Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 260–267 doi: 10.1111/j.2041-210X.2011.00170.x

Assessing transferability of ecological models: an underappreciated aspect of statistical validation

Seth J. Wenger[1]* and Julian D. Olden[2]

[1]*Trout Unlimited, 322 E. Front Street, Suite 401, Boise, ID 83702, USA; and* [2]*School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195, USA*

Summary

1. Ecologists have long sought to distinguish relationships that are general from those that are idiosyncratic to a narrow range of conditions. Conventional methods of model validation and selection assess in- or out-of-sample prediction accuracy but do not assess model generality or transferability, which can lead to overestimates of performance when predicting in other locations, time periods or data sets.

2. We propose an intuitive method for evaluating transferability based on techniques currently in use in the area of species distribution modelling. The method involves cross-validation in which data are assigned non-randomly to groups that are spatially, temporally or otherwise distinct, thus using heterogeneity in the data set as a surrogate for heterogeneity among data sets.

3. We illustrate the method by applying it to distribution modelling of brook trout (*Salvelinus fontinalis* Mitchill) and brown trout (*Salmo trutta* Linnaeus) in western United States. We show that machine-learning techniques such as random forests and artificial neural networks can produce models with excellent in-sample performance but poor transferability, unless complexity is constrained. In our example, traditional linear models have greater transferability.

4. We recommend the use of a transferability assessment whenever there is interest in making inferences beyond the data set used for model fitting. Such an assessment can be used both for validation and for model selection and provides important information beyond what can be learned from conventional validation and selection techniques.

Key-words: cross-validation, generality, niche model, performance, species distribution model, statistical

Introduction

A fundamental goal of ecology, as in other branches of science, is to identify relationships and patterns that are repeatable or general (Peters 1991). Although a relationship that is idiosyncratic to a narrow set of conditions may be interesting and informative, no ecologist would wish to mistake it for an association that is broadly applicable and constitutes a general rule. Such relationships or models can be said to have *generality* (Fielding & Haworth 1995; Olden & Jackson 2000), *generalizability* (Justice, Covinsky & Berlin 1999; Vaughan & Ormerod 2005) or *transferability* (Thomas & Bovee 1993; Randin *et al.* 2006) to data sets other than the one for which they were developed. However, conventional approaches to evaluating ecological models do not commonly provide inference into transferability. As a result, the generality of a model is often unknown, and the model selected as 'best' for a given data set may have worse transferability than an alternative, rejected one.

The issue of transferability has been the subject of intermittent ecological interest for a number of years, but this greatly increased with the rise of the field of species distribution modelling (Elith & Leathwick 2009) in the 2000s. Researchers have investigated whether a species model developed in one region can successfully predict in a different region (Peterson, Papeş & Kluza 2003; Randin *et al.* 2006; Peterson, Papeş & Eaton 2007; Barbosa, Real & Vargas 2009; Sundblad *et al.* 2009; Wenger *et al.* 2011a) and to a smaller extent whether models developed in one time period can predict a different time period with different weather or climatic conditions (Boyce *et al.* 2002; Araújo *et al.* 2005; Varela, Rodríguez & Lobo 2009; Buisson *et al.* 2010; Tuanmu *et al.* 2011). However, the question of model transferability is a general one that is common to questions other than those of species–environment relationships. It is equally important to consider the generality

*Correspondence author. E-mail: swenger@tu.org
Correspondence site: http://www.respond2articles.com/MEE/

© 2012 The Authors. Methods in Ecology and Evolution © 2012 British Ecological Society

of models of physical phenomena (e.g. models of temperature), of ecological processes (e.g. denitrification rates) or of population parameters (e.g. growth rates). The fundamental problem is that there can be considerable spatial or temporal heterogeneity in ecological relationships, and this heterogeneity can limit model generality.

Lack of model generality is often a result of overfitting (Chatfield 1995; Sarle 1995), which can be defined as accepting a predictor variable (or a form of a predictor variable, such as a squared term or interaction term) that is nominally correlated with the response variable in the data set, but which does not represent a relationship that holds generally. Overfitting may occur for two rather different reasons. First, weak correlations among variables arise as a result of random noise, and these may be incorrectly interpreted as legitimate relationships. Model selection criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are designed to minimize this kind of overfitting by penalizing models for excess complexity, resulting in the rejection of spurious relationships. By and large, these criteria are effective at this goal (Burnham & Anderson 2002, 2004) and have been widely adopted by ecologists. Traditional cross-validation techniques such as leave-one-out can also provide an unbiased assessment of model performance that does not favour such overfitted models (Olden & Jackson 2000; Olden, Jackson & Peres-Neto 2002). The second cause of overfitting is when there are statistical associations between predictor and response variables that are real in a given data set but do not occur under a wide range of conditions. For example, with the large data sets now commonly used in species distribution modelling, models with over 100 terms may be justifiable under traditional criteria, but such a precise description of a distribution in one location often transfers poorly to other locations (Tuanmu *et al.* 2011). This is especially true when using indirect predictors without a close, perceived mechanistic link to the response variable (Randin *et al.* 2006). Of course, underfitting is also possible. For example, a model predicting species occurrence as a function of elevation may be more parsimonious than a more complex one based on temperature and precipitation, but the elevation model will likely transfer poorly to other latitudes.

