, at the sample unit level, assessing the correspondence of field and model data aspatially within the sample unit.

Rating Class	Number of corresponding model and field ratings
H	3
M	2
L	2
N	1
	Score = 8/9 = 89%

Scoring for Continuous Numerical Rating Schemes

While scoring techniques are well developed for categorical data, methods for comparing the accuracy of continuous data are poorly developed, at least by comparison. The prevalent approach has been to simply categorize the continuous data into bins. For example using 0-1 model outputs from a logistic regression equation and classifying values >0.5 as presence, and values <0.5 as absence (e.g. Johnson . There are several unsatisfactory results of this approach. 1) Substantial information is lost, notably precision of the estimates, by collapsing a continuous value into a categorical bin (Boone and Krohn 2002). 2) Classifying continuous values creates arbitrary ‘breaks’ across the range

For continuous ratings on a bounded scale (e.g. 0-100) I recommend using the following equation to score individual samples:

Accuracy = 100[1] - |model rating – field rating|

When comparing a model prediction to a field score it is intuitive that the difference between the values is a measure of the degree of disagreement, or inaccuracy, between them. One minus that difference represents the degree of agreement, or accuracy, between them (Figure 2). To help understand this idea, consider suitability as a continuum from nil to optimal conditions. Any point along that continuum represents both how much better suitability is than nil, and how much worse suitability is than optimal. In this context accuracy can be viewed as how much of the continuum is outside of the two rating estimates. In the example in Figure 2, both the field rating and model prediction are in agreement that suitability is greater than 75 and also that suitability is less than 90. 100-|90-75| = 85% accuracy.

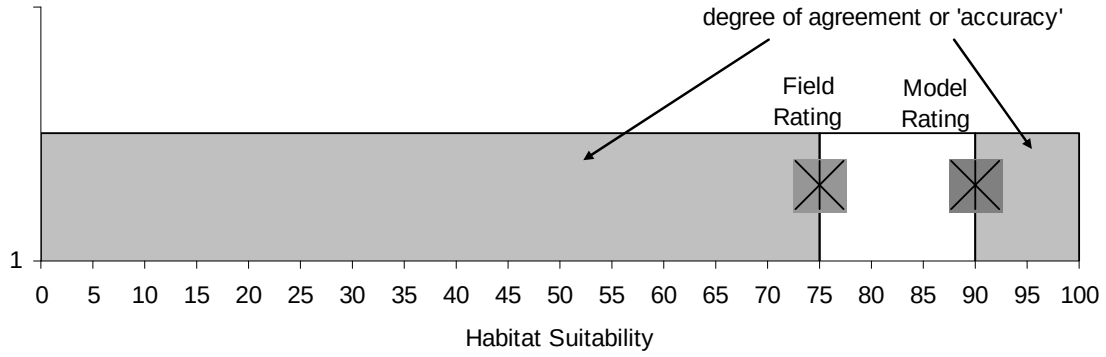

Figure 2. Rationale for accuracy scoring approach using a continuous rating scheme. Both the model and field ratings are in agreement that suitability is greater than 75 and less than 90. 75+10=85% accuracy.

To derive sample unit scores, subsample ratings are considered aspatially within each sample unit, similar to the ordinal rating scoring example. Field and model ratings should be ranked independently and then the re-ordered rank pairs are scored using the above equation. Example scoring summaries for one sample unit are shown below in Tables 4 and 5 at sub-sample and sample unit levels, respectively. Again, the mean of

[1] Assuming the scale is standardized to 100

sample unit scores can be used to estimate overall model score. Model bias can be examined by summarizing the average differences of the samples units. If the differences are calculated as model predictions minus field ratings, a negative difference indicates the model is overestimating suitability. If the differences are positive the model is underestimating suitability. It is also informative to display differences in a histogram such as Figure XX. The plot and statistics of a correlation between field ratings and model predictions may also be useful.

Table 5. Example accuracy scoring for one sample unit using a continuous rating scheme at the subsample level, which maintains explicit plot-level comparisons of model predictions and field ratings. Accuracy was calculated as 1 - |model prediction – field rating|.

