GfÖ
GfÖ Ecological Society of Germany,
Austria and Switzerland

Basic and Applied Ecology 14 (2013) 309–319

Basic and Applied Ecology

www.elsevier.com/locate/baae

Influences of temporal independence of data on modelling species distributions

Chia-Ying Ko[a,b,c], Chie-Jen Ko[d], Ruey-Shing Lin[e], Pei-Fen Lee[d,*]

[a] *Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA*
[b] *School of Forestry and Environmental Studies, Yale University, New Haven, CT 06511, USA*
[c] *Delta Electronics Foundation, Taipei 114, Taiwan*
[d] *Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei 106, Taiwan*
[e] *Endemic Species Research Institute, Nantou 552, Taiwan*

Received 22 June 2012; accepted 4 April 2013
Available online 8 May 2013

Abstract

Modelling species distributions has been widely used to understand present and future potential distributions of species, and can provide adaptation and mitigation information as references for conservation and management under climate change. However, various methods of data splitting to develop and validate functions of the models do not get enough attention, which may mislead the interpretation of predicted results. We used the Taiwanese endemic birds to test the influences of temporal independence of datasets on model performance and prediction. Training and testing data were considered to be independent if they were collected during different survey periods (1993–2004 and 2009–2010). The results indicated no significant differences of six model performance measures (AUC, kappa, TSS, accuracy, sensitivity, and specificity) among the combinations of training and testing datasets. Both species- and grid cell-based assessments differed significantly between predictions by the annual and pooled training data. We also found an average of 85.8% similarity for species presences and absences in different survey periods. The remaining dissimilarity was mostly caused by species observed in the late survey period but not in the early one. The method of data splitting, yielding training and testing data, is critical for resulting model species distributions. Even if similar model performance exists, different methods can lead to different species distributional maps. More attention needs to be given to this issue, especially when amplifying these models to project species distributions in a changing world.

Zusammenfassung

Die Modellierung der Verbreitung von Arten wurde weithin angewendet, um die gegenwärtige und die zukünftige potentielle Verbreitung von Arten zu verstehen. Sie kann auch Informationen zu Anpassung und Vorbeugung als Bezugspunkte für Naturschutz und Management in Hinblick auf den Klimawandel liefern. Indessen wird verschiedenen Methoden der Datenaufteilung für die Entwicklung und Validierung der Funktionen von Modellen nicht genügend Aufmerksamkeit geschenkt, was zu irrigen Interpretationen der Vorhersageergebnisse führen kann. Wir wählten die endemischen Vögel Taiwans, um den Einfluss zeitlicher Unabhängigkeit der Datensätze auf Modellleistung und -vorhersage zu prüfen. Die Trainings- und Testdaten wurden als unabhängig angesehen, wenn sie aus unterschiedlichen Erfassungsperioden stammten (1993–2004 bzw. 2009–2010). Die Ergebnisse zeigten für sechs Maßzahlen der Modellleistung (AUC, kappa, TSS, accuracy, sensitivity und specificity) keine signifikanten Unterschiede zwischen den getesteten Kombinationen von Trainings- und Testdatensätzen. Sowohl

[*] Corresponding author at: Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, 106, Taiwan. Tel.: +886 2 3366 2469; fax: +886 2 2362 3501.
E-mail address: leepf@ntu.edu.tw (P.-F. Lee).

http://dx.doi.org/10.1016/j.baae.2013.04.003

artbezogene als auch Rasterzellen-basierte Schätzungen differierten signifikant hinsichtlich der Vorhersagen, wenn Beobachtungen aus Einzeljahren bzw. über die Erfassungsperiode kumulierte Beobachtungen zugrundegelegt wurden. Wir fanden auch eine durchschnittliche Ähnlichkeit der Artenidentität von 85.8% zwischen den beiden Erfassungsperioden. Die verbleibende Unähnlichkeit wurde hauptsächlich durch einen Zugewinn an Arten in der späteren Erfassungsperiode verursacht. Die Vorgehensweise bei der Datenaufteilung, die die Trainings- und Testdatensätze ergibt, ist entscheidend für die aus dem Modell resultierenden Verbreitungen der Arten. Selbst bei gleicher Leistungsfähigkeit der Modelle können unterschiedliche Methoden zu unterschiedlichen Verbreitungskarten führen. Größere Aufmerksamkeit muss diesem Umstand gewidmet werden, insbesondere, wenn diese Modelle erweitert werden, um die Verbreitung von Arten in einer sich wandelnden Welt vorauszuberechnen.

Keywords: Temporal independence; Data splitting; Species distribution model; Model performance; Endemism

Introduction

Species distribution models are increasingly used and regarded as multidirectional applications for conservation and management, especially mitigation and adaptation strategies under climate change such as prioritization of sites for species and monitoring of species declines and expansions in range (Guisan & Zimmermann 2000; Hijmans & Graham 2006; Crawford & Hoagland 2010; Elith, Kearney, & Phillips 2010). Several studies have addressed the usage of different models including regression models, machine learning models, and maximum entropy, and compared advantages and disadvantages among them (Elith et al. 2006; Austin 2007; Meynard & Quinn 2007; Phillips et al. 2009). Simultaneously, these studies advance to the development of mature models, which are more adequate for different species groups.

With so much need to target conservation planning as efficiently as possible, optimizing the methodology becomes an imperative for species distribution models. A good set of ecological data is the basis for successful species distribution models, yet there is a surprising lack of guidance for such data use. For instance, should data collected from different years be used separately or unitedly? Besides, needs for independent data – data for different subjects that do not depend on each other – from the ecological data sets are of critical importance to be used for statistical hypothesis tests. The independent data also have a potential influence on model performance as well as subsequent interpretation of model predictions. However, in ecology, the independent data are difficult to define because the entire ecosystem is interrelated and can be regarded as one “subject”. Even if we assume that each species, site or survey period is an independent “subject” in the ecosystem, the restrictions on manpower, funding, and the time series for field surveys make “completely” independent data difficult to be obtained. Additionally, ecological literature appears to have paid less attention to how independent and dependent data may influence the accuracy of species distribution models.

