Contrasting determinants for the introduction and invasion success of exotic birds in Taiwan using decision trees models (#14802)

First submission

Please read the **Important notes** below, the **Review guidance** on page 2 and our **Standout reviewing tips** on page 3. When ready **submit online**. The manuscript starts on page 4.

Important notes

Editor and deadline

Michael Wink / 25 Dec 2016

Files	1 Figure file(s) 1 Raw data file(s) Please visit the overview page to **download and review** the files not included in this review PDF.
Declarations	No notable declarations are present

Review guidelines

Please read in full before you begin

How to review

When ready **submit your review online**. The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

To finish, enter your editorial recommendation (accept, revise or reject) and submit.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to **PeerJ standards**, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see **PeerJ policy**).

EXPERIMENTAL DESIGN

- Original primary research within **Scope of the journal**.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- Data is robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.
- Speculation is welcome, but should be identified as such.

The above is the editorial criteria summary. To view in full visit **https://peerj.com/about/editorial-criteria/**

7 Standout reviewing tips

The best reviewers use these techniques

Tip	***Example***
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that your international audience can clearly understand your text. I suggest that you have a native English speaking colleague review your manuscript. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Give specific suggestions on how to improve the manuscript	*Line 56: Note that experimental data on sprawling animals needs to be updated. Line 66: Please consider exchanging “modern” with “cursorial”.*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Contrasting determinants for the introduction and invasion success of exotic birds in Taiwan using decision trees models

Shih-Hsiung Liang [1] , **Bruno Andreas Walther** [2] , **Bao-Sen Shieh** [Corresp. 3]

[1] Department of Biotechnology, National Kaohsiung Normal University, Kaohsiung, Taiwan
[2] Master Program in Global Health and Development, College of Public Health, Taipei Medical University, Taipei, Taiwan
[3] Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung, Taiwan

Corresponding Author: Bao-Sen Shieh
Email address: bsshieh@kmu.edu.tw

Background. Biological invasions have become a major threat to biodiversity, and identifying determinants underlying success at different stages of the invasion process is essential for both prevention management and testing ecological theories. To investigate variables associated with different stages of the invasion process in a local region such as Taiwan, potential problems using traditional parametric analyses include too many variables of different data types (nominal, ordinal, and interval) and a relatively small data set with too many missing values. **Methods.** We therefore used five decision tree models instead and compared their performance. Our dataset contains 283 exotic bird species which were transported to Taiwan; of these 283 species, 95 species escaped to the field successfully (introductions success); of these 95 introduced species, 36 species reproduced in the field of Taiwan successfully (establishment success). For each species, we collected 22 variables associated with human selectivity and species traits which may determine success during the introduction stage and establishment stage. For each decision tree model, we performed three variable treatments: (I) including all 22 variables, (II) excluding nominal variables, and (III) excluding nominal variables and replacing ordinal values with binary ones. Five performance measures were used to compare models, namely, area under the receiver operating characteristic curve (AUROC), specificity, precision, recall, and accuracy. **Results.**The gradient boosting models performed best overall among the five decision tree models for both introduction and establishment success and across variable treatments. The most important variables for predicting introduction success were the bird family, the number of invaded countries, and variables associated with environmental adaptation, whereas the most important variables for predicting establishment success were the number of invaded countries and variables associated with reproduction. **Discussion.** Our final optimal models achieved relatively high performance values, and we discuss differences in performance with regard to sample size and variable treatments. Our results showed that, for both the establishment model and introduction model, the number of invaded countries was the most important or second most important determinant, respectively. Therefore, we suggest that future success for introduction and establishment of exotic birds may be gauged by simply looking at previous success in invading other countries. Finally, we found that species traits related to reproduction were more important in establishment models than in introduction models; importantly, these determinants were not averaged but either minimum or maximum values of species traits. Therefore, we suggest that in addition to averaged values, reproductive potential represented by minimum and maximum values of species traits should be considered in invasion studies.

Contrasting determinants for the introduction and invasion success of exotic birds in Taiwan using decision trees models

Shih-Hsiung Liang[1], Bruno Andreas Walther[2], Bao-Sen Shieh[3*]

[1] Department of Biotechnology, National Kaohsiung Normal University, 62 Sanchung Rd., Yanchao Township, Kaohsiung 824, Taiwan; [2] Master Program in Global Health and Development, College of Public Health, Taipei Medical University, 250 Wu-Hsing St., Taipei 110, Taiwan; [3] Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, 100 Shihchuan 1st Road, Kaohsiung 807, Taiwan

Corresponding author:

Bao-Sen Shieh

e-mail address: bsshieh@kmu.edu.tw

Tel: 886-7-3121101ext.2703

Fax: 886-7-3227508

Abstract

Background. Biological invasions have become a major threat to biodiversity, and identifying determinants underlying success at different stages of the invasion process is essential for both prevention management and testing ecological theories. To investigate variables associated with different stages of the invasion process in a local region such as Taiwan, potential problems using traditional parametric analyses include too many variables of different data types (nominal, ordinal, and interval) and a relatively small data set with too many missing values.

Methods. We therefore used five decision tree models instead and compared their performance. Our dataset contains 283 exotic bird species which were transported to Taiwan; of these 283 species, 95 species escaped to the field successfully (introductions success); of these 95 introduced species, 36 species reproduced in the field of Taiwan successfully (establishment success). For each species, we collected 22 variables associated with human selectivity and species traits which may determine success during the introduction stage and establishment stage. For each decision tree model, we performed three variable treatments: (I) including all 22 variables, (II) excluding nominal variables, and (III) excluding nominal variables and replacing ordinal values with binary ones. Five performance measures were used to compare models, namely, area under the receiver operating characteristic curve (AUROC), specificity, precision, recall, and accuracy.