Sample ID	Field Rating	Model Rating	Difference	Accuracy Score
T1-1	0.90	1.00	0.100	0.900
T1-2	0.90	1.00	0.100	0.900
T1-3	0.00	1.00	1.000	0.000
T1-4	1.00	0.00	1.000	0.000
T1-5	0.55	0.45	0.100	0.900
T1-6	0.70	0.675	0.025	0.975
T1-7	0.45	0.675	0.225	0.775
T1-8	0.30	0.45	0.150	0.850
T1-9	0.45	1.00	0.550	0.450
		Average	0.361	0.639

Table 6. Example accuracy scoring for one sample unit using a continuous rating scheme at the sample unit level. Prior to scoring, model predictions and field ratings were independently sorted and similar rank pair scores were calculated as 1 - |model prediction – field rating|.

Rank Order	Field Rating	Model Rating	Difference	Accuracy Score
1	1.00	1.00	0.000	1.000
2	0.90	1.00	0.100	0.900
3	0.90	1.00	0.100	0.900
4	0.70	1.00	0.300	0.700
5	0.55	0.675	0.125	0.875
6	0.45	0.675	0.225	0.775
7	0.45	0.45	0.000	1.000
8	0.30	0.45	0.150	0.850
9	0	0	0.000	1.000
		Average	0.111	0.889

An Example - Accuracy Assessment of a Northern Goshawk Nesting Habitat Suitability Index Model

An example verification project using these protocols will be included in the journal submission, but is not included here, pending permission of the Northern Goshawk Recovery Team to distribute and publish the results.

Acknowledgements

Funding for this paper was provided by the Forest Science Program through the Bulkley Valley Centre for Natural Resources Management and Research. Data used in the example was used with permission from the British Columbia Coastal Northern Goshawk Recovery Team.

Literature Cited

Brooks, R.P. 1997. Improving habitat suitability index models. Wildl. Soc. Bull. 25: 163-167.

Cameron, R.P. and T. Neily. 2008. Heuristic model for identifying the habitats of Eriderma peicellatum and other rare cyanolichens in Nova Scotia, Canada. The Bryologist111(4):650-658.

Congalton, R.G., R.G. Oderwald and R.A. Mead. 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques. Photogr. Eng. Rem. Sens. 49(1):69-74.

Congalton, R.G. and K. Green. 2009. Assessing the accuracy of remotely sensed data: principles and practices. 2nd Edition. CRC Press. Boca Raton, FL.

Fielding, A.H. 2002. What are appropriate characteristics of an accuracy measure? Chapter 21, pp 271-280. *In* J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson Eds. Predicting species occurrence, issues of accuracy and scale. Island Press. Washington, DC.

Fielding, A.H., and J.F. Bell. 1997. A review of methods for assessment of prediction erros in conservation presence/absence models. Environmental Conservation 24:38-49.

Grech, A. and H. Marsh. 2008. Rapid assessment of risks to a mobile marine mammal in an ecosystem-scale marine protected area. Conservation Biology 22(3):711-720.

Gopalm S. and C. Woodcock. 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogr. Eng. Rem. Sens. 60(2):181-188.

Mahon, T., E. McClaren, and F. Doyle. 2008. Parameterization of the Northern Goshawk (*Accipiter gentilis laingi*) Habitat Model for Coastal British Columbia. Nesting and Foraging Habitat Suitability Models and Territory Analysis Model. Unpubl. report for the BC Northern Goshawk Recovery Team, Ministry of Environment, Nanaimo, BC.

www.for.gov.bc.ca/hfd/library/fia/html/FIA2006MR105.htm [accessed June 20, 2008].

Manly, B.F.J., L.L. McDonald, D.L. Thomas, T.L. McDonald and W.P. Erickson. 2002. Resource selection by animals. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Marcot, B.G., R.S. Holthausen, M.G. Raphael, M.M Rowland, and M.J. Wisdom. 2001. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. Forest Ecology and Management 153:29-42.

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. Can. J. For. Res. 36: 3063-3074.