An estimate of transferability is especially important with the increased use of machine-learning modelling techniques such as neural networks (Lek & Guegan 1999), genetic algorithms (Stockwell & Noble 1992), maximum entropy (Phillips, Anderson & Schapire 2006), support vector machines (Drake, Randin & Guisan 2006), classification and regression trees (De'ath & Fabricius 2000) and random forests (RF) (Breiman 2001). These methods have the potential to match highly nonlinear, complex relationships, yielding in-sample and randomly cross-validated performance superior to that of traditional generalized linear modelling (Elith *et al.* 2006; Olden, Lawler & Poff 2008), but at the risk of limited transferability if model complexity is not constrained (Sarle 1995; Tuanmu *et al.* 2011). Assessment of generality of such models is critical if they are to be used in a predictive manner beyond the conditions under which they were trained – for example, if species distribution models are used to make projections under climate change conditions (Araújo *et al.* 2005).

Our primary objective in this paper is to present a general approach to estimating model transferability by extending and formalizing methods currently in use in the species distribution modelling literature. A secondary objective is to illustrate why transferability assessment can be important. We do this by fitting different kinds of models to an example data set and then comparing model transferability to traditional performance measures, showing how apparently good models can transfer very poorly. We then discuss practical aspects, limitations and appropriate use of the method.

The method: estimating transferability via non-random cross-validation

In species distribution modelling, transferability has often been estimated by splitting the data set into geographically distinct subsets, fitting the model with the first subset (called the training data set) and validating with the second (called the test data set). Then, the process is reversed, with the second subset used for fitting and the first for validation. This is nothing more than a form of cross-validation in which the subset membership is assigned non-randomly based on a relevant factor such as geography. We propose that this approach can be generalized to serve as a standard method for transferability assessment. We introduce the method by first reviewing conventional validation techniques.

A model's performance can be validated based on the error in its predictions of observed data. If these predictions involve the same data used to fit the model (i.e. the training and testing data sets are identical), then the errors are the model residuals and are called in-sample or resubstitution error. However, in-sample error underestimates true model error, especially for small sample sizes (Efron 1986; Fielding & Bell 1997; Olden & Jackson 2000; Burnham & Anderson 2002; Olden, Jackson & Peres-Neto 2002). An alternative approach is to use a fully independent validation data set, which provides an independent test of model error and a direct measure of transferability (Fielding & Bell 1997). The downside, of course, is that the test data are not used for model fitting.

A useful compromise is cross-validation, which uses all of the data but also can provide unbiased error estimates. With cross-validation, a portion of the data is withheld for model training, and a different portion is withheld as the test data set. In this way, all data are iteratively used for both training and testing. There are various types of cross-validation depending on what fraction of the data set is excluded from model training and used for validation. In fivefold cross-validation, the data are divided into five groups, and one-fifth of the data are withheld at a time; in n-fold or leave-one-out cross-validation, only one data point is withheld at a time. It has been shown that leave-one-out provides an unbiased estimate of model error, even at small sample sizes (Olden & Jackson 2000) and furthermore that when used as the basis for model selection, it is (asymptotically) consistent with the widely used AIC (Shao 1993). However, neither leave-one-out nor other forms of

cross-validation, as conventionally applied, provide an estimate of model transferability. This is because unless sample size is very small, subsamples randomly selected from the full range of data provide an unbiased estimate of the overall relationships in the full data set, but do not necessarily reflect the heterogeneity that may exist across space or time. The problem is compounded when there is autocorrelation in the data, such that for any given data point in the training data set, there is likely to be a similar, correlated data point in the validation data set (Araújo *et al.* 2005).