Results. The gradient boosting models performed best overall among the five decision tree models for both introduction and establishment success and across variable treatments. The most important variables for predicting introduction success were the bird family, the number of invaded countries, and variables associated with environmental adaptation, whereas the most

important variables for predicting establishment success were the number of invaded countries and variables associated with reproduction.

Discussion. Our final optimal models achieved relatively high performance values, and we discuss differences in performance with regard to sample size and variable treatments. Our results showed that, for both the establishment model and introduction model, the number of invaded countries was the most important or second most important determinant, respectively. Therefore, we suggest that future success for introduction and establishment of exotic birds may be gauged by simply looking at previous success in invading other countries. Finally, we found that species traits related to reproduction were more important in establishment models than in introduction models; importantly, these determinants were not averaged but either minimum or maximum values of species traits. Therefore, we suggest that in addition to averaged values, reproductive potential represented by minimum and maximum values of species traits should be considered in invasion studies.

Introduction

Biological invasions have become a major threat to biodiversity (*Pimentel, Zuniga & Morrison, 2005*). Hence, some studies of biological invasion have focused on how to prevent the invasion or how to eradicate the invasive species (*Dana, Jeschke & García-de-Lomas, 2014*). As more and more invasive species have spread into the wild, invasive species have also become important subjects in testing ecological theories in relation to niche and competition (e.g., *Broennimann et al., 2007; Alen et al., 2015*). Both prevention management and testing ecological theories require the identification of the key factors underlying success at different stages in the invasion process (*Duncan, Blackburn & Sol, 2003*); moreover, factors that are important to explain the invasion success have been suggested to be different at each stage of the invasion process (*Kolar & Lodge, 2002; Williamson, 2006; Dawson, Burslem & Hulme, 2009*).

Compared with other vertebrate taxa, birds have a higher number of invasive species and invasion success rates in a study focusing on Europe and North America (*Jeschke & Strayer, 2006*). Previous studies on exotic birds have identified two major categories of factors associated with their success at the introduction and establishment stages: human selectivity factors and species traits. Human selectivity factors consist of factors such as taxa and geography selected non-randomly by humans during the transport or introduction stages of exotic birds (*Duncan, Blackburn & Sol, 2003*). Species traits, on the other hand, then play an important role during the introduction and establishment stages (*Blackburn, Cassey & Lockwood, 2009*).

In Taiwan, at least 290 exotic species of pet birds have been imported, and a 9.7% rate of invasion success was estimated (*Shieh et al., 2006*). For the transport stage, non-random selectivity of exotic birds imported to Taiwan was associated with bird family, native geographic range, body size, and song production of species (*Su, Cassey & Blackburn, 2014*); as to the later stages of invasion, pet trade factors such as song attractiveness were significantly associated with introduction success but not establishment success (*Su, Cassey and Blackburn, 2016*).

For the exotic birds of Taiwan, species traits that help to avoid stochastic extinction or to constrain establishment (cf. *Sol, 2008*) have not been investigated with regard to their influences on different stages of the invasion process. To investigate the effects of these factors which are associated with both human selectivity and species traits onto different stages of the invasion process in a local region such as Taiwan, two potential problems using traditional parametric analyses have been identified as (1) a relatively small data set with too many missing values and (2) too many variables of different types (nominal, ordinal, and interval).

Machine learning is a new, advanced analytical method which overcomes many of the restrictions of traditional parametric analyses. We chose the decision tree method, a machine learning algorithm, because its advantages include no need to input data for missing values and no assumptions about the distribution of the data; therefore, this method is ideal for dealing with mixed data types, such as nominal, ordinal and interval variables (*Olinsky, Kennedy & Brayton Kennedy, 2014*). In studies of biological invasion, the decision tree method was first applied to investigating a data set of 45 fish species for risk assessment in the Great Lakes (*Kolar & Lodge, 2002*). In another recent study, *Chen, Peng & Yang (2015)* found that decision tree methods not only work best with nominal variables but also have higher performance values than traditional parametric methods in predicting alien herb invasion. In a comparative study of trait-based risk assessment for invasive species, *Keller, Kocev & Džeroski (2011)* found that random forests (an ensemble method that creates multiple decision tree sub-models) was one of the two best performing methods. To our knowledge, there are very few studies so far which have been using decision tree methods in invasion studies of exotic birds.

Consequently, we decided to use decision tree methods to assess factors associated with human selectivity and species traits which determine the success during the introduction and establishment stages of exotic birds in Taiwan. We used five decision tree models which differed in regard to resampling the data set and compared their performance. An optimal prediction model was chosen based on five performance measures, and the relative importance of factors in the optimal model for introduction success and for establishment success was examined and compared.

Materials & Methods

Species of the data set

The four stages of the invasion process were defined in *Duncan, Blackburn & Sol (2003)* as transport, introduction, establishment, and spread. In this study, we focused on the introduction and establishment stages. For a species to reach the introduction stage, it must have passed the transport stage. Therefore, we selected all the exotic species which had been transported to Taiwan main island (not including surrounding islands, such as Lanyu and Kinmen Island) as documented in *Shieh et al. (2006*) which included the results of *Chi* (1995), *Severinghaus* (1999) and *Lin* (2004). Whether a transported species has passed the subsequent stages of the invasion process was based (1) on escaping records in the field (introduction success) and (2) breeding record in the field (establishment success). We followed the detailed definitions and methods of how to establish introduction success and establishment success which were given in *Su, Cassey and Blackburn (2016*). However, *Su, Cassey and Blackburn (2016)* based their decision of establishment success on the respective species having been recorded to be breeding at least twice; instead, we based it on at least one record of fledglings actually having left the nest successfully.