Meidinger, D. 2003a. Protocol for accuracy assessment of ecosystem maps. BC Tech. Rep. 011. Research. Branch,. B.C. Ministry of Forest and Range. Victoria, BC. www.for.gov.bc.ca/hfd/pubs/Docs/Tr/Tr011.pdf [accessed January 2, 2009].

Meidinger, D. 2003b. Ecosystem mapping accuracy and timber supply applications. Unpubl. provincial policy statement. B.C. Ministry of Forest and Range, Victoria BC. www.for.gov.bc.ca/hre/becweb/Downloads/Downloads_PEM/PEMaccuracystatement2003.pdf [accessed January 2, 2009].

Moon, D. 2004. A protocol for assessing thematic map accuracy using small area sampling. Unpubl. Rep. submitted to the Cariboo Site Productivity Adjustment Working Group.

Olsson, O. and D.J. Rogers. 2009. predicting the distribution of a suitable habitat for the white stork in Southern Sweden: identifying priority areas for reintroduction and habitat restoration. Animal Conservation 12:62-70.

Resource Inventory Standards Committee. 1999. British Columbia Wildlife Habitat Rating Standards. Version 2.0. BC Ministry of Environment. Victoria, BC. http://www.env.gov.bc.ca/wildlife/whr/essentials.html [accessed January 2, 2009].

Roloff, G.J. and B.J Kernohan. 1999. Evaluating reliability of habitat suitability index models. Wildl. Soc. Bull. 27(4):973-985.

US Fish and Wildlife Service. 1981. Standards for the Development of Habitat Suitability Index Models. Ecological Services Manual 103. Dept of the Interior. Washington, DC.

Williams, N.S.G., A.K. Hahs, and J.W. Morgan. A dispersal-constrained habitat suitability model for predicting invasion of alpine vegetation. Ecological Applications 18(2):347-359.

Appendix 1. Description of rating interpretations and typical habitat conditions found across the gradient of nesting habitat suitability for Northern Goshawks in Coastal BC. The purpose of this table was to provide a basis for goshawk experts to standardize the criteria they used to decide on suitability ratings in the field. This is not a cookbook type lookup table that non-experts could use to generate reliable field calls.

Suitability Rating	0 – 0.25 (Nil)	0.25 – 0.50 (Low)	0.50 – 0.75 (Moderate)	0.75 – 1.00 (High)
Interpretation	Unsuitable. Habitat fails to provide minimum requirements.	Suitability Unknown. Habitat provides theoretical minimum requirements for supporting a nest, but use by goshawks is rarely observed. Suitability of two or more habitat variables is suboptimal, substantially reducing the overall suitability of the stand.	Suitable. Suitability of one or two habitat variables is lower than optimal conditions but minimum requirements still exceeded. Minority of nest sites expected to occur in Moderate class habitat.	Suitable. All habitat variables meet optimal conditions. Majority of nest sites are expected to occur in High class habitat.
Nest Platforms[1]	None	Very Limited	Somewhat Limited	Common
Subcanopy Flyways[2]	Either overdense stands with virtually no flyways or very open stands with few, interspersed trees with virtually no canopy	Flyways limited by multistoried stand structure or overdense stand (e.g. young forest)	Flyways somewhat limited by multistoried stand structure	Many clear flyways >30m in length below a closed overstory
Forest Spp	non-forested or forested bog	Yc, Pl, Bl, Cw	Ba, Hm, deciduous	Hw, Ss, Fd
Structural Stage	0 - 4	4 - 7	5 - 7	(5) 6 + 7
Height	<14m	14 - 20	20 - 26	> 26m
Canopy Closure	<20%	<35%	35 – 45%	≥ 45%
Expected % use[3]	0%	0-10%	10-25%	70-90%

[1] Branches large enough and in appropriate form to support a nest
[2] Flyways through the B2 and A3 layers to access nests and prey
[3] Expected distribution of a sample of nest areas at a regional level. Use of moderate and low quality habitats reflects heterogeneity of individual selection and issue of preference vs minimum requirement.