An alternative is to divide the data *non-randomly* into groups for cross-validation, such that any group used for validation differs from those used for training the model in the same way that an independent data set would. That is, we use heterogeneity within the data set as a surrogate for heterogeneity among data sets. For a species distribution model, for example, cross-validation based on dividing data into multiple geographic regions provides inferences into how the model will perform in an unsampled region (e.g. Olden & Jackson 2001; Kennard *et al.* 2007) or under future climate conditions in the same region (Vaughan & Ormerod 2005). This can readily be extended to other types of data sets. For example, in a data set with 5 years of annual observations, a full year of data could be withheld at a time; this ought to provide a reasonable estimate of the model's predictive ability in a future, as yet unobserved year. In general terms, we can define the transferability of a model as the accuracy of its predictions for an independent data set; an *estimate* of transferability (which we refer to as a 'transferability assessment') is provided by non-random cross-validation.

An intuitive extension of assessing transferability is to use the results to select among competing models. Model validation and model selection are two sides of the same coin; it is reasonable to rank models based on their validated predictive performance and select the best performing model or a confidence set of good performing models (Arlot & Celisse 2010). Evaluating models based on transferability should provide a highly robust method of identifying relationships with predictor variables that are truly general, thus greatly reducing the risk of overfitting and increasing model utility.

Example: invasive trout in the western United States

Brook trout (*Salvelinus fontinalis* Mitchell) and brown trout (*Salmo trutta* Linnaeus) are introduced species in the western United States, where they are considered invasive and a threat to the persistence of native trout (Thurow, Lee & Rieman 1997; Dunham *et al.* 2002; McHugh & Budy 2006). Relationships between the species' distributions and climatic and landscape variables are of interest for predicting future invasions, as well as the species' potential response to projected climate change. We used a data base of 9890 presence/absence fish collection records from the interior west of the United States (Fig. 1) to model brook and brown trout occurrence as a function of predictor variables that were selected *a priori* to be likely influences on the species distributions: mean summer air temperature, winter high flow frequency, mean flow, slope, presence/absence of a road within a kilometre of the stream and distance to the nearest unconfined valley. Details on the data and variables are in Wenger *et al.* (2011b); here, we summarize the statistical analysis methods used here.

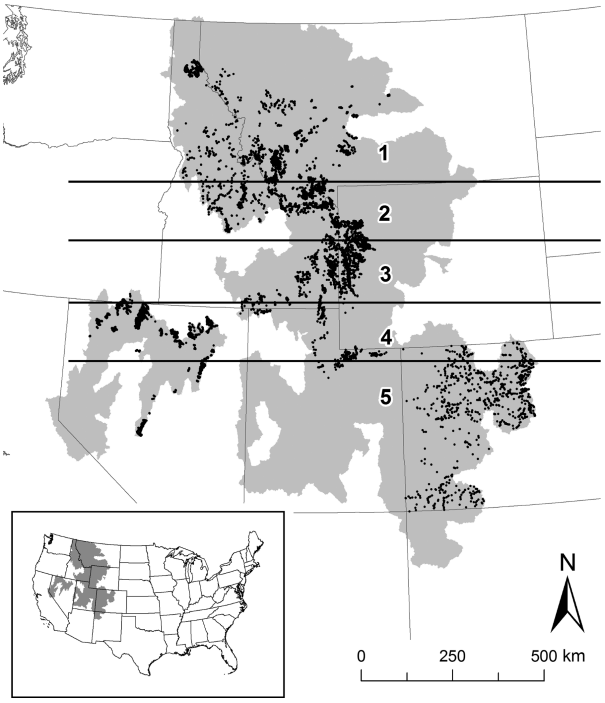

Fig. 1. Collection sites (black dots) and study area (grey shading) in the western United States used in the example. For the fivefold transferability assessment, sites were assigned non-randomly to five groups (labelled with numbers) based on latitudinal bands (delineated by heavy black lines). For the 10-fold cross-validation, each of these bands was divided by latitude into two equal-sized groups, producing 10 groups. For the twofold cross-validation, the entire data set was divided by latitude into two equal-sized groups.

Three modelling approaches were employed: (i) multilevel generalized linear modelling [or generalized linear mixed modelling (GLMM)] with a logit link; (ii) artificial neural networks (ANN) and (iii) the RF classifier. In the GLMM modelling, we used a multilevel analysis because sites were not distributed randomly across the landscape, but were often clustered; a multilevel approach (with sites nested within watersheds) reduces the bias caused by such spatial autocorrelation (Raudenbush & Bryk 2002; Gelman & Hill 2007). We used AIC to select the best model for each species from among a candidate set of GLMM models with different combinations of predictor variables. The second method, ANN, is a widely used machine-learning technique that can account for nonlinearities and complex interactions among variables (Olden, Lawler & Poff 2008). For the ANN modelling, we included all six predictor variables for each species and specified three types of network architecture of increasing complexity: one with six hidden nodes, one with 12 hidden nodes and one with 18 hidden nodes, all in a single layer. The third method, RF, is a type of sophisticated classification and regression tree analysis (see De'ath & Fabricius 2000) that has been shown to display