In order to record all the escaping and breeding records of bird species up to 2015, we continuously (1) checked information from the Chinese Wild Bird Federation (http://www.bird.org.tw/) database which is the main collector of wild bird data in Taiwan, as well as other Taiwanese websites dedicated to natural history observations of birds, (2) remained in contact with local ornithologists, birdwatchers and bird societies, and (3) included any relevant publications (e.g., *Walther, 2011; Walther 2014* for red-whiskered bulbul, *Pycnonotus jocosus*; *Fan et al., 2009* for white-rumped shama, *Copsychus malabaricus*, or *Shieh, Lin & Liang, 2016* for Asian glossy starling, *Aplonis panayensis*). Most of this updated information was published recently in a project report for the Taiwan Forestry Bureau (*Liang & Shieh, 2015*).

Despite following the methods as described in *Su, Cassey and Blackburn (2016*), we independently collected all the data used in this analysis beginning in 2004 and ending in 2015. Our dataset thus contains 283 full species (although we entered subspecies in our dataset, for this analysis, we only used full species) which were transported to Taiwan (see above). Of these 283 species, 95 species escaped to the field successfully (introductions success). Of these 95 species, 36 species reproduced in the field of Taiwan successfully (establishment success) (see supplementary file Table S1 for species list).

Variables

We collected 22 variables for each species, including two nominal ones (order and family taxa), six ordinal ones (latitude overlap with Taiwan: 0～2, migration pattern: 0～3, nesting location: 0～3, feeding: 1～3, diet: 1～6, and habitat: 0～6), three binary ones (hole nest, Taiwan genus_resident, dichromatism), and 11 interval ones (clutch size: Clutch, maximum clutch size: Mclutch, incubation days: Incubation, minimum incubation days: Minincubation, body length: Length, maximum body length: Mlength, body mass: Mass, maximum body mass:

Mmass, the number of invaded countries: Invcontry_Max, distribution range (km^2): Range, the number of subspecies: subspecies) (see supplementary file Table S2 for code descriptions of variables). The variable Taiwan genus_resident was based on the information in *Hsiao & Li (2014)*. For the other variables, we gathered species information from the books of *del Hoyo et al. (1992-2011)*, *Dunning Jr. (1993)*, and internet databases of IUCN (www.iucn.org) and BirdLife International Datazone (http://datazone.birdlife.org) (see supplementary file Table S1 for associated information of each species and Table S2 for code descriptions of variables). When we collected the values for reproduction and body size for each species, we usually found a given range instead of fixed values in the references. In order to account for the maximum adaptation and reproduction potential in the invasion process, we used maximum values such as maximum body mass or minimum values such as minimum incubation days in addition to averaged values. To determine the number of invaded countries (Invcontry_Max), we counted the total (or maximum) number of countries in which occurrences of feral populations of each respective species were reported.

Decision trees models and variable treatments

To investigate the possible effects of nominal variables (family and order) and ordinal variables on the performance of the decision tree models, we conducted three variable treatments for modeling: (I) including all variables, (II) excluding nominal variables, and (III) excluding nominal variables and replacing ordinal values with binary ones; e.g. changing habitat values of 0～4 to 0 (natural habitats) and habitat values of 5～6 to 1 (artificial habitats).

For each variable treatment, we used five decision tree models (DT_no bagging, DT_bagging 90%, DT_bagging 100%, gradient boosting, and HP forest) to predict the outcomes of introduction and establishment, respectively. Modeling processes and comparisons of model performance were implemented using SAS Enterprise Miner 13.1 (for diagrams of process flow, see Supplementary Figure S1). Because of the small data set, no data partition was implemented; that is, all data were used as training data. Instead, other methods, such as bagging and cross validation, which have been suggested for the use with small data sets *(SAS Institute Inc., 2013)*, were used in the present study.

DT_no bagging is the traditional classification tree method by constructing a layered tree model with the following settings: splitting rule = Gini, cross validation with 10 subsets and 100 repeats. The DT_bagging 90% and DT_bagging 100% methods used the same setting of splitting rule and cross validation as the DT_no bagging method but with bagging 90% or 100% of the data set for 50 times, respectively. Gradient boosting is a boosting method that resamples the data set to produce a series of decision trees which together form a single predictive model which has been found to be less prone to overfitting the data than a single decision tree (*Georges, 2009*). HP Forest is the random forest method which builds many parallel trees forming a forest; a tree in the forest is a sample without replacement from all the available observations, and the input variables that are considered for splitting a node are randomly selected from all the available inputs *(Hall et al., 2014)*.

We calculated five performance measures to compare models, namely, the area under the receiver operating characteristic curve (AUROC), the specificity which measures the fraction of negative events that were correctly labeled, the precision which measures the fraction of positively labeled outcomes that were correctly labeled, the recall which measures the fraction of positive events that were correctly labeled, and the accuracy which measures the fraction of all events that were correctly labeled (accuracy = 1- misclassification rate) (*Sohngen, Chang & Schomburg, 2011*). The higher the values of these five performance measures are, the better the model performs; therefore, we summed up the five values (called the "total score" from hereupon) and chose the model with the highest sum as our final optimal model. We then compared the relative importance of each of the variables in the optimal introduction model and establishment model.

For illustrative purposes, we chose the visual output of the resulting trees of DT_no bagging of variable treatment I for our figures (Figures 2-3). Such visual outputs are not possible for the other four methods (namely, DT_bagging 90%, DT_bagging 100%, gradient boosting, and HP forest).

We used the decision tree models described above to build various versions of two kinds of models: (1) introduction success prediction models and (2) establishment success prediction models. However, for brevity's sake, from hereon we will call them introduction models and establishment models, respectively.

