Diversity and Distributions, (Diversity Distrib.) (2011) 1–11

BIODIVERSITY RESEARCH

Trait-based risk assessment for invasive species: high performance across diverse taxonomic groups, geographic ranges and machine learning/statistical tools

Reuben P. Keller[1]*, Dragi Kocev[2] and Sašo Džeroski[2]

[1]*Program on the Global Environment, University of Chicago, 5828 S. University Ave., Chicago, IL 60637, USA,* [2]*Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia*

*Correspondence: Reuben P. Keller, Program on the Global Environment, University of Chicago, 5828 S. University Ave., Chicago, IL 60637. USA.
E-mail: rpkeller@uchicago.edu

ABSTRACT

Aim Trait-based risk assessment for invasive species is becoming an important tool for identifying non-indigenous species that are likely to cause harm. Despite this, concerns remain that the invasion process is too complex for accurate predictions to be made. Our goal was to test risk assessment performance across a range of taxonomic and geographical scales, at different points in the invasion process, with a range of statistical and machine learning algorithms.

Location Regional to global data sets.

Methods We selected six data sets differing in size, geography and taxonomic scope. For each data set, we created seven risk assessment tools using a range of statistical and machine learning algorithms. Performance of tools was compared to determine the effects of data set size and scale, the algorithm used, and to determine overall performance of the trait-based risk assessment approach.

Results Risk assessment tools with good performance were generated for all data sets. Random forests (RF) and logistic regression (LR) consistently produced tools with high performance. Other algorithms had varied performance. Despite their greater power and flexibility, machine learning algorithms did not systematically outperform statistical algorithms. Geographic scope of the data set, and size of the data set, did not systematically affect risk assessment performance.

Main conclusions Across six representative data sets, we were able to create risk assessment tools with high performance. Additional data sets could be generated for other taxonomic groups and regions, and these could support efforts to prevent the arrival of new invaders. Random forests and LR approaches performed well for all data sets and could be used as a standard approach to risk assessment development.

Keywords
Artificial intelligence, biological invasions, logistic regression, machine learning, random forests, receiver-operator curve, traits.

INTRODUCTION

It is widely accepted that the most cost-effective way to reduce impacts from harmful invasive species is to prevent their importation (Lodge *et al.*, 2006; Keller *et al.*, 2007a). Preventing all species imports is not desirable because many more species are imported than become invasive, and most intentionally imported species are environmentally benign and/or provide economic and social benefits (e.g. as pets or garden plants; Reichard & Hamilton, 1997; Smith *et al.*, 1999). This has spurred ecologists to develop risk assessment tools for predicting which species pose a high risk of causing harm if they are imported. Accurate risk assessment tools can support policy and management efforts to reduce the overall impacts from harmful invaders while allowing importation of beneficial species (Keller & Drake, 2009).

Recent efforts at risk assessment have followed a number of paradigms (Keller & Drake, 2009). Here, we consider the quantitative approach (*sensu* Keller & Drake, 2009). This paradigm holds that there are multiple steps in the invasion

DOI: 10.1111/j.1472-4642.2011.00748.x
http://wileyonlinelibrary.com/journal/ddi

A Journal of Conservation Biogeography
Diversity and Distributions

process: to become invasive, a species must first survive transport to be introduced, must then begin reproducing to become established and must finally spread and cause harm to be considered invasive (Kolar & Lodge, 2001). A species can fail or succeed at each step. The quantitative approach considers that success at any step is a function of species traits (e.g. biological traits, invasion history, history of domestication) and that patterns in those traits can be used to explain success or failure. Patterns are generally searched for with statistical algorithms, such as logistic regression (LR). If strong patterns are found, they can be used to assess the risk posed by species that have not yet been introduced. Other risk assessment paradigms are generally also based on species traits and have similar goals to the quantitative approach. These include the scored questionnaire approach, which is the basis of the Australian weed risk assessment (Pheloung, 1995), and individual species literature reviews (e.g. Mandrak & Cudmore, 2004). These latter two approaches are not considered further in this paper, but see Keller & Drake (2009) for a review.

To create a quantitative risk assessment tool, the assessor begins by choosing a taxonomic group, geographic range and a step in the invasion process. Previous risk assessments have been made for fishes in California passing through the introduced to established transition (Marchetti *et al.*, 2004) and for molluscs in the Laurentian Great Lakes passing through the established to invasive transition (Keller *et al.*, 2007b; see Kolar & Lodge, 2001; Hayes & Barry, 2008; Keller & Drake, 2009 for reviews of additional tools). Next, the assessor chooses species traits that they believe are related to success at the invasion step and collects data for these traits for all species. The final step is to use a discrimination algorithm to search for patterns in traits that are associated with success or failure at the invasion step. The logic of risk assessment is that robust patterns in historical data can be applied to future species introductions to determine the likelihood that they will pass through the invasion step (Keller & Drake, 2009).

Choice of taxonomic and geographic ranges and invasion step are generally guided by the ecological or policy question posed. In contrast, the algorithm used to search for patterns in data generally depends on the experience and skill of the assessor and the type and extent of data available. Although the range of algorithms used to search for patterns in trait data has recently increased, it remains quite small in comparison with the large number of methods available. Risk assessments have generally been created using LR, discriminant analysis and occasionally classification trees (Keller & Drake, 2009). Machine learning, a branch of artificial intelligence (and more broadly computer science), has developed a wide range of extremely flexible and powerful methods for finding patterns in data sets. These models are generally nonparametric, make few assumptions of the data (e.g. normality) and have been developed to find complex patterns in large data sets (Witten & Frank, 2005). There is reason to believe that their increased flexibility and computational power could find more patterns in trait data and thus lead to more accurate risk assessments.