excellent in-sample predictive performance (e.g. Lawler *et al.* 2006; Holden, Morgan & Evans 2009). Random forest classifiers are a model-averaging or ensemble-based approach in which multiple classification or regression tree models are built using random subsets of the data and predictor variables (Cutler *et al.* 2007). We grew a forest of 1000 classification trees by sampling with replacement randomized subsets of the original observations (using default software settings). We included all six predictor variables. Models were fit in the R Statistical Package (http://www.R-project.org) using the packages *lme4*, *nnet* and *randomForest*.

Model performance was evaluated in three ways. First, we calculated in-sample model performance by using the models to predict the training data used for fitting. Secondly, we performed a traditional type of fivefold cross-validation in which data points were randomly assigned to groups. We iteratively trained the model using 4/5 of the data and validated it using 1/5 of the data. Third, we assessed transferability by partitioning the data non-randomly into latitudinal bands. We examined 2-, 5- and 10-fold cross-validation (Fig. 1 shows fivefold group assignments as an example), iteratively fitting with one group withheld at a time and evaluating performance in predicting the withheld data. The use of latitudinal bands was designed to roughly separate the data climatically, providing inference into transferability to a future climate (Wenger *et al.* 2011b). For the GLMMs, only the fixed effects were used to make predictions for validation, as it is not possible to estimate random effects for regions outside the fitting data set. Because ANNs and RF are subject to random variability in model fitting, we ran 100 iterations of each validation. For the traditional cross-validation, different training and fitting subsets were selected randomly at every iteration, but for the transferability assessments, group assignments were constant across iterations. For each validation, we calculated area under the curve (AUC), the AUC of the receiver-operator characteristic plot, a widely used and unbiased summary metric of model performance for binary data (Guisan & Zimmermann 2000; Manel, Wiilams & Ormerod 2001; but see Lobo, Jimenez-Valverde & Real 2008 for limitations). We explored alternative performance measures, but as all gave results that were essentially identical to AUC, we report only the latter.

Results showed that the RF models had the highest performance based on in-sample and random cross-validation, followed by models developed using ANNs and GLMMs (Table 1). Among the three ANN models, the more complex formulations (models with many nodes) tended to have better in-sample and random cross-validation performance than the simpler ones (with few nodes). The comparison of model transferability among the methods showed nearly opposite trends. The GLMM models displayed the highest transferability, while the RF and ANN models exhibited substantially lower transferability. The ANN models with simpler formulations (in terms of the number of nodes) had greater transferability than the more complex versions. Note that for brook trout, even the GLMM transferability was not very good, but the performances of the other modelling methods were even worse. The 2-, 5- and 10-fold transferability assessments all showed the same trends, but the twofold transferability assessment produced the lowest AUC scores, followed by 5- and 10-fold.

Examination of the predictor–response relationships from the different models sheds insight into the cause of the poor transferability performance of the machine-learning approaches (RF and ANN). As an example, consider the temperature response for the RF and GLMM models. In the GLMM models, this was represented by a quadratic relationship, temperature + temperature^2 (Fig. 2), selected *a priori* as a candidate relationship because it represents the classical species niche association with temperature as an ecological resource (Magnuson, Crowder & Medvick 1979; Austin 2002). By contrast, the RF model empirically describes the observed relationships in the data, with no assumptions of form, as shown in the partial dependence plot (Hastie, Tibshirani & Friedman 2001) of occurrence in response to temperature (Fig. 3). The jagged shape of the response curve

Table 1. Area under the curve (AUC) of the receiver-operator characteristic plot for models based on in-sample validation, random cross-validation and transferability assessment (non-random cross-validation)

Species Model	In-sample	Random CV	Transferability 10-fold	Transferability fivefold	Transferability twofold
Brown trout					
Random forests	**0·918**	**0·912**	0·749	0·717	0·711
ANN – 18 hidden nodes	0·906	0·848	0·729	0·711	0·646
ANN – 12 hidden nodes	0·892	0·838	0·738	0·719	0·665
ANN – 6 hidden nodes	0·865	0·834	0·756	0·740	0·703
Multilevel GLMM	0·822	0·820	**0·788**	**0·783**	**0·757**
Brook trout					
Random forests	**0·884**	**0·873**	0·625	0·603	0·516
ANN – 18 hidden nodes	0·783	0·745	0·618	0·596	0·551
ANN – 12 hidden nodes	0·764	0·738	0·627	0·604	0·553
ANN – 6 hidden nodes	0·728	0·717	0·640	0·618	0·563
Multilevel GLMM	0·674	0·673	**0·650**	**0·653**	**0·574**

Best models for each species for each performance measure are shown in bold. An AUC score of 0·5 is no better than random; scores >0·7 are good, and scores >0·9 are excellent. Each value is the mean of 100 iterations.
ANN, artificial neural networks.