Results

Across the three variable treatments and for both the introduction models (Table 1) and establishment models (Table 2), the gradient boosting models always achieved the highest score among the five decision tree models (i.e., it performed best overall). However, this overall best performance did not mean that gradient boosting always performed best when comparing values of the five performance measures. For instance, Table 1 (see also supplementary file Fig. S3 for receiver operating characteristic curves, and supplementary file Fig. S5 for classification charts) shows that gradient boosting only performed best for accuracy in variable treatment I and II; otherwise, other models always performed better using the other four performance measures. Nevertheless, across all three treatments, the total score is always highest for gradient boosting for the introduction models (Table 1).

For the establishment models (Table 2, see also supplementary file Fig. S4 for receiver operating characteristic curves, and supplementary file Fig. S6 for classification charts), however, gradient boosting has the highest total score for all the three treatments and also for most of the five performance measures (the only exceptions being specificity and precision in variable treatment II). Therefore, we considered gradient boosting the optimal model for both the introduction models and establishment models and only considered its results from hereupon.

Looking across the three different variable treatment methods I-III, gradient boosting performed best with variable treatment I for the introduction models (Table 1) as well as the establishment models (Table 2). For variable treatments II and III, the total score decreased by only 0.169 (4%) and 0.128 (3%), respectively. We also note that this decreasing trend across variable treatments is maintained for most of the five performance measures. Furthermore, the values of the performance measures are all > 0.7 and 60% are > 0.9, which means that the performance was consistently high or very high.

In the optimal introduction model, family and the number of invaded countries (Invcontry_Max) were the most important variables, and their relative importance values were 1 and 0.888, respectively (Fig. 1). The top six variables with an importance value > 0.3 also included maximum body mass (0.394), order (0.384), latitude overlap with Taiwan (0.354), and distribution range (0.345). For the introduction model based on the classification tree method (Fig. 2), the number of invaded countries was the most important determinant, as it appeared at the top of the tree, which means that the 84 species with any record of invading other countries had a 66.7% chance of successful introduction to Taiwan. Among these 84 species, the 72 species which had a migration pattern categorized as sedentary (0), local movement (1) or partial migration (2) had a 73.6% chance of successful introduction, while the 12 species categorized as migrants (3) had only a 25.0 % chance of successful introduction. Among the 199 species which no record of invading other countries, family was chosen as an important determinant of successful introduction.

In the optimal establishment model, the number of invaded countries and distribution range were the most important variables, and their relative importance values were 1 and 0.826, respectively (Fig. 2). The top six variables with an importance value > 0.6 also included minimum incubation days (Minincub, 0.647), migration pattern (Migration, 0.633), clutch size (Clutch, 0.62), and habitat type (Habitat, 0.616). The relative importance of the variable family decreased to 0.569 which is therefore much lower than in the optimal introduction model (see above). For the establishment model based on the classification tree method (Fig. 3), the number of invaded countries was again the most important determinant, as it appeared at the top of the tree. However, in this case it means that the 39 species with a record of invading at least two countries had a 59.0% chance of successful establishment in Taiwan, while the 56 species with a record of invading fewer than two countries had only a 23.2% chance of successful establishment. Among the 39 species noted above, the 21 species with a maximum clutch size (Mclutch) < 5.5 had an 81.0% chance of successful establishment, while the other 18 with a maximum clutch size of ≥ 5.5 had only a 33.3% chance of successful establishment. Finally, among the 56 species noted above, the eight species with a body length (Length) ≥ 36.5 cm had a 62.5% chance of successful establishment.

Discussion

Model comparisons and variable treatment comparisons

Our results showed that for the complete data set of 283 transported species or for the data set of 95 introduced species, the gradient boosting method performed better than the other four decision tree methods. While we calculated five performance measures, the only other study which used the decision tree method on a bird data set was *Keller, Kocev & Džeroski (2011)* who calculated only AUROC and accuracy values. Considering AUROC values first, the AUROC values of gradient boosting of our study were over 0.919 in the introduction models and over 0.971 in the establishment models; thus, they were all higher than our values for the random forests method. This is in contrast to the results of *Keller, Kocev & Džeroski (2011)* who found that, based on the AUROC values, random forests performed better than gradient boosting for both their New Zealand and Australia bird data sets. Specifically, AUROC values for gradient boosting for their New Zealand (79 species with 11 traits) and Australia (52 species with 11 traits) data sets were 0.682 and 0.681, respectively, whereas AUROC values for random forests were 0.731 and 0.745, respectively. *Pearce & Ferrier (2000)* suggested that AUROC values between 0.7 and 0.9 indicate a reasonable discrimination ability of models, and values higher that 0.9 indicate a very good discrimination ability of models. The higher AUROC values of our study might have resulted from the inclusion of more variables (up to 22 variables) rather than larger samples used for analysis. In our study, both the introduction model and establishment model used 22 variables, and we found higher AUROC values (0.971-0.985) in the smaller data set (namely, the establishment model with 95 species) than in the larger data set (namely, the introduction model with 283 species) (AUROC values 0.919-0.936). We therefore suggest that even a small data set (less than 100 species) with up to 22 variables can achieve a prediction model of good performance using the gradient boosting method.

Comparing the performances of variable treatment I with variable treatments II and III, we found little difference on model performance. Treatment II excluded nominal variables, and treatment III changed ordinal variables of species traits into binary variables, but neither one of these changes really had much discernable influence on overall performance. Our results therefore provided evidence to support the use of ordinal variables of species traits, and that there is no need to convert ordinal variables of species traits to binary ones for their use in decision tree models.

Predictors of introduction and establishment success in exotic birds

Perhaps the most interesting and novel result of our study is that, for both the establishment model and introduction model, the number of previously invaded countries was the most important or second most important determinant in all the models presented in our Results. Therefore, our study suggests that future success for introduction and establishment of birds can be gauged by simply looking at previous success in invading other countries or regions. Future studies should include this variable to confirm our supposition because it might be a very simple and straightforward way to predict the potential invasion success of a species: if it has been successful before, it will probably be successful again. While this variable could not have been established a few decades ago, we now have a global track record of successful species invasions, and we might therefore be able to use it to predict future local or regional invasions.