A simple way to define “independent data” for modelling species distributions is to assume samples from different years are independent and to develop/train models by such samples. Museum records, private collections and historical literature that cover long periods of time and contain a vast source of information on species distributions are commonly used to develop/train species distribution models (Newbold 2010). However, these consecutive-years data are often combined together for training models instead of using data from individual years. For using the consecutive-years data together, it has clearly been biased spatially, environmentally, temporally, and taxonomically in the data and there are still major gaps in our knowledge (Soberon, Llorente, & Onate, 2000; Newbold 2010). Results analyzed by questionable data splitting, such as general uses of the aforementioned data, may reflect sampling effort more than real ecological phenomena (Newbold 2010). Thus, splitting data adequately and correctly before using them to develop/train models is important to avoid biases and misinterpretation.

Model performance is an important indicator for the credibility of each species distribution model (Guisan & Zimmermann 2000). Recent studies have compared the performance of different models that are expected to predict species distributions more precisely to face rapid anthropogenic habitat destruction (Fielding & Bell 1997; Segurado & Araujo 2004; Allouche, Tsoar, & Kadmon, 2006; Austin 2007). It has been suggested that evaluating the nature of prediction errors (*i.e.* false positives/false negatives or commission/omission error) to assess prediction success might be more advantageous than evaluating the overall percentage accuracy, which has a restricted set of error measures and has been commonly used in the late 20th century. The generally accepted methods are ROC (receiver operating characteristics) plots and measures derived from a binary confusion matrix (Swets 1996; Fielding & Bell 1997; Zou, O'Malley, & Mauri 2007). Moreover, goals of model predictions in application to conservation and/or management and requirements for certain accuracy further affect choices and reliability of the above methods. For example, if a model is used to predict the impacts on endangered species, a false positive might be of greater concern. However, if species distributions are predicted as references for environmental or urban development, then false negatives need to be taken into account. Using different methods for model performance evaluation is, therefore, helpful for understanding overall predictive capability of models.

In the process of model performance evaluation, data used in the assessment of models (*i.e.* reference/testing data) have major impacts on the accuracy and interpretation of model results (Foody 2011). These reference/testing data can be independent or dependent of the data used for developing/training models (*i.e.* training data). Using data from different survey periods for model training and model testing, respectively, is a simple way to ensure independence. However, temporally independent reference/testing data are always not easy to be additionally obtained. Besides, as far as we know differences between the usages of temporally independent and dependent reference/testing data have not been thoroughly compared.

In Taiwan, a breeding bird survey has been implemented for over 15 years, and a systematic investigation, which is called "BBS Taiwan" and formed by non-governmental organizations, academic institutions, and government agencies, has been established in 2009 (Hsu, Yao, Lin, Yang, & Lai 2004; Koh, Lee, & Lin 2006; Lee et al. 2010). A core concept of the BBS Taiwan is to encourage the public to participate in the studies of ecological science and conservation and to promote a large-scale and long-term bird survey. Observers have been well trained prior to the survey, and they continue to investigate bird species' populations two to three times during the breeding season. Thus, the BBS Taiwan is a suitable database to not only monitor changes of species populations and distributions but also to estimate influences of sub-sets of data on the species distribution models.

Our study focused on the influences of temporal independence of datasets on model performance and predictions. The training data were considered to be dependent if they were extracted using records pooled for all years, different with using annual records. On the other hand, the testing data were considered to be independent if they were collected during different survey periods.

Materials and methods

Study area

Taiwan Island, which lies across the ocean from mainland China, has frequent geological activities. The Taiwan Island experienced numerous disturbances in the Earth's crust, which formed various landscapes and over 100 mountains 3000 m above sea level. The geographic coordinates (21°53′–25°18′ N latitude and 120°08′–122°01′ E longitude; Fig. 1) of the Taiwan Island span an area of 36,000 sq km. Two-thirds of the island is mountainous, and one-third is lowland. Because (1) the Taiwan Island is located at the transition zone between the Holarctic and Palaeotropical Kingdoms, and (2) it has a high productivity generated by year-round high temperatures and heavy rainfall, a high level of biodiversity is formed. Accordingly, the Taiwan Island has been regarded as the diversity center of East Asia and owns a wide variety of endemic species (Lei, Qu, Lu, & Yin 2003).

Sampling species

Seventeen Taiwanese endemic bird species belonging to ten families were used in this study (Table 1). We categorized these 17 species as common, uncommon, and rare species based on number of grid cells where species had been sighted and heard (Ko, Lin, & Lee 2010). The common species are present in more than 200 grid cells, the uncommon species are present in 100–200 grid cells, and the rare species are present in fewer than 100 grid cells. In this grid cell system, the Mikado Pheasant and Swinhoe's Pheasant are categorized as rare species, which is congruent with the IUCN Red List Category (IUCN 2012), which lists both pheasant species as near-threatened species. These endemic bird species are distributed across all ranges of elevation and territorial areas in Taiwan and generally occupy habitats with high vegetation cover and low human disturbance, except for the Styan's Bulbul, which favors high-road density areas (Ko et al., 2010).