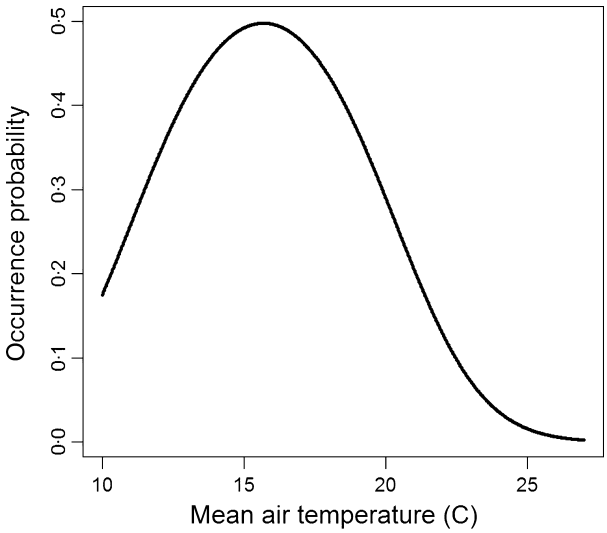

Fig. 2. Plot of the probability of brook trout occurrence in response to air temperature, with other variables held to their mean values, based on the best-supported GLMM model.

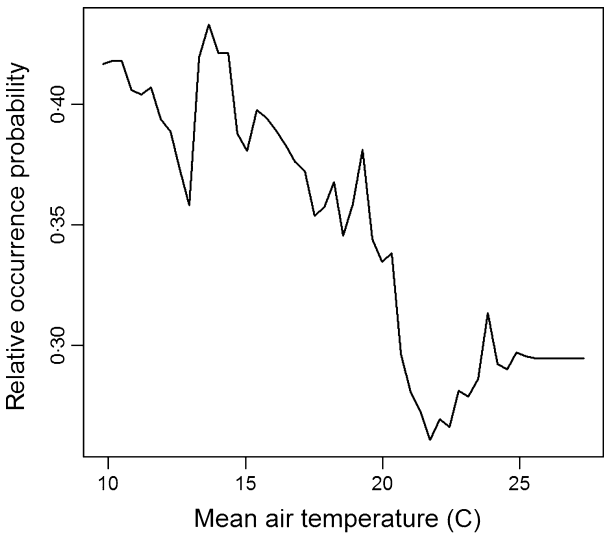

Fig. 3. Partial dependence plot for the probability of brook trout occurrence in response to air temperature, based on random forests model. The *Y*-axis is 0·5 × the logit of the occurrence probability; for practical purposes, this may be viewed as relative occurrence probability. It is the shape of the plot that should be compared with Fig. 2.

matches the fitting data set extremely well, but likely has no basis in biology, and it is perhaps no wonder that it fails to transfer to other data sets.

Discussion and practical guidance

Our example demonstrates the importance of considering model transferability, in addition to traditional measures of model accuracy, when assessing model performance. Based on in-sample validation and conventional cross-validation, the RF models for both brook trout and brown trout appeared to be excellent, and it would be tempting to use them as a forecasting tool – such as for projections of future invasion potential or distributional responses to projected climate change. However, the transferability assessment indicates this could be a mistake. The RF and many ANN models performed relatively poorly when just a fifth of the data was non-randomly withheld, suggesting possible overfitting and the need for great caution in making inferences in new locations or new climates. Our case study illustrated that simpler models can, at least in some cases, be more transferable.

It is now common to see machine-learning methods like RF applied to a range of ecological data analyses, including projections of species distributions under climate change scenarios, without any assessment of transferability (e.g. Ledig *et al.* 2010). Because RF is, by design, immune to overfitting associated with random noise (Breiman 2001), researchers may make the incorrect assumption that it is also immune to overfitting caused by heterogeneity in predictor–response relationships. Of course this cannot be true; for example, no reasonable ecologist would interpret the complex relationship shown in Fig. 3 to be a general one that can be applied with high accuracy in other locations. It may be surprising that a method like RF that is robust to overfitting in the conventional sense should suffer poor transferability. However, RF, like all analytical methods, is designed to seek the best fit for a data set as a whole. It cannot distinguish between predictor–response relationships with high generality from those that have less generality, but are nevertheless legitimate relationships in the data set. The only way to gain insight into the degree of model generality or transferability is either to test the model with a new, independent data set or to cross-validate it using non-random subsets, as we advocate here.