Another important determinant was family taxon. While family was the most important variable in the optimal introduction model, it dropped to being only the seventh most important variable in the optimal establishment model. In other words, family was an important determinant of introduction but not establishment in Taiwan. Our results thus differ from those of a global study which found that bird family was also a good predictor for establishment success (*Lockwood 1999)*. The discrepancy between this study and our study could result from the fact that exotic birds in Taiwan are primarily introduced for aesthetic reasons but not for hunting (*Shieh et al., 2006; Su, Cassey & Blackburn, 2016*), while the global data set included many hunted species.

Several species traits were also chosen as important determinants for the introduction and establishment models. For the optimal introduction model, the top three selected species traits were maximum body mass (Mmass), latitude overlap with Taiwan (Overlap), and distribution range (Range). Among these three variables, maximum body mass was ranked the most important, and it also had a relative importance greater than that of two other closely related measures, specifically, the averaged body mass (Mass) and body length (Length). Therefore, we suggest that including maximum body mass may be important in order not to miss a potentially important determinant. For example, *Su, Cassey & Blackburn (2016*) did not find that body mass had any influence on introduction success. However, they only used averaged body mass, and perhaps their result would have been different if they had also included maximum body mass. Furthermore, *Cassey's (2001)* global study found that averaged body mass was significantly correlated with introduction success which further supports the role of some measure of body mass being an important determinant of introduction success.

Finally, several species traits related to reproduction were also important, such as minimum incubation days (Minincub), clutch size (Clutch), dichromatism, and nesting locations (Nesting); however, these determinants were more important in establishment success than in introduction success. Furthermore, given that some top ranking variables were associated with maximum or minimum values of species traits, we suggest that in addition to averaged values, reproductive potential represented by minimum and maximum values of species traits should be considered in prediction models of invasion studies.

We conclude that decision tree models are efficient for the analysis of small data sets with mixed types of variables, including nominal, ordinal and interval variables in predicting the invasion success of exotic birds. Our results further demonstrate that the most important determinants in predicting introduction success of exotic birds in Taiwan were the bird family, the number of invaded countries, and variables associated with environmental adaptation, whereas the most important determinants in predicting establishment success were the number of invaded countries and variables associated with reproduction.

Data availability

File: Supplementary S1 (Table S1-S2).xls

Table S1. Species list and associated variable information

Table S2. Code description of variable

Supplemental Information

File: Supplementary S2 (Figure S1-S6).pdf

Figure S1. Diagrams of modeling introduction success using SAS Enterprise Miner: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

Figure S2. Diagrams of modeling establishment success using SAS Enterprise Miner: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

Figure S3. Receiver operating characteristic (ROC) curves of five introduction models: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

Figure S4. Receiver operating characteristic (ROC) curves of five establishment models: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

Figure S5. Classification charts of five introduction models: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

Figure S6. Classifications charts of five establishment models: (a) variable treatment I, (b) variable treatment II, and (c) variable treatment III.

References

Allen CR, Angeler DG, Moulton MP, Holling CS. 2015. The importance of scaling for detecting community patterns: success and failure in assemblages of introduced species. *Diversity* **7:** 229-241.

Blackburn TM, Cassey P, Lockwood JL. 2009. The role of species traits in the establishment success of exotic birds. *Global Change Biology* **15:** 2852-2860.

Broennimann O, Treier UA, Müller-Schärer H, Thuiller W, Peterson AT, Guisan A. 2007. Evidence of climatic niche shift during biological invasion. *Ecology letters* **10:** 701-709.

Cassey P. 2001. Are there body size implications for the success of globally introduced land birds? *Ecography* **24:** 413-420.

Chen L, Peng S, Yang B. 2015. Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. *Biological Invasions* **17:** 2187-2198.

Chi W-L. 1995. An investigation report on pet bird trade in Taiwan. Green Consumer's Foundation. Taipei, Taiwan. (In Chinese)

Dana ED, Jeschke JM, García-de-Lomas J. 2014. Decision tools for managing biological invasions: existing biases and future needs. *Oryx* **48:** 56-63.

Dawson W, Burslem DF, Hulme PE. 2009. Factors explaining alien plant invasion success in a tropical ecosystem differ at each stage of invasion. *Journal of Ecology* **97:** 657-665.

del Hoyo J, Elliott A, Sargatal J, Christie DA. 1992-2011. *Handbook of the birds of the world, vol. 1–16.* Barcelona: Lynx Edicions.

Duncan RP, Blackburn, TM, & Sol D .2003. The ecology of bird introductions. *Annual Review of Ecology, Evolution, and Systematics* 71-98.

Dunning Jr, JB. 1993. *CRC handbook of avian body masses*. CRC press, Boca Raton, Florida.

Fan M-W, Lin R-S, Fang W & Lin Y-H. 2009. The distribution and abundance of the alien invasive white-rumped shama (*Copsychus malabaricus*) in Taiwan. *Taiwania* **54:** 248-254.

Hall P, Dean J, Kabul IK, Silva J. 2014. *An Overview of Machine Learning with SAS® Enterprise Miner™*. Cary: SAS Institute Inc.

Hsiao M-C, Li C-L. 2014. *Field Guide of Wild Bird in Taiwan.* Taipei: Wild Bird Society of Taiwan.

Jeschke JM, Strayer DL. 2006. Determinants of vertebrate invasion success in Europe and North America. *Global Change Biology* **12:** 1608-1619.