Here, we have assembled six risk assessment data sets from the literature, representing a range of taxonomic groups, geographic ranges and invasion steps. We have analysed each with a range of statistical and machine learning algorithms. We aim to test three hypotheses. First, we hypothesize that machine learning methods will produce higher performing risk assessment models than conventional statistical approaches because they are able to gather more information from available data sets and because they have proven superior for ecological applications in the past (e.g. Elith *et al.*, 2006). Second, we hypothesize that the smaller the recipient geographic range considered by the risk assessment, the more likely it is that traits associated with invasiveness will be the same across species, leading to more accurate risk assessment models. The logic for this hypothesis is that smaller geographic ranges will contain less ecological variability, leading to a narrower set of traits that promote invasion for introduced species. Third, we hypothesize that the models created using the risk assessment data sets with relatively more species, and with relatively more traits for each species, will perform better because they contain more information. We caveat our tests for hypotheses two and three by noting that the limited number of data sets ($n = 6$) and the heterogeneity in taxonomic groups, geographic regions and variable selection mean that our analysis probably has low power.

Additionally, we are interested in the basic question of how well risk assessment models can predict future invasions. There remains debate in the literature about whether useful predictions can be made (Smith *et al.*, 1999). We aim to investigate this by testing performance across a range of data sets and classification tools. Additionally, we only include species data that are available before introduction so that the resulting tools can address the question of how accurately species invasions can be predicted.

METHODS

Data sets

Six data sets were chosen from the literature based on completeness of data (i.e. few missing values), our informal judgment of their quality (based largely on the extent to which they had been referenced by others) and ease of access (i.e. data sets had been published in full). The number of data sets was limited based on these criteria and to remain a tractable number for our analyses (Table 1; Full data sets used are in Tables S6–S11 in Supporting Information). The six data sets span a range of taxonomic and geographic scales, invasion step transitions and size. The broadest taxonomic group considered is Phylum Mollusca (*MollGL*) and the narrowest is the tree Genus *Pinus* (*PinusG*). Other taxonomic groups are at the Class level of Aves (birds; BirdNZ, BirdAU) or Osteichthyes (bony fishes; *FishCA*). The remaining data set (*FishGL*) includes 45 species, all but one of which is in Class Osteichthyes. The final species in this data set is a lamprey, in Class Petromyzontidae.

Table 1 Details of invasion data sets used.

Data set name	Taxa/ Geography	Transition	Proportion successful*	# of species	# of traits†	Traits	Source
BirdNZ	Birds/ New Zealand	Introduced to established	0.342	79	5D, 6C	1) Female body length, 2) female mass, 3) geographical range outside NZ, 4) migratory, 5) #months insects part of diet, 6) herbivorous, omnivorous, carnivorous, 7) clutch size, 8) broods per season, commonly found in 9) woodlands, 10) uplands, 11) wetlands	Veltman *et al.*, 1996
BirdAU	Birds/Australia	Introduced to established	0.365	52	6D, 5C	1) Female mass, 2) plumage dichromatism, 3) migratory, 4) flocking, 5) herbivorous, omnivorous, carnivorous, 6) clutch size, 7) broods per season, 8) days egg incubation, 9) uses human-dominated habitats, 10) range outside Australia, 11) established non-indigenous species elsewhere.	Duncan *et al.*, 2001
FishGL	Fish/Laurentian Great Lakes	Introduced to established	0.533	45	5D, 13C	1) Egg diameter, 2) larval length, 3) adult length, percent adult length at age 4) 1 year, 5) 2 years, 6) incubation time, 7) annual fecundity, 8) longevity, 9) age at maturity, 10) max. spawns during female lifetime, 11) extent of parental care, 12) native range size, 13) diet breadth, 14) minimum and 15) maximum temp. tolerance, 16) human use of species, 17) established NIS elsewhere.	Kolar & Lodge, 2002
FishCA	Fish/California	Introduced to established	0.563	87	3D, 4C	1) Parental care, 2) maximum adult length, 3) physiological tolerance, 4) minimum distance between CA and species native range, 5) trophic status, 6) size of native range, 7) number of countries where species is non-indigenous established.	Marchetti *et al.*, 2004
MollGL	Molluscs/ Laurentian Great Lakes	Established to invasive	0.278	18	4D, 4C	1) Mode of reproduction, 2) egg brooding, 3) maximum adult size, 4) annual fecundity, 5) longevity 6) non-indigenous elsewhere, 7) latitude range, 8) larval stage.	Keller *et al.*, 2007b
PinusG	*Pinus*/Global	Introduced to invasive	0.703	37	5D, 9C	1) Seed mass, 2) mode of seed dispersal, 3) serotiny, 4) generation time, 5) interval between large seed crops, mean 6) elevation 7) latitude and 8) rainfall in native range, 9) rarity, 10) length of juvenile period, 11) fire tolerance, 12) variation in seed crop, 13) seed-wing loading index, 14) functional group.	Grotkopp *et al.*, 2004; Richardson *et al.*, 1990

*Proportion of species in data set that successfully transited invasion transition.

†D, discrete (e.g. is the species herbivorous, carnivorous or omnivorous?); C, continuous (e.g. body length).

Geographic range varies from global (*PinusG*) to the Laurentian Great Lakes (*MollGL*, *FishGL*). Other data sets cover the US state of California (*FishCA*), the continent of Australia (*BirdAU*) and the large island system of New Zealand (*BirdNZ*).