We suggest that a transferability assessment be conducted whenever there is interest in making projections or inferences beyond the data set used for model fitting. If results suggest that transferability is substantially worse than in-sample performance, simpler alternative models should be considered. With RF, it is possible to simplify by reducing the number of parameters used in modelling, reducing the maximum number of nodes per tree and specifying a minimum number of cases per node (although our attempts to use these resulted in minimal improvements, as evidenced by partial dependence plots and transferability; S.J. Wenger, unpublished data). Other classification algorithms such as boosted regression trees have alternative settings to manage complexity (Elith, Leathwick & Hastie 2008). Neural networks offer multiple ways to control complexity, including managing the number of nodes (as we did here), stopping the algorithm early and weight decay or weight elimination (Bishop 1995; Sarle 1995; Olden & Jackson 2002). We possibly could have achieved higher transferability in the ANN models in our example if we had used these techniques to more aggressively control complexity. For general additive models, the order and number of knots in the splines may be limited on a variable-by-variable basis; for GLMMs and GLMs, complexity may be similarly controlled by limiting higher-order terms; and for both general additive models and GLMMs, interactions may be specified or not.

Another factor that may influence transferability is the distance of the causal links between the predictor variables and the response variable. Predictor–response relationships that have a sound ecological basis and direct causal linkages are likely to be more transferable than those based on indirect relationships or pure correlation (Austin 2002; Sundblad *et al.* 2009). For this reason, we suggest selecting predictor variables and possibly even the form of the expected response (e.g. positive, negative and quadratic) on the basis of reasonable *a priori* hypotheses, unless the goal of the modelling is purely exploratory. This is perhaps best performed using methods such as GLMs, GLMMs and GAMs that offer a high degree of user control. Such models benefit further from well-developed theory and methodologies for addressing autocorrelation (Cressie 1993; Lichstein *et al.* 2002; Dormann *et al.* 2007), a problem that is generally ignored in the application of machine-learning methods. Of course, GLMs and their variants are not immune to overfitting, and GLMs with excessive parameters, higher-order terms or interactions can suffer from decreased transferability (Wenger *et al.* 2011a).

In devising a transferability assessment, the researcher must make several key decisions requiring a degree of professional judgment. The first of these is deciding how many groups into which to divide the data set (i.e. the number of folds of *k*-fold cross-validation), which is essentially a decision on how conservative a test to run. In our example, we found that the fewer the groups, the more conservative the assessment. We expect this to be a general rule and to be true regardless of the size of the data set, which stands in contrast to random cross-validation, in which the number of groups becomes irrelevant as the size of the data set grows arbitrarily large (Olden & Jackson 2000). To date, transferability assessments in the field of species distribution modelling have tended to use twofold cross-validation (e.g. Randin *et al.* 2006; Peterson, Papeş & Eaton 2007; Barbosa, Real & Vargas 2009). We suspect this will be overly conservative for many applications. On the other hand, cross-validation with more than 10-fold may be too liberal, so we recommend between 3- and 10-fold for most applications. The choice depends largely on how projections are to be used. If the data set covers most of the area of potential inference, a more liberal test is reasonable; if the coverage of the data set is small relative to the area of inference, then a more conservative test is appropriate. For example, consider a case where researchers wish to parameterize survival estimates in a population model based on a targeted study involving six populations, and then apply the results to forecasts of 60 populations across a large region. In such a case, a conservative transferability assessment based on twofold or threefold cross-validation would be essential to avoid overly optimistic predictions of the generality of the observed relationships. If there is difficulty in deciding how many groups to use, it is perfectly reasonable and usually quite practical to run multiple tests with different numbers of groups, as we did.