Lin R-S. 2004. The black terrors of Hitchcock – current status of invasive birds in Taiwan. *Nature Conservation Quarterly* **48:** 38-43. (In Chinese)

Keller RP, Kocev D, Džeroski S. 2011. Trait-based risk assessment for invasive species: high performance across diverse taxonomic groups, geographic ranges and machine learning/statistical tools. *Diversity and Distributions* **17:** 451-461.

Kolar CS, Lodge DM. 2002. Ecological predictions and risk assessment for alien fishes in North America. *Science* **298(5596):** 1233-1236.

Liang S-H, Shieh B-S. 2015. Application and anaysis of the life history database to assess invasion impact of exotic animals (in Chinese with English abstract). *Available at http://conservation.forest.gov.tw/File.aspx?fno=61675*.

Lockwood JL (1999). Using taxonomy to predict success among introduced avifauna: relative importance of transport and establishment. *Conservation Biology* **13(3):** 560-567.

Olinsky A, Kennedy K, Brayton Kennedy B. 2014. Assessing gradient boosting in the reduction of misclassification error in the prediction of success for actuarial majors. *Case Studies In Business, Industry And Government Statistics* **5:** 12-16.

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling* **133:** 225-245.

Pimentel D, Zuniga R, Morrison D. 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological economics* **52:** 273-288.

SAS Institute Inc. 2013. *SAS® Enterprise Miner™ 13.1 Extension Nodes: Developer's Guide.* Cary: SAS Institute Inc.

Severinghaus L-L. 1999. Exotic birds in Taiwan. *Wild Birds* **7:** 45-58.

Shieh B, Lin Y, Lee T, Chang C, Cheng K. 2006. Pet trade as sources of introduced bird species in Taiwan. *Taiwania* **51:** 81-86.

Shieh B-S, Lin C-J, Liang S-H. 2016. Breeding biology of the invasive Asian Glossy Starling (*Aplonis panayensis*) in urban parks of Kaohsiung City, southern Taiwan. *Taiwan J For Sci* **31:** 63-70.

Söhngen C, Chang A, Schomburg D. 2011. Development of a classification scheme for disease-related enzyme information. *BMC bioinformatics* **12:** 329.

Sol D. 2008. Do successful invaders exist? Pre-adaptations to novel environments in terrestrial vertebrates. In *Biological invasions* (pp. 127-141). Springer Berlin Heidelberg.

Su S, Cassey P, Blackburn TM. 2014. Patterns of non-randomness in the composition and characteristics of the Taiwanese bird trade. *Biological Invasions* **16:** 2563-2575.

Su S, Cassey P, Blackburn TM. 2016. The wildlife pet trade as a driver of introduction and establishment in alien birds in Taiwan. *Biological Invasions* **18:** 215-229.

Walther, B.A. (2011) First documented nesting of the red-whiskered bulbul *Pycnonotus jocosus* in Taiwan. TW J. *Biodivers* **13:** 121-133.

Walther, B.A. (2014) Second documented nesting of the red-whiskered bulbul *Pycnonotus jocosus* in Taiwan. *Feather* **266:** 20-25.

Williamson, M. (2006). Explaining and predicting the success of invading species at different stages of invasion. *Biological Invasions* **8:** 1561-1568.