Each data set includes trait data for the defined group of species that came to a transition in the invasion process within the specified geographic area. Four of six data sets cover the transition from introduced (i.e. present in the region, but not necessarily found beyond captivity/cultivation) to established (i.e. wild, self-sustaining population). One data set covers the transition from established to invasive (i.e. causing negative impacts) and one from introduced to invasive (Table 1).

The number of species per data set ranges from 18 to 87, and the number of traits for each data set ranges from 7 to 17, with most data sets having a roughly equal number of discrete and continuous trait variables (Table 1). Most species traits are strictly biological and are specific to the taxonomic group in question. Four data sets also include an indication of whether species are established in other areas beyond their native range. Because we were interested in determining how well trait-based risk assessment can predict future invasions, we removed all traits that can only be assessed after a species is introduced (e.g. area that the species eventually occupies in the recipient region). This makes our input variables different than for the original analyses of these data sets and means that our results are not directly comparable. The dependent variable for each data set is the binary description of whether the species did or did not successfully transit the invasion step in question. Across data sets, the proportion of species that were successful at the invasion step ranges from 0.278 to 0.702.

Machine learning methods

Seven machine learning and two statistical algorithms were used to search for patterns in trait data that explain the success of individual species at the transition step. Each combination of data set and algorithm created a separate classification model, referred to as a 'classifier'. In the following, we briefly describe the machine learning tools used. The two statistical methods [LR and linear discriminant analysis (LDA)] are not described in detail because they are commonly used and because we used their standard implementations. All algorithms except LDA were implemented using standard procedures (except where noted below) of the software package Weka (Witten & Frank, 2005). LDA is not available in Weka; instead, we used its implementation in the MASS package of R (Venables & Ripley, 2002). For LDA, data were transformed using the internal filters in Weka as for LR and then exported for analysis in R. Missing data points were replaced with the average value of the variable for that data set.

Decision trees (DT): DTs (Quinlan, 1993) begin by splitting the data set at the threshold in one of the predictor variables that maximizes class homogeneity of the resulting two subgroups. Each group can be split further and splitting continues until a user-defined limit (usually the minimum number of species per subgroup) is reached. Each subgroup is formed at the 'node' of a tree, and the final subgroups exist at the 'leaves'. Predictions are made by sorting new species down the DT until a leaf is reached. The new species is predicted to succeed at the invasion transition step if greater than half of the training data set species that reached the same leaf were successful.

Decision trees have previously been used in ecology (De'ath & Fabricius, 2000), including for invasive species risk assessment (e.g. Kolar & Lodge, 2002; Keller *et al.*, 2007b). Because the stopping criterion can affect classifier performance, we tried minimum leaf sizes of 2, 4, 6,….., 16 species for each data set and chose the best classifier.

Random forests (*RF*) *and Boosted Decision Trees* (*BDT*) are ensemble methods that create multiple DT submodels. The predicted class of a new case is a combination of the predictions made by each submodel. Ensemble methods have been shown to produce better results than single models (Breiman, 1996, 2001; Dietterich, 2000; Caruana & Niculescu-Mizil, 2006; Kocev *et al.*, 2007).

In the case of RFs (Breiman, 2001), DTs are constructed from data sets that contain bootstrap replicates of the training cases. A bootstrap replicate is constructed by applying random selection with replacement on species in the data set. RFs then use a randomized decision tree algorithm; instead of considering all explanatory variables when selecting each split, the algorithm is constrained to only consider a subset of explanatory variables. RF predicts the class membership of a new example by running it through all DT and predicting that it belongs to the class most commonly predicted ('majority voting scheme'). We followed Breiman's (2001) recommendations and created RF classifiers with 100 trees and a number of randomly selected variables for each DT equal to the logarithm (base 2) of the total number of variables in the data set.

Boosted decision trees are constructed with a boosting algorithm (Freund & Schapire, 1996) and begin by creating a DT. Cases (i.e. species) are weighted by whether the initial model correctly predicted their class. If the model did not correctly predict a given case, the weight of that case is increased. The next submodel is created so that it preferentially includes splits that lead to correct prediction for cases that were previously predicted incorrectly. This process continues until a stopping criterion is reached (e.g. a predefined number of iterations/trees). The final BDT predicts the class membership of a new example with a weighted voting scheme, whereby the voting power of each model is proportional to its accuracy. We created 100 trees for each data set.

Instance-based Learning (IBL) classifiers (Aha *et al.*, 1991) use a 'nearest neighbour(s)' algorithm to estimate class membership. To classify a new example, the IBL algorithm finds the training example(s) that are most similar, usually in Euclidean space. The prediction is the most common class of the k closest examples (species in our case), where k is a parameter that can take values from 1 to the number of training examples. For each data set, we chose the classifier that

performed best from classifiers created with k values of 1, 3, 5,…., 17.

Naïve Bayes Classifier (*NB*) is a probabilistic algorithm that uses Bayes' theorem with a naïve independence assumption; the influence of the value of an explanatory variable on a given class is independent of the values of the other variables. NB uses the training examples to learn probabilistic relationships between the predictor and response variables. The prior and conditional probabilities are combined to give the posterior class probability, which is the predicted class membership of a new example.

Support Vector Machines (SVM) treat the training examples as two sets of vectors (i.e. species successfully passed invasion step, species failed) in n-dimensional space. In the simplest case of two explanatory variables, the classes are plotted in two-dimensional space and the SVM algorithm finds the line ('hyper-plane') that maximizes the margin between the two classes. SVMs are generalized linear classifiers; by using appropriate nonlinear kernels, they can be applied to nonlinear classification tasks (Boser *et al.*, 1992). This is carried out by nonlinear mapping of the examples to a higher dimensional space where a linear classifier can be applied.