Sampling methods

To generate dependent and independent datasets, we compiled inventories of the avifauna during two time periods: 1993–2004 (Hsu et al. 2004; Koh et al. 2006) and 2009–2010 (Lee et al. 2010). All data were obtained by point- and transect-counting techniques based on distance sampling methods (Buckland, Anderson, Burnham, & Laake 1993; Buckland, Goudie, & Borchers 2000). The point-counting method involved a 1500- to 2000 m-long transect with 10 sampling sites 150 m apart in 1993–2004 and to 6–10 sampling sites 200 m apart since 2009. Birds were recorded in each sampling site for a 6 min period. The transect-counting method used a fixed observation route, which was 3 km long and walked at a consistent speed of 1.5 km/hour without stopping. In 1993–2004, the sites were sampled once per breeding season (*i.e.* March to June in Taiwan). In 2009–2010, the sites were sampled mainly during the breeding season. Because the altitude led to slight differences in the bird breeding season in Taiwan, the sampling season in 2009–2010 was divided into three subdivisions: the areas reaching maximum altitudes below 1500 m a.s.l. were surveyed during March and May, the areas situated between 1500 and 2500 m a.s.l. were surveyed during April and June, and the areas above altitudes of 2500 m a.s.l. were surveyed during May and June, respectively. Although the two census methods differ in precision of the bird density estimate (Buckland 2006), two advantages have been noted for the simultaneous use of the different methods: (1) increasing overall precision concerning density, and (2) providing the possibility to decompose components during the detection processes (Nichols, Thomas, & Conn 2008). We did not

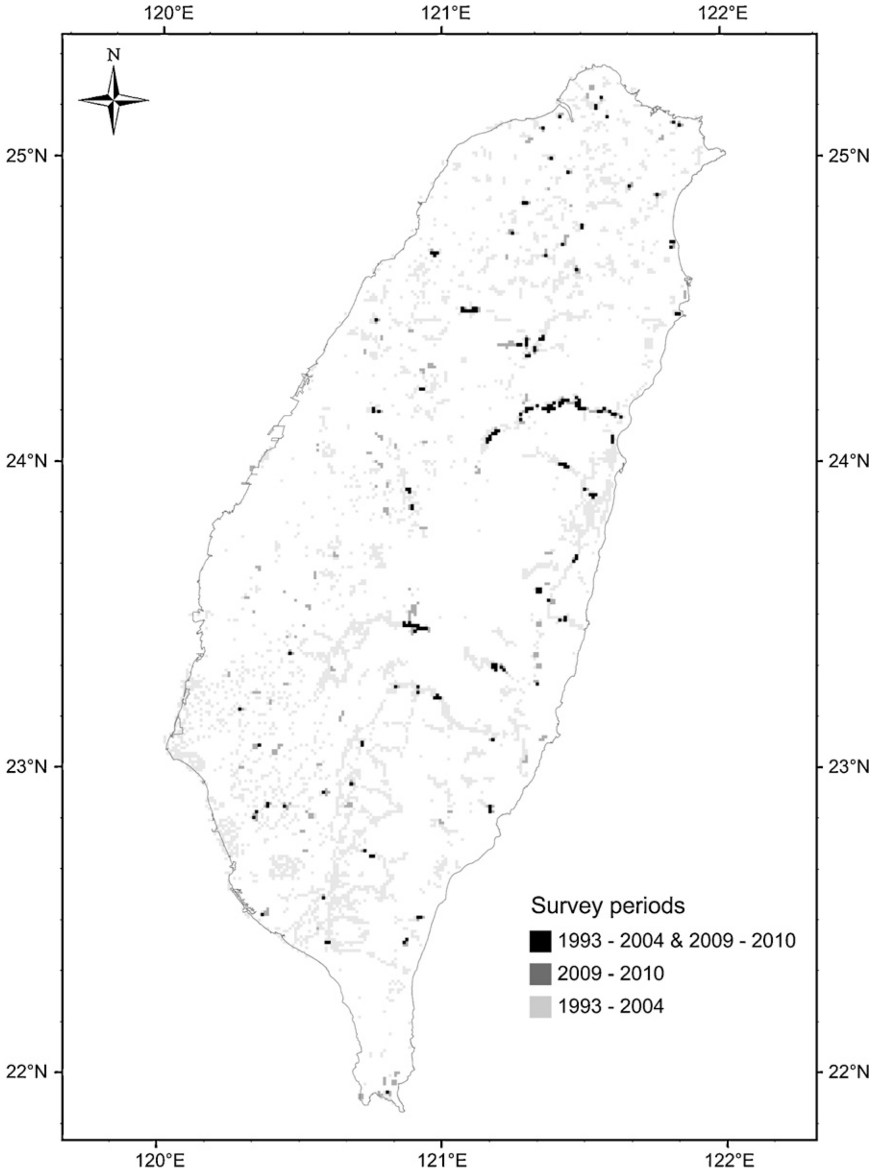

Fig. 1. Geographical location of the Taiwan Island and 1 km resolution grid cells investigated in 1993–2004 and 2009–2010.

differentiate between results obtained by the two census methods.

Observations of birds and geographic coordinates of individual sampling sites were recorded. All geographic coordinates were then transferred to a grid of 1 km × 1 km cells. In total, 4082 grid cells were sampled.

Modelling species distributions

We used a presence-only model: Maximum Entropy (Maxent) for modelling in this study. Its algorithm assigns a non-negative probability of species occurrence to each grid cell in the study area (Phillips, Dudik, & Schapire 2004; Phillips, Anderson, & Schapire 2006). We chose 22 environmental variables (Su 1992; Koh et al. 2006; Ko, Lin, Ding, Hsieh, & Lee 2009; Ko, Root, & Lee 2011; see Appendix A: Table 1 for full information on the variables) to run the Maxent model. Most environmental variables had a low correlation with any of the others estimated by univariate analysis (see Appendix A: Table 2), and all variables were left for modelling.