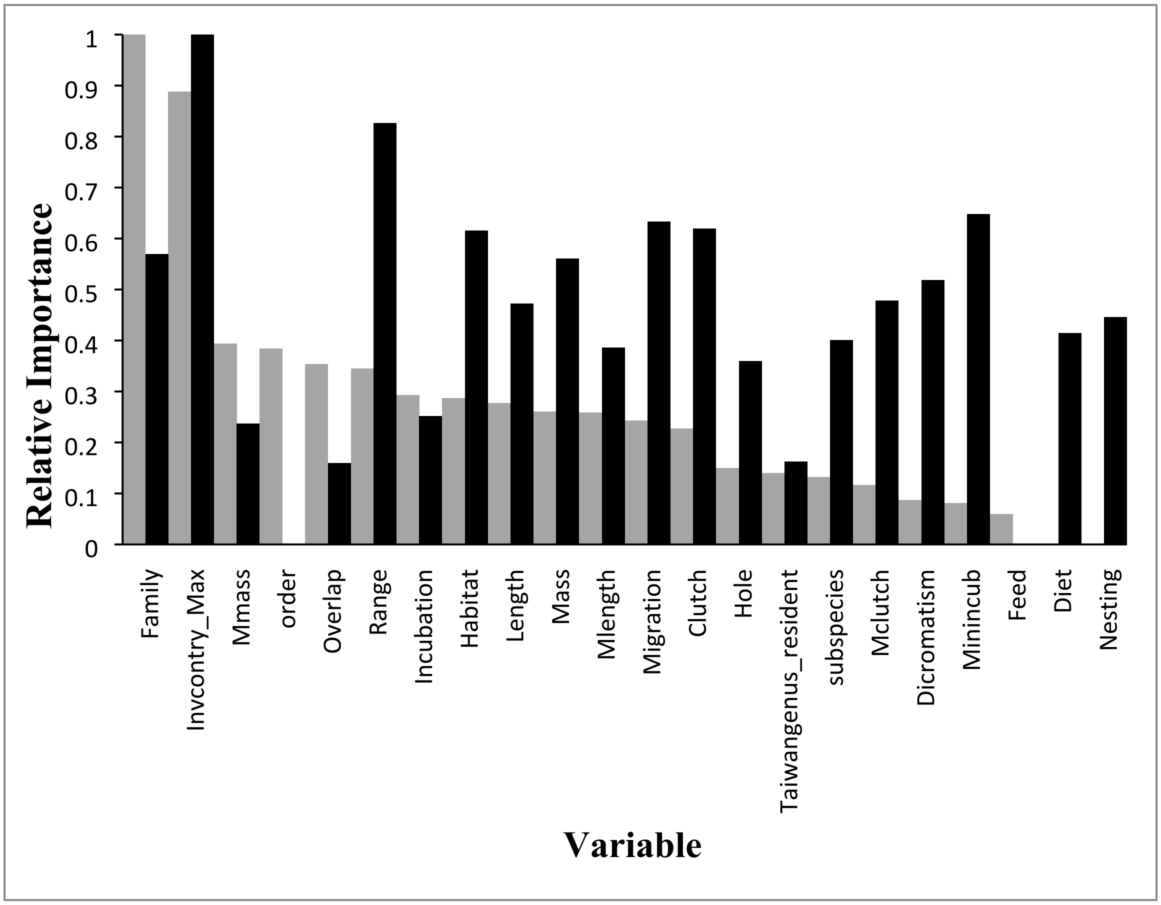

Fig. 1. Relative importance of variables in the prediction models using gradient boosting approach (grey bars for introduction models and black bars for establishment models). For descriptions of codes for variables, see supplementary Table S2.

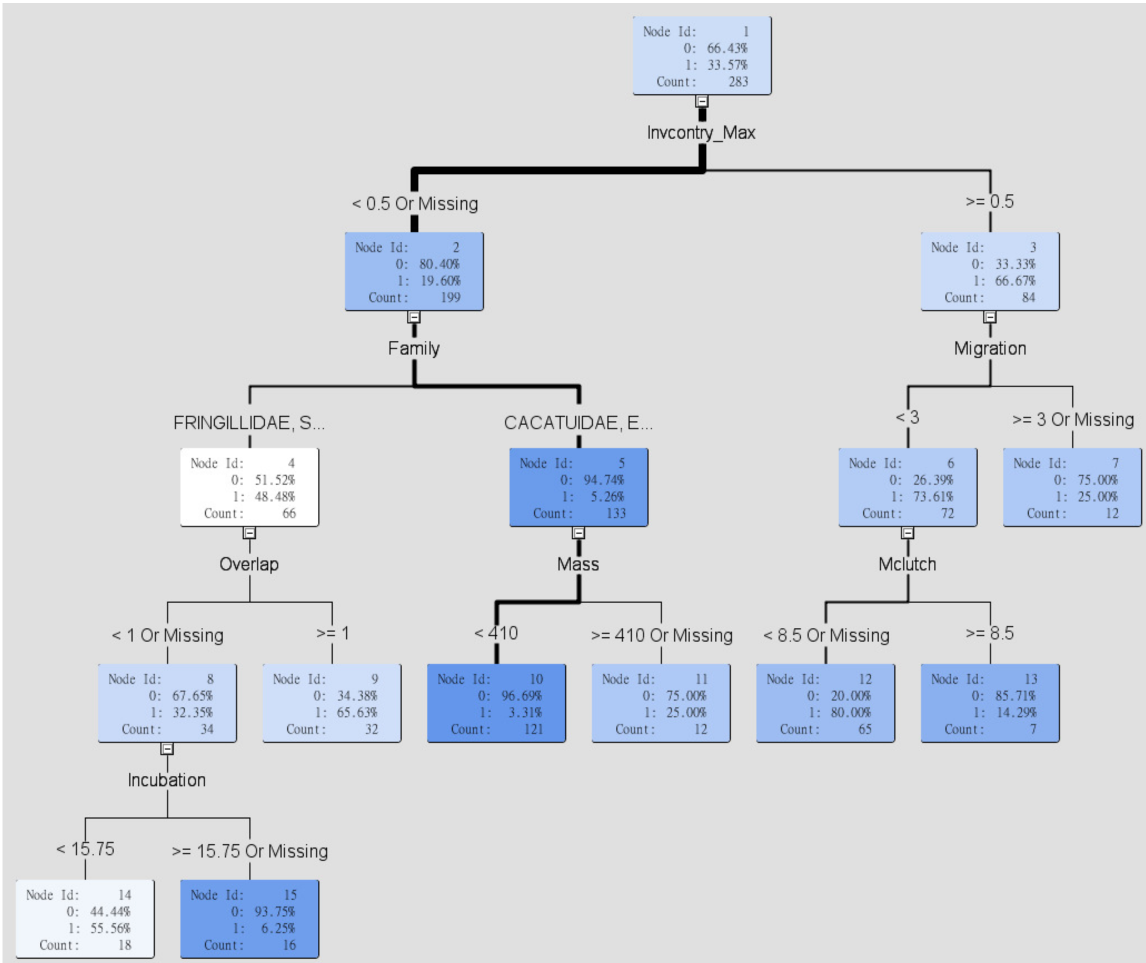

Fig. 2. The visual output of the introduction model based on the classification tree method for exotic birds of Taiwan generated from the dataset of 283 transported species, of which 95 species successfully escaped in the field (see supplementary file Table S1 for associated information of each species and Table S2 for code descriptions of variables)

.