Many algorithms are available for constructing the hyper-plane. We used the Sequential Minimal Optimization (SMO) algorithm (Platt, 1999). We tuned the algorithm by testing different kernel types and parameters for each data set – polynomial kernel with exponents 1, 2 and 3; normalized polynomial kernel with exponents 1, 2 and 3; and radial basis function with gamma values 0.01, 0.02, 0.03, 0.1, 0.5, 1.0.

Classification Rule (CR) algorithms use the training examples to create a set of if/then statements (rules). The rules have the form 'IF *conditions* THEN *prediction*'. The rules can be constructed directly from the training examples or from other classifiers (e.g. a decision tree can be transformed into a set of rules). Each rule within a CR can apply to fewer cases than are in the full data set. We used the RIPPER (Cohen, 1995) CR learner that allowed us to set the minimum number of cases covered by a rule. We tested across minima of 2, 4, 6,….., 16 cases and chose the best classifier.

Analysis

Each algorithm produced a classifier for each data set. We used leave-one-out cross-validation to estimate classifier performance on unseen cases. This removes one species ('hold-out') from the data set, calculates the best classifier from the remaining data and tests it on the hold-out species. This is repeated for every species in the data set.

Our principal performance metric for each classifier was the area under the receiver-operator characteristic curve (AUROC; Fawcett, 2006; Flach, 2003). AUROC values higher than 0.7 indicate a good fit of model to data; values higher than 0.9 indicate extremely good fit (Pearce & Ferrier, 2000). We chose AUROC because it gives an estimate of classifier performance with respect to both outcomes (i.e. species does or does not transit invasion step) and is not sensitive to the prior distribution of outcomes. Additionally, we calculated classifier accuracy as the proportion of times that the category of the hold-out species was correctly predicted. For algorithms that give categorical predictions (DT, RF, BDT, IBL, SVM, CR), it is straightforward to determine whether the predicted class of the hold-out species was correct. For algorithms that give probabilistic predictions (LR, LDA, NB), we used a threshold of 0.5 to discriminate between predictions of success (> 0.5) and failure (< 0.5).

Three algorithms (DT, CR, SVM) have parameters that can be tuned to optimize performance. We performed parameter tuning with the MultiScheme package in Weka. MultiScheme creates classifiers over the range of possible parameter values, tests each for performance and chooses the best (according to AUROC) as the final classifier (see Appendix S1 in Supporting Information for full details, and Table S1 in Supporting Information for the parameter values used in the final classifiers).

We followed Demšar (2006) and used Friedman tests (Friedman, 1940; Iman & Davenport, 1980) to search for significant differences in performance among algorithms. This ranks algorithms according to performance on each data set and then compares average ranks. When significant differences were found, we used Nemenyi post hoc tests (Demšar, 2006) to locate the differences. This test finds a critical distance that must exist between the average ranks of two algorithms for them to be significantly different. We used a significance level of $P = 0.05$ for Nemenyi tests. The same procedures (i.e. Friedman and Nemenyi tests) were followed to search for significant differences in performance among the six data sets.

As described earlier, different algorithms used have very different ways of selecting and utilizing variables to produce classifiers. This makes a comparison across all algorithms of the number of variables chosen, and the actual variables chosen, impractical. DT and CR are the only algorithms that use variables in a comparable way. We compared the classifiers created by these algorithms to determine the extent to which they use different variables and different numbers of variables.

RESULTS

Area under the receiver-operator characteristic curve values for combinations of statistical/machine learning algorithms and data sets ranged from 0.436 to 0.941 (Table 2). The best performing algorithms were RFs and LR, each of which achieved a good fit (i.e. AUROC > 0.7) for all data sets. BDTs also had a high average AUROC but were less consistent. CRs and SVMs were the worst performing classification methods. There was not a significant difference in performance among the algorithms according to the AUROC results (Friedman test $P = 0.104$; Fig. 1a; ranks in Table S2 in Supporting Information).

Accuracy results are presented in Table 3 and generally follow the pattern of AUROC results. Again, no significant

Table 2 Classifier performance presented as area under the ROC (AUROC) for each combination of algorithm and data set (see Methods for algorithm, and Table 1 for data set, acronyms).

Data set	LR	LDA	DT	RF	BDT	IBL	NB	SVM	CR	Average
BirdNZ	0.726	0.635	0.691	0.731	0.682	0.680	0.594	0.523	0.436	0.633
BirdAU	0.724	0.845	0.583	0.745	0.681	0.785	0.864	0.808	0.581	0.735
FishCA	0.728	0.737	0.691	0.709	0.670	0.648	0.724	0.587	0.616	0.679
FishGL	0.806	0.673	0.662	0.782	0.871	0.468	0.804	0.667	0.699	0.715
MollGL	0.846	0.769	0.846	0.877	0.815	0.900	0.692	0.800	0.831	0.820
PinusG	0.808	0.781	0.741	0.941	0.930	0.811	0.895	0.696	0.825	0.825
Average	0.773	0.740	0.702	0.798	0.775	0.715	0.762	0.680	0.665	0.734

AUROC, area under the receiver-operator characteristic curve; BDT, boosted decision trees; CR, classification rule; IBL, instance-based learning; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes Classifier; RF, random forests; SVM, support vector machines.