Four combinations of data splitting to understand how temporal independence may influence model performance and predictive distributions of species were used and compared in the analyses. For training models, we randomly selected 80% of presences of individual species in 1993–2004. The 80% data were extracted using two methods: annual records (YY) and records pooled for all years (AYs). The 80% presences of a species for each year were separately selected and then combined together in the YY-method. However, in the AYs-method, all presences of a species were combined

Table 1. List of the 17 Taiwanese endemic bird species considered in this study. The species are categorized as common (>200; C), uncommon (100–200; U) or rare (<100; R) based on number of grid cells are occupied (# cells). Nomenclature follows Clements et al. (2011).

Family	English name	Scientific name	Cat.	# cells
Corvidae	Formosan Magpie	*Urocissa caerulea*	U	156
Megalaimidae	Taiwan Barbet	*Megalaima nuchalis*	C	1657
Megaluridae	Taiwan Bush-Warbler	*Bradypterus alishanensis*	U	135
Muscicapidae	Collared Bush-Robin	*Tarsiger johnstoniae*	C	202
Paridae	Yellow Tit	*Macholophus holsti*	U	151
Phasianidae	Taiwan Partridge	*Arborophila crudigularis*	C	408
	Mikado Pheasant	*Syrmaticus mikado*	R	30
	Swinhoe’s Pheasant	*Lophura swinhoii*	R	95
Pycnonotidae	Styan’s Bulbul	*Pycnonotus taivanus*	C	414
Regulidae	Flamecrest	*Regulus goodfellowi*	U	305
Timaliidae	Taiwan Hwamei	*Garrulax taewanus*	C	482
	White-whiskered Laughingthrush	*Garrulax morrisonianus*	C	207
	White-eared Sibia	*Heterophasia auricularis*	C	779
	Steere’s Liocichla	*Liocichla steerii*	C	673
	Taiwan Yuhina	*Yuhina brunneiceps*	C	772
	Taiwan Barwing	*Actinodura morrisoniana*	U	102
Turdidae	Formosan Whistling-Thrush	*Myophonus insularis*	C	481

together and then 80% of the data were randomly selected. For model testing, we used the remaining 20% of data from 1993 to 2004 or all data from 2009 to 2010. The models, including both training and testing, were repeated 100 times. We used the default setting in Maxent, including all features (*i.e.* linear features, quadratic features, product feature, threshold features, and hinge features) and using logistic format as output values (Phillips & Dudik 2008).

In short, the training data selected from the annual records for the 1993–2004 survey period (YY-method) and the testing data recorded during a different survey period (2009–2010) were regarded as temporally independent data. On the other hand, the training data selected randomly from the data pooled for all years 1993–2004 (AYs-method) and the remaining data from the same survey period, which were used for model testing, were regarded as temporally dependent data.

Model evaluation, comparison and statistics

We used six measures to estimate model performance and compared differences among 2 (the training data extracted according to the YY- or AYs-method) × 2 (the testing data from the 1993–2004 and 2009–2010 survey period, respectively) dataset combinations. Values of an area under the ROC curve (AUC), kappa, accuracy, sensitivity, specificity, and true skills statistic (TSS) were evaluated. The six measures individually place different emphases, such as quantifying omission/commission error or both, on the model results (details see below; Allouche et al. 2006), which can reveal advantages and disadvantages of predictive results modelled by each dataset combination.

The AUC is a threshold as well as prevalence independent accuracy measure of species distribution models and equivalent to the probability that a species distribution model will rank a randomly chosen species presence site higher than a randomly chosen absence site (Swets 1996; Zou et al. 2007). The remaining five measures are generated from a confusion/error matrix (Fielding & Bell 1997; see Appendix A: Tables 3 and 4), which includes true and false positives and negatives by applying a certain threshold to transform the probabilities into a dichotomous set of presence–absence predictions and constructing the matrix. Kappa and TSS take into account both omission and commission errors in one parameter and their relative tolerance to zero values in the confusion matrix. Both of them range from +1 to −1, where +1 indicates perfect agreement and values of zero or less indicate a model performance no better than random (Cohen 1960). The main difference between kappa and TSS is dependency on prevalence. TSS is proposed as being immune to prevalence and to have all of the advantages of kappa (Allouche et al. 2006). Accuracy (*i.e.* overall accuracy) is calculated as an overall proportion of observed presences and absences that are predicted correctly. Sensitivity and specificity are calculated, respectively, as the proportion of observed presences or observed absences that are predicted correctly. Simply, sensitivity quantifies omission error, and specificity quantifies commission error. The above three measures are independent of each other and of prevalence. Values of accuracy, sensitivity and specificity are in the range from 0 to +1, where +1 represents a good model performance (*i.e.* a perfect species distribution prediction). A sensitivity-specificity sum maximization threshold of each species was used to transform probabilistic predictions into presence–absence predictions.