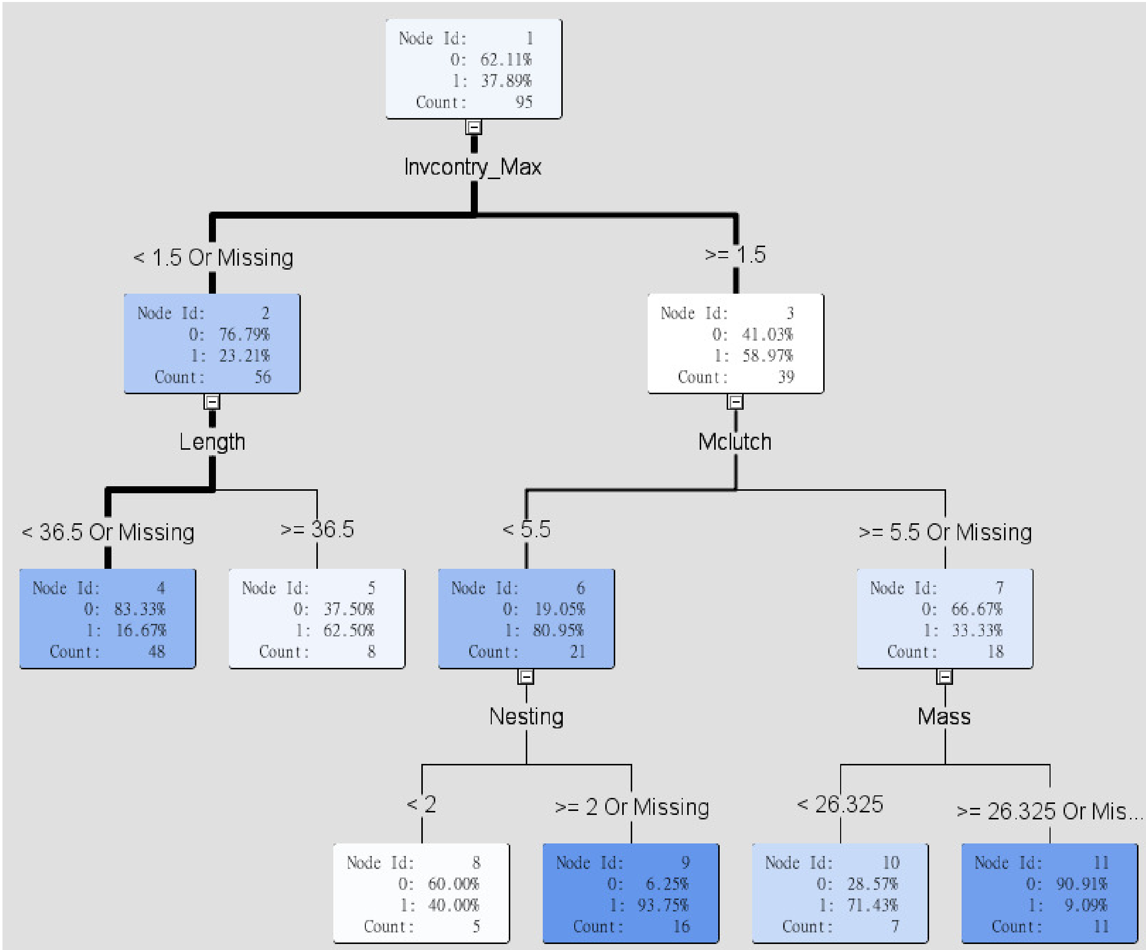

Fig. 3. The visual output of the establishment model based on the classification tree method for exotic birds of Taiwan generated from the dataset of 95 introduced species, of which 36 species successfully reproduced in the field (see supplementary file Table S1 for associated information of each species and Table S2 for code descriptions of variables)

Table 1. Comparison of five performance measures among five introduction models of exotic birds in Taiwan, separately for three variable treatments (see Methods for details).

Variable Treatment I

Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.894	0.830	0.722	0.874	0.845	4.164
DT_bagging 90%	0.970	0.936	0.782	0.453	0.774	3.914
DT_bagging 100%	0.976	0.910	0.742	0.516	0.777	3.921
Gradient Boosting	0.936	0.941	0.869	0.768	0.883	4.398
HP Forest	0.903	0.963	0.873	0.505	0.809	4.053

Variable Treatment II

Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.904	0.872	0.765	0.821	0.855	4.217
DT_bagging 90%	0.949	0.899	0.683	0.432	0.742	3.705
DT_bagging 100%	0.955	0.910	0.742	0.516	0.777	3.900
Gradient Boosting	0.924	0.915	0.816	0.747	0.859	4.261
HP Forest	0.894	0.963	0.848	0.411	0.777	3.893

Variable Treatment III

Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.910	0.888	0.781	0.789	0.855	4.224
DT_bagging 90%	0.946	0.910	0.691	0.400	0.739	3.685
DT_bagging 100%	0.953	0.888	0.700	0.516	0.763	3.820
Gradient Boosting	0.919	0.926	0.827	0.705	0.852	4.229
HP Forest	0.888	0.957	0.840	0.442	0.784	3.912

Table 2. Comparison of five performance measures among five establishment models of exotic birds in Taiwan, separately for three variable treatments (see Methods for details).

Variable Treatment I						
Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.839	0.898	0.806	0.694	0.821	4.059
DT_bagging 90%	0.945	0.932	0.800	0.444	0.747	3.869
DT_bagging 100%	0.963	0.949	0.842	0.444	0.758	3.957
Gradient Boosting	0.985	1.000	1.000	0.861	0.947	4.793
HP Forest	0.901	0.983	0.875	0.194	0.684	3.638
Variable Treatment II						
Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.839	0.898	0.806	0.694	0.821	4.059
DT_bagging 90%	0.942	0.932	0.800	0.444	0.747	3.866
DT_bagging 100%	0.963	0.949	0.842	0.444	0.758	3.957
Gradient Boosting	0.976	0.983	0.969	0.861	0.937	4.726
HP Forest	0.914	1.000	1.000	0.167	0.684	3.765
Variable Treatment III						
Model	AUROC	Specificity	Precision	Recall	Accuracy	Total
DT_no bagging	0.839	0.898	0.806	0.694	0.821	4.059
DT_bagging 90%	0.936	0.932	0.800	0.444	0.747	3.860
DT_bagging 100%	0.940	0.949	0.842	0.444	0.758	3.934
Gradient Boosting	0.971	1.000	1.000	0.778	0.916	4.665
HP Forest	0.912	1.000	1.000	0.139	0.674	3.725