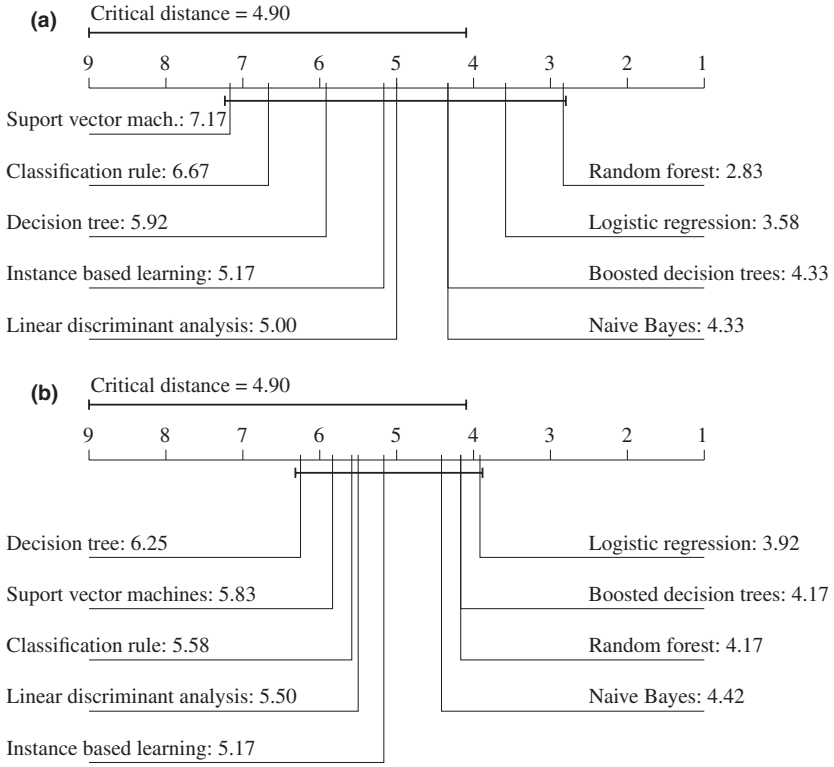

Figure 1 Average ranks diagram for the nine classification algorithms compared by (a) area under the receiver-operator characteristic curve and (b) accuracy. Algorithms that did not perform significantly differently [i.e. difference in average ranks is less than critical distance (P-value = 0.05)] are connected by a line. Numbers next to algorithm names are average ranks.

Table 3 Classifier performance presented as accuracy for each combination of algorithm and data set (see Methods for algorithm, and Table 1 for data set, acronyms).

Data set	LR	LDA	DT	RF	BDT	IBL	NB	SVM	CR	Average
BirdNZ	64.6	62.0	75.9	68.4	70.9	63.3	64.6	59.5	60.8	65.6
BirdAU	75.0	78.8	50.0	76.9	69.2	76.9	76.9	82.7	67.3	72.6
FishCA	70.1	70.1	54.0	66.7	59.8	59.8	71.3	59.8	63.2	63.9
FishGL	77.8	64.4	77.8	71.1	75.6	48.9	68.9	66.7	71.1	69.1
MollGL	88.9	83.3	88.9	83.3	88.9	94.4	77.8	88.9	88.9	87.0
PinusG	70.3	67.6	78.4	81.1	89.2	81.1	86.5	75.7	73.0	78.1
Average	74.5	71.0	70.8	74.6	75.6	70.7	74.3	72.2	70.7	72.7

BDT, boosted decision trees; CR, classification rule; IBL, instance-based learning; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes Classifier; RF, random forests; SVM, support vector machines.

difference was found in performance among algorithms (Friedman $P = 0.902$; Fig. 1b; ranks in Table S3 in Supporting Information), and the best performers were BDTs, RFs and LR, with CRs and SVMs performing worst.

Hypothesis 1 – performance of classification methods

We reject our first hypothesis that statistical algorithms (LR, DA) would be consistently outperformed by machine learning algorithms. Despite rejecting this hypothesis, we note that the *P*-value for the Friedman test of AUROC values is almost significant ($P = 0.104$) and that three methods (RFs, BDTs and LR) clearly stand out.

Hypothesis 2 – geographic range of data sets

Our second hypothesis was that data sets based on smaller recipient geographic ranges would contain organisms more likely to share traits associated with invasion and thus produce better classifiers. There were significant differences among data sets for both AUROC (Table 2; Friedman test $P < 0.0001$, ranks in Table S4 in Supporting Information) and accuracy results (Table 3; Friedman test $P < 0.0001$, ranks in Table S5 in Supporting Information). Figure 2 shows the average ranks diagram for comparison of data sets (critical distance from Nemenyi post hoc tests). *MollGL* and *FishGL* have the smallest geographic range and *PinusG* the largest. Our hypothesis predicts that the former two data sets would perform better than *PinusG* because they are based on a smaller area where the factors associated with invasion are more likely to be consistent. This trend did not occur. *PinusG*, with the largest geographic range, was the best performing data set according to AUROC. The hypothesized pattern was not evident across results for the other data sets.

Hypothesis 3 – number of species and traits

We found no positive relationship between classifier performances based on the number of species in the training data set. Data sets with highest performance (*PinusG* and *MollGL*) were the smallest (Tables 1 and 2), and data sets with the worst performance were some of the largest. Likewise, we saw no relationship between the number of traits in a data set and performance (Tables 1 and 2).