A second key decision is how to assign data to the groups. Two principles should guide this process. The first is that all of the fitting data sets should cover a large portion of the range of variability of the predictor variables of interest. For example, in building a species–climate model, if all high elevation locations are placed in a single group, that group will likely be poorly predicted because it lies outside the range of variability of the other groups. This would be overly conservative, so it is preferable to assign those high elevation sites to at least two groups, so some of them are always available for model training. We used latitudinal bands in our example because the large elevational gradient in this region produced a climatic range of a magnitude at least as great as that produced by the latitudinal gradient, preserving the range of variability in predictor variables when a band is removed (except in the very conservative case of twofold validation). However, such an assignment would probably not be appropriate for a data set from the Great Plains of the United States, which have little elevational gradient. The second principle for guiding group assignments is that the heterogeneity among the groups (in terms of predictor–response relationships) should be in the range of the expected heterogeneity between the full data set and other locations or data sets for which inferences are of interest. In our example, we were interested in the temporal climatic variability between current conditions and future conditions and made the assumption that climatic variability across latitudinal bands provided a reasonable surrogate. These guidelines notwithstanding, concern over optimizing group assignments should not become an obstacle to performing a transferability assessment, as any rational method of group assignments is likely to yield useful information, especially for large data sets. With small data sets, where it is possible for a particular grouping to significantly affect the outcome, it may be useful to repeat the transferability assessment multiple times with different group assignments in a form of ensemble prediction (Araújo & New 2007). We explored this with our example data set (S.J. Wenger, unreported data), but found it had little effect on the results, likely due to the relatively large sample size.

A transferability assessment is not necessary or appropriate for all data sets and all circumstances. If projections and inferences do not extend beyond the conditions represented by the data used to fit the model (e.g. Evans & Cushman 2009), transferability is less relevant. Most models of data from tightly controlled experiments also would not benefit from a transferability assessment because heterogeneity is either limited or is itself the focus of study. For very small data sets, transferability assessments may be infeasible because models cannot be effectively fit unless all the data are used, or because the data do not cover a sufficient range of conditions. Where it is appropriate, we regard a transferability assessment as a useful tool that complements existing model performance measures and selection methods. It has some clear limitations. Dividing the data into subsets provides some inferences into how a model will perform with a new data set (e.g. a different region or time period), but the actual performance could be substantially better or worse. Using a transferability assessment based on geographic units to provide inferences into performance under a future climate requires assumptions that regional climatic differences are of a similar magnitude to differences between current and future climates in a single area, which cannot really be

known. In some cases, climates will shift to novel ones that lack current analogues (Williams, Jackson & Kutzbacht 2007), limiting the value of a spatial transferability assessment. Even under such circumstances, we argue that a transferability assessment provides important additional information beyond what can be learned from traditional performance assessment methods. Whenever there is interest in extending inferences to data sets beyond the one used to fit an ecological model, the researcher is better off armed with some insight into potential transferability (however imperfect) than proceeding under the untested assumption that the model will perform with the same error rate in new data sets as the one used to create it.

Acknowledgements

S.J.W. was supported in part by grant G09AC00050 from the US Geological Survey and a contract from the Forest Service Rocky Mountain Research Station. J.D.O. acknowledges funding support from the U.S. EPA Science To Achieve Results (STAR) Program (Grant No. 833834), USGS Status and Trends Program and the USGS National Gap Analysis Program. The manuscript was improved by helpful comments from Daniel Dauwalter, Mary Freeman, Daniel Isaak and two anonymous reviewers.

References

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42–47.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Arlot, S. & Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statistical Surveys*, **4**, 40–79.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Barbosa, A.M., Real, R. & Vargas, J.M. (2009) Transferability of environmental favourability models in geographic space: the case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. *Ecological Modelling*, **220**, 747–754.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, UK.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference*. Springer, New York, NY.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference – understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**, 261–304.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **158**, 419–466.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*. Wiley, New York.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., *et al*. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Drake, J.M., Randin, C. & Guisan, A. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, **43**, 424–432.

Dunham, J.B., Adams, S.B., Schroeter, R.E. & Novinger, D.C. (2002) Alien invasions in aquatic ecosystems: toward an understanding of brook trout invasions and potential impacts on inland cutthroat trout in western North America. *Reviews in Fish Biology and Fisheries*, **12**, 373–391.

Efron, B. (1986) How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, **81**, 461–470.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Evans, J.S. & Cushman, S.A. (2009) Gradient modeling of conifer species using random forests. *Landscape Ecology*, **24**, 673–683.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fielding, A.H. & Haworth, P.F. (1995) Testing the generality of bird-habitat models. *Conservation Biology*, **9**, 1466–1481.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, USA.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hastie, T.J., Tibshirani, R.J. & Friedman, J.H. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, NY, USA.

Holden, Z.A., Morgan, P. & Evans, J.S. (2009) A predictive model of burn severity based on 20-year satellite-inferred burn severity data in a large southwestern US wilderness area. *Forest Ecology and Management*, **258**, 2399–2406.

Justice, A.C., Covinsky, K.E. & Berlin, J.A. (1999) Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, **130**, 515–524.