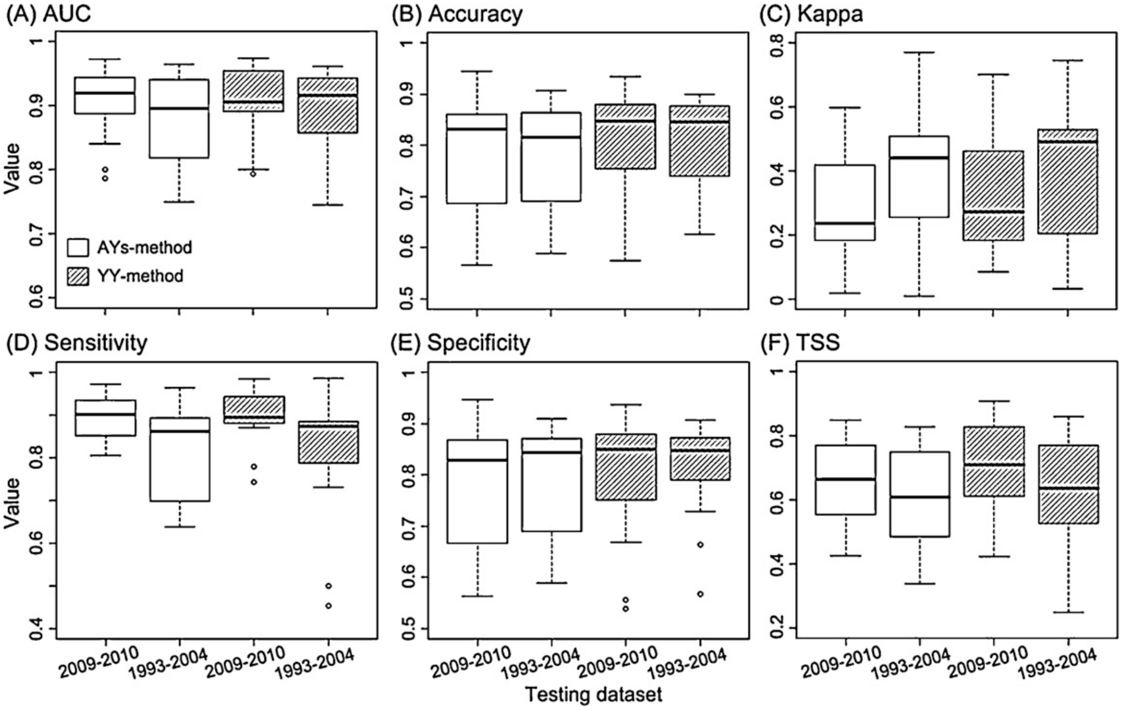

Fig. 2. Six model performance measures, including AUC, accuracy, kappa, sensitivity, specificity, and TSS, calculated for four combinations of datasets. The training data were extracted using two methods: annual records (YY-method) and records pooled for all years (AYs-method). The testing data were surveyed in 1993–2004 and 2009–2010, respectively.

To compare predicted results of the models trained by the YY- and AYs-method, we used both species- and grid cell-based assessments. The sum of probabilistic presences of individual species, range sizes of individual species, average of probabilistic presences of all endemic species of each grid cell, and predicted species richness of each grid cell were calculated to assess effects of using the temporally independent or dependent data.

The sum of probabilistic presences of individual species represents an entire probability of species' presences in an area while the range sizes represent the extent of occurrence of a species that is often measured by a minimum convex polygon of the present occurrence of the species (Gaston & Fuller 2009). In other words, the range sizes of a species can be considered the amount of sub-areas within the area (the Taiwan Island in our case) that are occupied by the species. Similarly, the average of probabilistic presences of all endemic species of each grid cell indicates the probability of all endemic species being present in a grid cell. The predicted species richness of each grid cell is the number of species predicted to be present in that grid cell.

We conducted the six model performance measures among the combinations using ANOVA. We estimated comparisons between predictions by the YY- and AYs-method using a paired student's *t* test and calculated correlations at the level of species- and grid cell-based assessments, respectively. The model running and evaluation and all statistical analyses were done in R 2.12 (R Development Core Team 2010).

Results

There were 4082 and 745 (out of 37,552) grid cells being investigated in 1993–2004 and 2009–2010, respectively (Fig. 1). Among these grid cells, 280 grid cells were investigated in both survey periods.

Among the six model performance measures, AUC represented the highest averaged values, followed by sensitivity, specificity, accuracy, TSS, and kappa values (Fig. 2). There were no significant differences of the values of each measure among the combinations (ANOVA, d.f. = 3, all $p > 0.05$). Predictions by the AYs-method had slightly higher model performance than those by the YY-method, while each measure estimated by the testing data in 1993–2004 and 2009–2010, respectively, showed incongruent model performance (Table 2).

The sum of probabilistic presences and the range sizes of individual species demonstrated significant differences between predictions by the YY- and AYs-method (paired *t* test, both $p < 0.001$). Although few (4 out of 17) species showed greater sum of probabilistic presences predicted by the YY-method than the AYs-method, most of species (9 out of 17) were predicted to occupy larger range sizes by the YY-method than by the AYs-method (Fig. 3A and B).

Table 2. Six model performance measures estimated separately by the temporally independent and dependent datasets as the mean for the 17 Taiwanese endemic bird species.

Performance measures	Training data (1993–2004)		Testing data	
	AYs-method	YY-method	1993–2004	2009–2010
AUC	0.90	0.89	0.91	0.89
Accuracy	0.82	0.79	0.79	0.81
Kappa	0.36	0.35	0.32	0.39
Sensitivity	0.86	0.85	0.90	0.82
Specificity	0.81	0.79	0.79	0.82
TSS	0.68	0.65	0.68	0.64

In other words, a species having a high sum of probabilistic presences might not be predicted to occupy large range sizes after transforming the probabilistic predictions into the presence–absence predictions. The sum of probabilistic presences of individual species predicted by the YY- and AYs-method, respectively, was highly correlated (Fig. 3A; $r^2 = 0.91$) whereas the predicted range sizes of individual species indicated a low correlation between the two methods (Fig. 3B; $r^2 = 0.15$). There were no specific trends among the rare, uncommon, and common species. Likewise,