Comparison of classifiers produced by DT and CR algorithms shows large differences between the number of variables, and the actual variables, selected for classifiers. For example, the DT classifier for *PinusG* is based on one variable, while the CR is based on two separate variables. Also of interest is that the DT for *BirdsNZ* is based on four variables, while the CR includes no variable and simply predicts that all species will fail to establish. This corresponds with the low proportion of

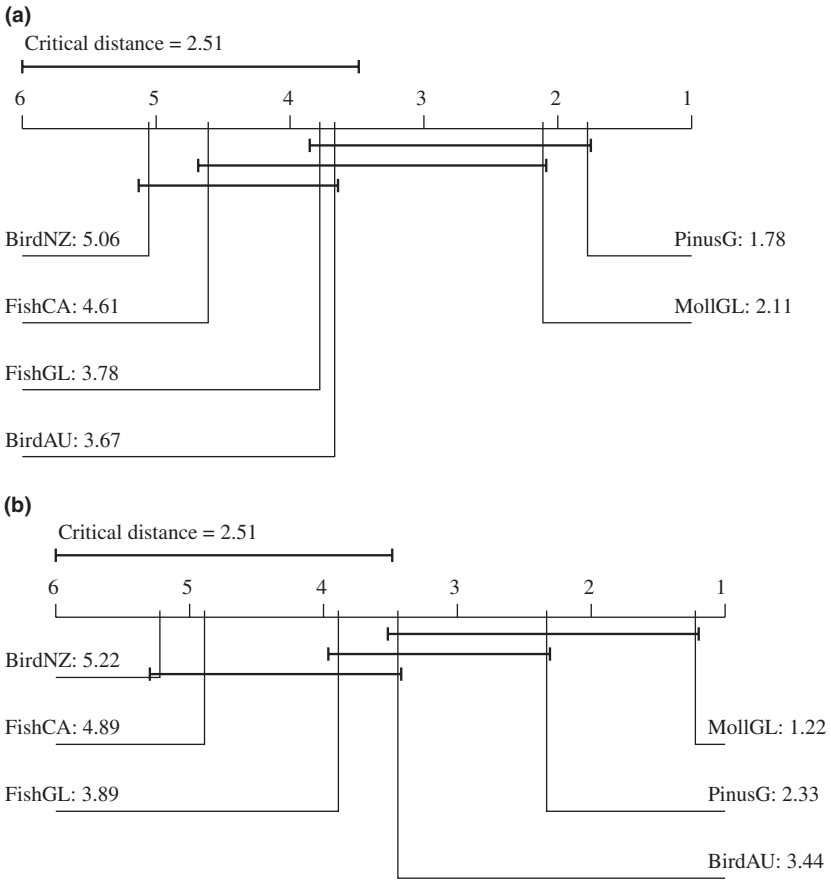

Figure 2 Average ranks diagram for the six data sets compared by (a) area under the receiver-operator characteristic curve and (b) accuracy. Data sets that did not produce classifiers with significant differences in average rank [i.e. difference in average ranks is less than critical distance (*P*-value = 0.05)] are connected by a line. Numbers next to algorithm names are average ranks.

successful establishers in that data set (Table 1; see Appendix S2 in Supporting Information for full analysis of CR and DT classifiers).

DISCUSSION

Across all data sets, there was good discrimination between successful and unsuccessful species (i.e. AUROC > 0.7) whenever LR or RF algorithms were used. The average AUROC value from all analyses (0.734, Table 2) further indicates strong relationships between species traits and success at passing through steps in the invasion sequence. In six cases, AUROC showed very strong concordance between model and data (AUROC > 0.85). This suggests that risk assessment tools based only on information that is available prior to species introduction can produce good models for determining whether a species will become invasive in the future.

Elith *et al.* (2006) performed a similar test to ours, assessing the performance of several machine learning and statistical tools for predicting the total geographic range of a species based on a subset of observed records. They found that machine learning tools, many similar to the tools used here, performed better than more traditional methods. In contrast, we found no significant differences in performance among the algorithms used. The most likely explanation for this comes from the size of the data sets used here, which are all much smaller than those used by Elith *et al.* (2006). Machine learning algorithms are generally developed and refined using data sets that have more than 1000, and often hundreds of thousands, of cases. Our largest data set contained 87 species. It is likely that in many instances the machine learning algorithms were over-fitting the data. That is, the classifiers created may have relied on patterns expressed by only a small number of species, and those patterns did not hold when the classifiers were applied to hold-out species.

Recent theory for invasive species risk assessment has emphasized that data sets assembled for smaller geographic regions should result in better performing models (Kolar & Lodge, 2001; Keller & Drake, 2009). The best performing data set in our analysis had the largest geographic range (*PinusG*), and there was no clear relationship between geographic range and performance for the remaining data sets. Recent theory has also suggested that smaller taxonomic groups should perform better. The smallest taxonomic group assessed here was also the best performing (*PinusG*). It is interesting to note that the largest taxonomic group (*MollGL*) performed extremely well and that it was also the most geographically restricted. These results hint that either small geographic range or small taxonomic group lead to high performance, but further analyses on other data sets would be required to explore this further.

Data set size, and the number of variables in each data set, had no apparent relationship to classifier performance. All else being equal, it is reasonable to expect that data sets with more species and/or variables will perform better because they contain more information and offer better insurance against over-fitting. Our results suggest that any effect of this is swamped by other factors. We emphasize here that our study is likely limited by the number of data sets ($n = 6$) and thus probably has low power to reject hypotheses two and three. Additionally, the data sets used are heterogeneous in size, taxonomic and geographic coverage and perhaps also in the quality of data. These could each confound any comparison across data sets. With regard to hypotheses two and three, we conclude that further research is needed to conclusively determine whether they are true and that they should not be rejected based solely on results presented here.