Kennard, M.J., Olden, J.D., Arthington, A.H., Pusey, B.J. & Poff, N.L. (2007) Multiscale effects of flow regime and habitat and their interaction on fish assemblage structure in eastern Australia. *Canadian Journal of Fisheries and Aquatic Sciences*, **64**, 1346–1359.

Lawler, J.J., White, D., Neilson, R.P. & Blaustein, A.R. (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology*, **12**, 1568–1584.

Ledig, F.T., Rehfeldt, G.E., Saenz-Romero, C. & Flores-Lopez, C. (2010) Projections of suitable habitat for rare species under global warming scenarios. *American Journal of Botany*, **97**, 970–987.

Lek, S. & Guegan, J.F. (1999) Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, **120**, 65–73.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Lobo, J.M., Jimenez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Magnuson, J.J., Crowder, L.B. & Medvick, P.A. (1979) Temperature as an ecological resource. *American Zoologist*, **19**, 331–343.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

McHugh, P. & Budy, P. (2006) Experimental effects of nonnative brown trout on the individual- and population-level performance of native Bonneville cutthroat trout. *Transactions of the American Fisheries Society*, **135**, 1441–1455.

Olden, J.D. & Jackson, D.A. (2000) Torturing data for the sake of generality: how valid are our regression models? *Ecoscience*, **7**, 501–510.

Olden, J.D. & Jackson, D.A. (2001) Fish-habitat relationships in lakes: gaining predictive and explanatory insight using artificial neural networks. *Transactions of the American Fisheries Society*, **130**, 878–897.

Olden, J.D. & Jackson, D.A. (2002) Illuminating the "black box": understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**, 135–150.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a comment on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Olden, J.D., Lawler, J.J. & Poff, N.L. (2008) Machine learning methods without tears: a primer for ecologists. *Quarterly Review of Biology*, **83**, 171–193.

Peters, R.H. (1991) *A Critique of Ecology*. Cambridge University Press, Cambridge, UK.

Peterson, A.T., Papeş, M. & Eaton, M. (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30**, 550–560.

Peterson, A.T., Papeş, M. & Kluza, D.A. (2003) Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science*, **51**, 863–868.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Raudenbush, S.W. & Bryk, A.S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, London, UK.

Sarle, W.S. (1995) Stopped training and other remedies for overfitting. *Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics*, pp. 352–360. Interface Foundation of North America, Fairfax Station, VA, USA.

Shao, J. (1993) Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

Stockwell, D.R.B. & Noble, I.R. (1992) Induction of sets of rules from animal distribution data – a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, **33**, 385–390.

Sundblad, G., Harma, M., Lappalainen, A., Urho, L. & Bergstrom, U. (2009) Transferability of predictive fish distribution models in two coastal systems. *Estuarine Coastal and Shelf Science*, **83**, 90–96.

Thomas, J.A. & Bovee, K.D. (1993) Application and testing of a procedure to evaluate transferability of habitat suitability criteria. *Regulated Rivers: Research & Management*, **8**, 285–294.

Thurow, R.F., Lee, D.C. & Rieman, B.E. (1997) Distribution and status of seven native salmonids in the Interior Columbia Basin and Portions of the Klamath River and Great Basins. *North American Journal of Fisheries Management*, **17**, 1094–1110.

Tuanmu, M.-N., Viña, A., Roloff, G.J., Liu, W., Ouyang, Z., Zhang, H. & Liu, J. (2011) Temporal transferability of wildlife habitat models: implications for habitat monitoring. *Journal of Biogeography*, **38**, 1510–1523.

Varela, S., Rodríguez, J. & Lobo, J.M. (2009) Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. *Journal of Biogeography*, **36**, 1645–1655.

Vaughan, I.P. & Ormerod, S.J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720–730.

Wenger, S.J., Isaak, D.J., Dunham, J.B., Fausch, K.D., Luce, C.H., Neville, H.M., Rieman, B.E., Young, M.K., Nagel, D.E., Horan, D.L. & Chandler, G.W. (2011a) Role of climate and invasive species in structuring trout distributions in the Interior Columbia Basin. *Canadian Journal of Fisheries and Aquatic Sciences*, **68**, 988–1008.

Wenger, S.J., Isaak, D.J., Luce, C.H., Neville, H.M., Fausch, K.D., Dunham, J.B., Dauwalter, D.C., Young, M.K., Elsner, M.M., Rieman, B.E., Hamlet, A.F. & Williams, J.E. (2011b) Flow regime, biotic interactions and temperature drive differential declines of trout species under climate change. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 14175–14180.

Williams, J.W., Jackson, S.T. & Kutzbacht, J.E. (2007) Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5738–5742.

Received 27 June 2011; accepted 11 November 2011
Handling Editor: Robert Freckleton