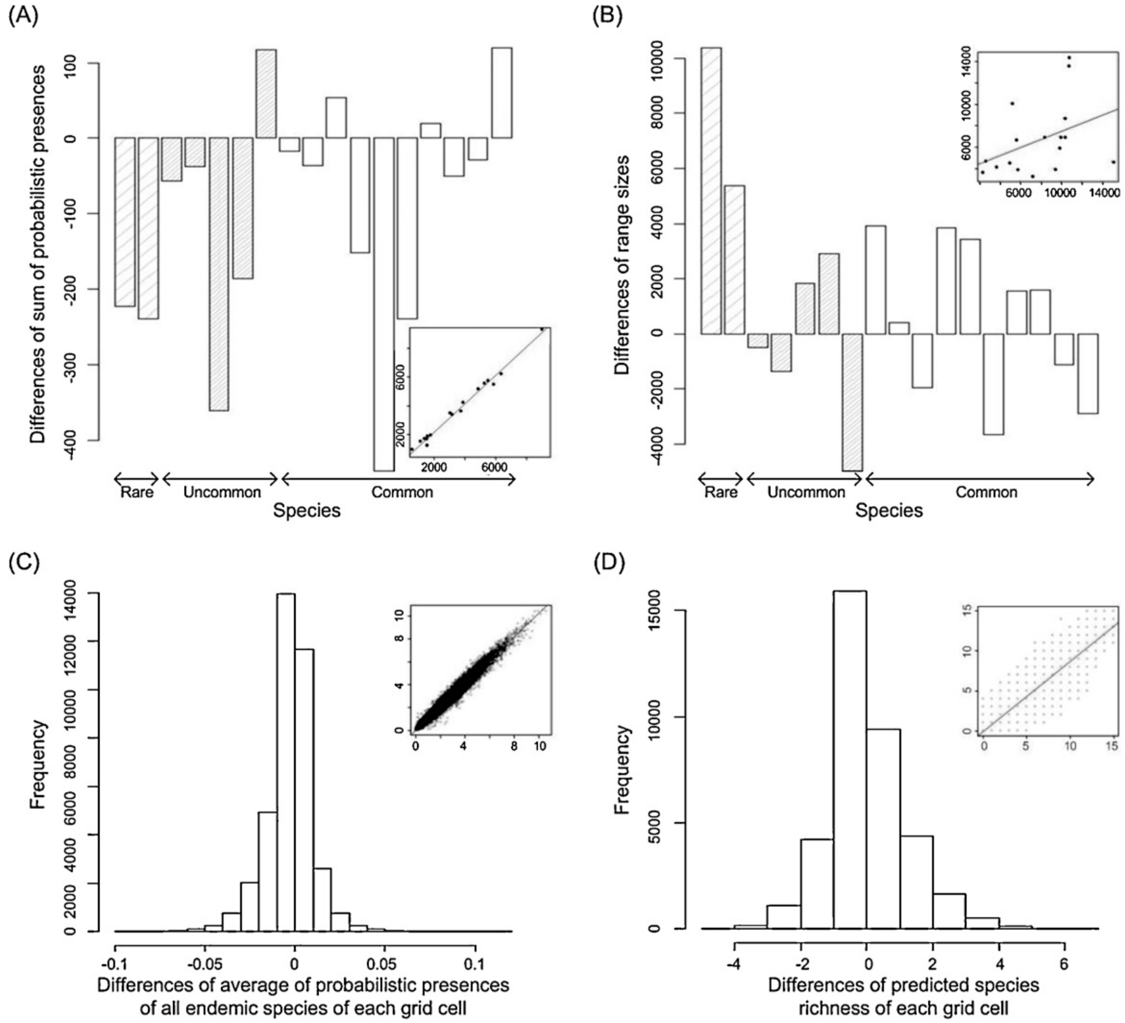

Fig. 3. Differences in (A) sum of probabilistic presences and (B) range sizes of individual species and frequencies of differences in (C) average of probabilistic presences of 17 endemic bird species per grid cell and (D) predicted species richness of each grid cell between the YY- and AYs-method. The differences were calculated as "YY-AYs". Insert figures show correlations between predicted results of the models trained by the YY- and AYs-method. Values of the *x*-axis were predicted by the YY-method and values of the *y*-axis were predicted by the AYs-method. Red lines represent the 1:1 relationship. Species from left to right in (A) and (B) can refer to Fig. 4 (top to bottom).

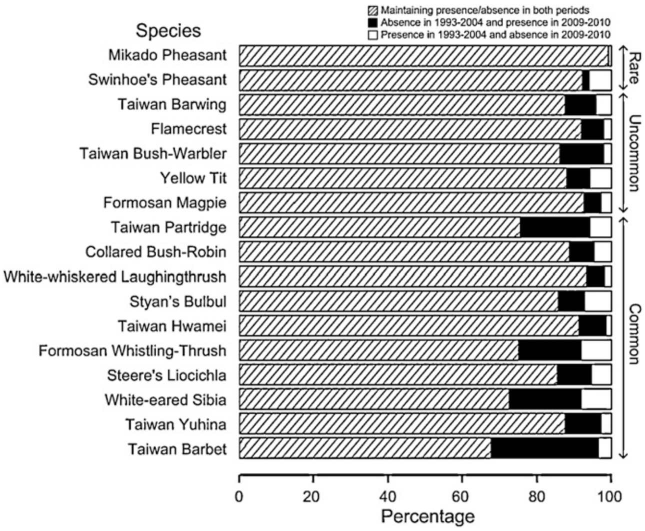

Fig. 4. Patterns of species presences and absences in 1993–2004 and 2009–2010 based on records from 280 grid cells that were investigated in both survey periods.

species distributional maps, including both probabilistic presence and presence-absence maps, predicted by the YY- and AYs-method had significant differences (paired *t* test, both $p < 0.001$). In most grid cells, the average probabilistic presences of all species and species richness showed higher values predicted by the AYs-method than by the YY-method (Fig. 3C and D). Differences of the average probabilistic presences of all species of each grid cell between the YY- and AYs-method were within ± 0.1 and differences of the predicted species richness of each grid cell were between −4 and 5 species. The predicted results showed high correlations between predictions by the YY- and AYs-method (Fig. 3C, 3D; $r^2 = 0.92$ and 0.88, average of probabilistic presences and predicted species richness, respectively).