Comparison of DT and CR results shows that these algorithms produced classifiers based on different actual traits, and on different numbers of traits, for five of six data sets (Appendix S2 in Supporting Information). This demonstrates that interactions between species biology (i.e. traits) and algorithm can lead to very different classifiers, often with very different performances (Tables 1 and 2). The exception is the Great Lakes mollusk data set, for which both DT and CR classifiers perform very well (AUROC = 0.846 and 0.831, respectively) and are based on the single trait of annual fecundity per female. This is the same single trait that was chosen by LR and DT algorithms in the original analysis of this data set (Keller *et al.*, 2007b). This consistency suggests that, at least in some cases, the basic biology of the taxonomic group in question can override the large differences in how different algorithms search for patterns in the data. In turn, this means that some combinations of taxonomic group and geographic range may be more tractable for risk assessment. Further research in this area using additional data sets may discover patterns that could guide future risk assessment development.

A major goal of invasive species risk assessment is to support policies that identify and exclude species posing a high risk of becoming invasive. For the present study, we removed all variables that required data that could not be gathered prior to species introduction. This makes our assessed performances and accuracies relevant to the problem of the decision-maker working to design import policy based on ecological predictions. Keller *et al.* (2007a) constructed an economic model for determining when it is financially beneficial to apply a risk assessment as policy and applied the model to the Australian trade in ornamental plants. They found that applying risk assessment tools with accuracy > 70% creates net financial benefits over reasonable assumptions of discount rate and time horizon. Assessed by accuracy, the Californian fish data set was the worst performing, with a maximum accuracy of 71.3%. Accuracies for other data sets and algorithms were as high as 94.4%. Although these results suggest that all six data sets could provide financial benefits if incorporated into policy, we outline three reasons in the following to be cautious in this interpretation.

First, the data sets analysed here contain high proportions of species that successfully transited the invasion step in question

(0.278–0.703). These proportions are referred to as base-rates, and depending on taxonomic group and geographic region, true base-rates are generally < 10% for plants and < 50% for animals (Keller & Drake, 2009). This means that the risk assessment tools developed here will have a bias towards correctly identifying successful species. If applied by policy makers to a group of species proposed for introduction, this would result in high rates of false positives, which would reduce the financial benefits of using the risk assessment tool by preventing benign species from entering trade (see Smith *et al.* (1999) for a detailed discussion of base-rates and risk assessments).

Second, we used internal leave-one-out cross-validation to assess the performance of risk assessment tools. This method is appropriate here because the data sets are small. A more robust test, available when data sets are larger, is to reserve a subset of the data for testing (e.g. use 75% of species to generate risk assessment, test performance on remaining 25% of species). Such a test may show poorer performance of the classifiers produced. A third limitation of our data compounds this issue. All species in each data set were introduced in the past, while patterns of future species introductions may be quite different. For example, it is reasonable to expect that new suites of species will be transported around the globe as trading patterns change and especially as new regions enter the global economy. These new species may have different traits than the species analysed here, reducing the performance of the classifiers generated in this paper. In total, these three points indicate that the risk assessment performances reported here may be higher than the performance that would result from applying these tools in policy.

Predicting the impacts of introduced species is notoriously difficult, and it has been argued that the complexity of the ecological processes involved makes it essentially impossible. In this study, we took six data sets from the literature, trimmed them to only include data that are available prior to species introduction and analysed them with a range of algorithms. Our results include high performing models for each of the data sets and suggest that a reasonable fallback is to either use LR or ensemble (e.g. RF) algorithms. If only one algorithm is to be used on a particular data set, then LR is probably the best of those considered here because it has consistently high performance, is widely taught, and the results are relatively easy to interpret. Ideally, however, multiple algorithms should be used because our results show that LR is not always the highest performing.

ACKNOWLEDGEMENTS

This work was supported by the National Center for Ecological Analysis and Synthesis (NCEAS) workshop *Machine Learning for the Environment*. D.K. and S.D. are currently supported by the Slovenian Research Agency (through the research project Data Mining for Integrative Data Analysis in Systems Biology under grant J2-2285) and the European Commission (through the FP7 project PHAGOSYS Systems biology of phagosome formation and maturation – modulation by intracellular pathogens under grant number HEALTH-F4-2008-223451). S.D. is also supported by the Slovenian Research Agency (through the research program Knowledge Technologies under grant P2-0103 and the research projects Advanced machine learning methods for automated modelling of dynamic systems under grant J2-0734). He is also supported by the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins (operation no. OP13.1.1.2.02.0005 financed by the European Regional Development Fund (85%) and the Slovenian Ministry of Higher Education, Science and Technology (15%)), as well as the Jožef Stefan International Postgraduate School in Ljubljana.

REFERENCES

Aha, D., Kibler, D. & Albert, M. (1991) Instance-based learning algorithms. *Machine Learning*, **6**, 37–66.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Association for Computing Machinery (ACM) Conference on Computational Learning Theory* (ed. by D. Haussler), pp. 144–152. ACM Press, New York, USA.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

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

Caruana, R. & Niculescu-Mizil, A. (2006) An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning* (ed. by W. Cohen and A. Moore), pp. 161–168, ACM Press, New York, USA.

Cohen, W.W. (1995) Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning* (ed. by A. Prieditis and S. Russell), pp. 115–123, ACM Press, New York, USA.

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

Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, **7**, 1–30.

Dietterich, T. (2000) Ensemble methods in machine learning. *Proceedings of the 1st International Workshop on Multiple Classifier Systems* (ed. by J. Kittler and F. Roli), pp. 1–15, Springer-Verlag, Berlin, Germany.

Duncan, R.P., Bomford, M., Forsyth, D.M. & Conibear, L. (2001) High predictability in introduction outcomes and the geographical range size of introduced Australian birds: a role for climate. *Journal of Animal Ecology*, **70**, 621–632.