When examining the 280 grid cells investigated in both survey periods, the species presences and absences were generally quite similar, with an average of 85.8% similarity (Fig. 4). But among the 17 species, 0.4–28.9% of the survey grid cells where species were not observed in 1993–2004 but observed in 2009–2010 and less than 10% of survey grid cells showed species presences in 1993–2004 but not in 2009–2010. Three species with the highest percent changes of grid cells from species absences in the early survey period to species presences in the late survey period were the Taiwan Barbet (28.9%), the White-eared Sibia (19.3%), and the Taiwan Partridge (18.9%), all of which were categorized as common species.

Discussion

Using temporally independent or dependent datasets in species distribution models as training or testing data influences the results of species predicted distributions, especially when estimating range sizes of a species. There was no statistical significance observed among model performance measures, but models yielded increasing variances when using different methods to train and test models. Each model performance measure did not show congruent patterns in comparison with all dataset combinations, and thus it can be misleading to interpret model accuracy while reporting any measure alone. Additionally, developing models based on different methods to select training datasets does result in different predictions according to both species- and grid cell-based assessments that may further lead to model uncertainty. The manner in which to use past and current species distribution data to develop species distribution models, to evaluate model performance, and to project future species distribution requires further exploration.

The problems caused by errors in the location of species presence records and bias in sampling effort, which often occur in different survey periods, have received a great deal of theoretical treatment (Newbold 2010). The methods we proposed in this study (*i.e.* the YY- and AYs-method) were to empirically test the effect of these on the accuracy of species distribution models. We paired *t* test found that when spatial, taxonomical, and sampling effort was consistent, there was no difference in model performance between training-data extraction methods (*e.g.* the YY- and AYs-method), but predictions in detail still existed. The weights given to environmental variables in the models using the annual and pooled training data caused differences in the sum of probabilistic presences of individual species and average of probabilistic presences of all endemic species of each grid cell. However, we expected that the range sizes of individual species and the predicted species richness in each grid cell should produce similar results regardless of the methods used. Nevertheless, the results were incongruent. We detected 0.5- to 3-fold differences in the range sizes of individual species between predictions trained by the YY- and AYs-method, which might provoke completely different conservation assessment and management. Given the considerable investment in time and money necessary to conduct surveys of species presences, it is important to ensure that the species records are not biased spatially, environmentally, temporally, and taxonomically with respect to the environmental variables used. For instance, false absences, which can occur when a species could not be detected although it was present, or when the species is not yet/no more present but the environment is in fact suitable (*e.g.* due to dispersal limitation or metapopulation dynamics), are seen as one of the main and common drivers of uncertainty in species distribution models and may seriously bias analyses (Barry & Elith 2006; Pearson 2007; Hanspach, Kühn, Schweiger, Pompe, & Klotz 2011). In addition, spatial autocorrelation may also violate the assumption of independently distributed errors in the models and affect evaluation of explanatory variables resulting in commission error being inflated (Legendre 1993; Diniz-Filho, Bini, & Hawkins 2003; Kissling, Field, & Böhning-Gaese 2008). However, we could not measure the spatial autocorrelation or intrinsic properties of explanatory variables in the study due to in the absence of systematical classification of different-source spatial data.

Similarly, the quantity and quality of the testing data used in the assessments of presence–absence models have major impacts on the interpretation of model results (Foody 2011). Although there were no significant differences between using the temporally independent (data in 2009–2010) and dependent (data in 1993–2004) testing data among the model performance measures, variances of most measures, including AUC, kappa, sensitivity, and TSS, among the 17 species were higher when estimating by the testing data selected from the 1993–2004 survey period than from the 2009–2010 survey period. The species, such as the Taiwan Barbet, the White-eared Sibia, and the Taiwan Partridge, with high competition ability and high tolerance to environmental changes are likely to increase their presences and may positively or negatively affect the evaluation of model performance.

In both ecology and clinical epidemiology, prevalence is an important factor that is inherent in assessments of model performance and predictive accuracy (Lantz & Nebenzahl 1996; McPherson, Jetz, & Rogers 2004; Jiménez-Valverde, Lobo, & Hortal 2009). We further estimated relationships between prevalence and the values of six model performance measures. In our study, the accuracy, sensitivity, specificity, and TSS were influenced less by prevalence (mostly $r^2 < 0.2$) in both linear and exponential trends that were similar to those demonstrated by McPherson et al. (2004) and Allouche et al. (2006). Although a high correlation was suggested by those studies when using the dependent datasets as the testing data, the correlation might be resulted from the original data itself instead of the prevalence. The kappa values were the most dependent on prevalence, with r^2 ranging from 0.23 to 0.31 among different dataset combinations. The AUC values indicated slightly different patterns in the two testing datasets. Regardless of the training datasets and the linear and exponential trends, the aforementioned r^2 values which were estimated by the temporally dependent datasets were higher than those estimated by the temporally independent datasets. These result indicated a high and low response to prevalence, respectively, and also different from known AUC, which should be independent of prevalence. Yet we have not discovered a good explanation for the findings. We also found slightly negative effects of prevalence on the AUC, accuracy, sensitivity, and specificity, which implied some common species (*i.e.* widespread species) might have lower values of predictive accuracy (*i.e.* greater overall errors) than uncommon species (*i.e.* restricted-range species). A possible explanation is that the local availability of environmental resources may be of overriding importance in limiting the distributions of some common species (Guisan & Hofer 2003; Segurado & Araujo 2004). The negative relationships between model performance and species prevalence require further tests at a local scale in Taiwan.

In all, our comparisons of data splitting in the species distribution models identified differences among model evaluation and predicted species distributions; to be successful, establishing long-term species sampling networks and monitoring species distributional changes are important for the usages of species distribution models and further proper interpretation of predictions.

Acknowledgements

We thank the anonymous reviewers and editors for comments that improved the manuscript. We are particularly grateful to the Endemic Species Research Institute, Taiwan and Spatial Ecology lab at the Institute of Ecology and Evolutionary Biology, National Taiwan University for providing and compiling years of bird data, and all of the volunteers and researchers for involving in the bird surveys. The study was supported by the Delta Electronics Foundation and Yu Chi-Chung Culture and Educational Foundation fellowships (to C.-Y. Ko) and National Science Council, Taiwan (to P.-F. Lee).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.baae.2013.04.003.

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232.

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, *200*. I–I9

Barry, S., & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, *43*, 413–423.

Buckland, S. T. (2006). Point transect surveys for songbirds: Robust methodologies. *The Auk*, *123*, 345–357.

Buckland, S. T., Anderson, D. R., Burnham, K. P., & Laake, J. L. (1993). *Distance sampling: Estimating abundance if biological populations*. London: Chapman and Hall.

Buckland, S. T., Goudie, I. B. J., & Borchers, D. L. (2000). Wildlife population assessment: Past developments and future directions. *Biometrics*, *56*, 1–12.

Clements, J.F., Schulenberg, T.S., Iliff, M.J., Sullivan, B.L., Wood, C.L., & Roberson, D. (2011) The Clements checklist of birds of the world: Version 6.6. Downloaded from http://www.birds.cornell.edu/clementschecklist/Clements%206.6.xls

Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Crawford, P. H. C., & Hoagland, B. W. (2010). Using species distribution models to guide conservation at the state level: the endangered American burying beetle (Nicrophorus americanus) in Oklahoma. *Journal of Insect Conservation*, *14*, 511–521.

Diniz-Filho, J. A. F., Bini, L. M., & Hawkins, B. A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, *12*, 53–64.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Jin, L., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J., Peterson, McC., & Phillips, A. T. S. J. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*, 330–342.

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

Foody, G. M. (2011). Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. *Global Ecology and Biogeography*, *20*, 498–508.

Gaston, K. J., & Fuller, R. A. (2009). The sizes of species' geographic ranges. *Journal of Applied Ecology*, *46*, 1–9.

Guisan, A., & Hofer, U. (2003). Predicting reptile distributions at the mesoscale: Relation to climate and topography. *Journal of Biogeography*, *30*, 1233–1243.

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

Hanspach, J., Kühn, I., Schweiger, O., Pompe, S., & Klotz, S. (2011). Geographical patterns in prediction errors of species distribution models. *Global Ecology and Biogeography*, *20*, 779–788.

Hijmans, R. J., & Graham, C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, *12*, 2272–2281.

Hsu, F. H., Yao, C. T., Lin, R. S., Yang, C. C., & Lai, S. J. (2004). Avian species composition and distribution along elevation gradient in the southern Taiwan. *Endemic Species Research*, *6*, 41–66.

IUCN. (2012). *The IUCN Red List of Threatened Species. Version 2012.1*. http://www.iucnredlist.org

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2009). The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, *10*, 196–205.

Kissling, W. D., Field, R., & Böhning-Gaese, K. (2008). Spatial patterns of woody plant and bird diversity: Functional relationships or environmental effects? *Global Ecology and Biogeography*, *17*, 327–339.

Ko, C. Y., Lin, R. S., Ding, T. S., Hsieh, C. H., & Lee, P. F. (2009). Identifying biodiversity hotspots by predictive models: A case study using Taiwan's endemic bird species. *Zoological Studies*, *48*, 418–431.

Ko, C. Y., Lin, R. S., & Lee, P. F. (2010). Macrohabitat characteristics and distribution hotspots of endemic bird species in Taiwan. *Taiwania*, *55*, 216–227.

Ko, C. Y., Root, T. L., & Lee, P. F. (2011). Movement distances enhance validity of predictive models. *Ecological Modelling*, *222*, 947–954.

Koh, C. N., Lee, P. F., & Lin, R. S. (2006). Bird species richness patterns of northern Taiwan: Primary productivity, human population density, and habitat heterogeneity. *Diversity and Distributions*, *12*, 546–554.

Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, *49*, 431–434.

Lee, P. F., Ko, C. J., Huang, K. H., Kao, W. H., Wu, T. Y., Lin, H. S., Chen, W. C., Lin, R. S., Fan, M. W., Hsieh, C. F., & Yu, W. T. (2010). *Taiwan breeding bird survey in 2009–2010*. Nantou, Taiwan: Endemic Species Research Institute., 11 pp.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*, 1659–1673.

Lei, F. M., Qu, Y. H., Lu, J. L., & Yin, Z. H. (2003). Conservation on diversity and distribution patterns of endemic birds in China. *Biodiversity and Conservation*, *12*, 239–254.

McPherson, J. M., Jetz, W., & Rogers, D. J. (2004). The effects of species' range size on the accuracy of environmental distribution models – ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, *41*, 811–823.

Meynard, C. N., & Quinn, J. F. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, *34*, 1455–1469.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, *34*, 3–22.

Nichols, J. D., Thomas, L., & Conn, P. B. (2008). Inferences about landbird abundance from count data: Recent advances and future directions. *Journal of Ecological and Environmental Statistics*, *3*, 201–236.

Pearson, R. G. (2007) *Species' distribution modeling for conservation educators and practitioners. Synthesis*. American Museum of Natural History. Available at: http://ncep.amnh.org

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

Phillips, S. J., & Dudik, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*, 161–175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*, 181–197.

Phillips, S. J., Dudik, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on machine learning* (pp. 655–662).

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Segurado, P., & Araujo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, *31*, 1555–1568.

Soberon, J. M., Llorente, J. B., & Onate, L. (2000). The use of specimen-label databases for conservation purposes: An example using Mexican Papilionid and Pierid butterflies. *Biodiversity and Conservation*, *9*, 1441–1466.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlbaum Press.

Su, H. J. (1992). Vegetation in Taiwan: Mountain vegetation and geographic and climatic zones. *Botanical bulletin of Academia Sinica*, *11*, 39–53.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*, 654–657.