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

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

Flach, P.A. (2003) The geometry of ROC space: understanding machine learning metrics through ROC isometrics. *Proceedings of the 12th International Conference on Machine Learning* (ed. by A. Prieditis and S.J. Russell), pp. 194–201, Morgan Kaufmann, San Francisco, CA, USA.

Freund, Y. & Schapire, R. (1996) Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning* (ed. by L. Saitta), pp. 148–156, Morgan Kaufmann, San Francisco, CA, USA.

Friedman, M. (1940) A comparison of alternative tests of significance for the problem of *m* rankings. *Annals of Mathematical Statistics*, **11**, 86–92.

Grotkopp, E., Rejmánek, M., Sanderson, M.J. & Rost, T.L. (2004) Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. *Evolution*, **58**, 1705–1729.

Hayes, K.R. & Barry, S.C. (2008) Are there any consistent predictors of invasion success? *Biological Invasions*, **10**, 483–506.

Iman, R.L. & Davenport, J.M. (1980) Approximations of the critical region of the Friedman statistic. *Communications in Statistics*, **9**, 571–595.

Keller, R.P. & Drake, J.M. (2009) Trait-based risk assessment for invasive species. *Bioeconomics of invasive species* (ed. by R.P. Keller, D.M. Lodge, M.A. Lewis and J.F. Shogren), pp. 44–62, Oxford University Press, New York.

Keller, R.P., Lodge, D.M. & Finnoff, D.C. (2007a) Risk assessment for invasive species produces net bioeconomic benefits. *Proceedings of the National Academy of Sciences USA*, **104**, 203–207.

Keller, R.P., Drake, J.M. & Lodge, D.M. (2007b) Fecundity as a basis for risk assessment of nonindigenous freshwater molluscs. *Conservation Biology*, **21**, 191–200.

Kocev, D., Vens, C., Struyf, J. & Džeroski, S. (2007) Ensembles of multi-objective decision trees. *Proceedings of the 18th European Conference on Machine Learning* (ed. by J.N. Kok, J. Koronacki, R.L. de Mantaras, S. Matwin, D. Mladenič and A. Skowron), pp. 624–631, Springer-Verlag, Berlin, Germany.

Kolar, C.S. & Lodge, D.M. (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology and Evolution*, **16**, 199–204.

Kolar, C.S. & Lodge, D.M. (2002) Ecological predictions and risk assessment for alien fishes in North America. *Science*, **298**, 1233–1236.

Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., Carlton, J.T. & McMichaell, A. (2006) Biological invasions: recommendations for U.S. policy and management. *Ecological Applications*, **16**, 2035–2054.

Mandrak, N.E. & Cudmore, B. (2004) *Risk assessment for Asian carps in Canada*. Department of Fisheries and Oceans, Ottawa, Canada.

Marchetti, M.P., Moyle, P.B. & Levine, R. (2004) Invasive species profiling? Exploring the characteristics of non-native fishes across invasion stages in California. *Freshwater Biology*, **49**, 646–661.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Pheloung, P.C. (1995) *Determining the weed potential of new plant introductions to Australia*. Agriculture Protection Board, Perth, Australia.

Platt, J. (1999) Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel methods: support vector learning* (ed. by B. Schölkopf, C.J.C. Burges and A.J. Smola), pp. 185–208, MIT Press, Cambridge, MA, USA.

Quinlan, R. (1993) *C4.5: programs for machine learning*. Morgan Kaufmann, San Francisco, CA, USA.

Reichard, S.H. & Hamilton, C.W. (1997) Predicting invasions of woody plants introduced into North America. *Conservation Biology*, **11**, 193–203.

Richardson, D.M., Cowling, R.M. & Le Maitre, D.C. (1990) Assessing the risk of invasive success in *Pinus* and *Banksia* in South African mountain Fynbos. *Journal of Vegetation Science*, **1**, 629–642.

Smith, C.S., Lonsdale, W.M. & Fortune, J. (1999) When to ignore advice: invasion predictions and decision theory. *Biological Invasions*, **1**, 89–96.

Veltman, C.J., Nee, S. & Crawley, M.J. (1996) Correlates of introduction success in exotic New Zealand birds. *The American Naturalist*, **147**, 542–557.

Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*. Springer, New York, NY, USA.

Witten, I.H. & Frank, E. (2005) *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, USA.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Parameter setting methods for machine learning tools.

Appendix S2 Comparison of DT and CR classifiers.

Table S1 Tuning parameters selected for machine learning algorithms.

Table S2 Ranks of algorithm performance by AUROC.

Table S3 Ranks of algorithm performance by accuracy.

Table S4 Ranks of data set performance by AUROC.

Table S5 Ranks of data set performance by accuracy.

Tables S6–S11 Full data sets used (see Table 1).

BIOSKETCH

Reuben Keller is Henry Chandler Cowles Lecturer in the Program on the Global Environment at the University of Chicago. His research focuses on interactions between ecology and globalization that lead to invasions, predicting the ecological and economic impacts of invasive species and developing bioeconomic models to inform intervention strategies. Dragi Kocev is a research assistant at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. His PhD research concerns machine learning and data mining, prediction of structured outputs and their applications in environmental and life sciences. Sašo Džeroski is a scientific councillor at the Department of Knowledge Technologies, Jožef Stefan Institute, and the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, and a professor at the Jožef Stefan International Postgraduate School, all in Ljubljana, Slovenia. His research interests include data mining and machine learning and their applications in environmental sciences (ecology) and life sciences (biomedicine).

Author contributions: R.P.K. and S.D. identified questions and designed research. R.P.K. gathered data. D.K. performed data analysis and prepared figures. R.P.K. wrote the manuscript with input from D.K. and S.D.

Editor: Mark Burgman