SEXING BIRDS USING DISCRIMINANT FUNCTION ANALYSIS:
A CRITICAL APPRAISAL

FRANÇOIS-XAVIER DECHAUME-MONCHARMONT,1,3 KARINE MONCEAU,1
AND FRANK CEZILLY1,2

1Université de Bourgogne, Equipe Ecologie Évolutive, UMR CNRS 5561 Biogéosciences, Dijon, France; and
2Institut Universitaire de France

ABSTRACT.—Discriminant function analysis (DFA) based on morphological measurements is a quick, inexpensive, and efficient method for sex determination in field studies on cryptically monomorphic bird species. However, behind the apparent standardization and relative simplicity of DFA lie subtle differences and pitfalls that have been neglected in some studies. Most of these concerns directly affect assessment of the discriminant performance, a parameter of crucial importance in practice because it provides a measure of the quality of an equation that may be used in later field studies. Using results from 141 published studies and simulations based on a large data set collected on adult Zenaida Doves (Zenaida aurita), we assessed the effects of sexual dimorphism, sample size, and validation methods on discrimination rates. We compared the three most common methods used to estimate the proportion of correctly classified males and females by DFA: resubstitution, jackknife, or sample splitting. Results from simulations indicate that these procedures may lead to opposite conclusions, especially when the sample size is small. In particular, the resubstitution techniques appear to be over-optimistic, and we therefore recommend that DFA accuracy be estimated by the jackknife cross-validation procedure. In addition, we show that most previous studies failed to present DFA accuracy with 95% confidence intervals, which hampers comparisons among studies. Finally, our results suggest that large sample sizes should be preferred over repeated measurements of the same individuals, because random measurement error is likely to have only a weak effect on the accuracy of the discriminant rate. Received 17 May 2010, accepted 11 November 2010.

Key words: cross-validation, DFA, measurement errors, morphological measurements, sample size effect, sexual dimorphism, Zenaida aurita.

Identification du sexe des oiseaux par analyse factorielle discriminante: une évaluation critique

RÉSUMÉ.—L’analyse factorielle discriminante (AFD) à partir de mesures morphologiques constitue une méthode rapide, peu chère et efficace pour un sexage sur le terrain d’oiseaux appartenant à des espèces cryptiquement monomorphes. Cependant, l’apparente standardisation et la relative simplicité de l’AFD cachent en réalité plusieurs différences et écueils négligés dans de nombreuses études. La plupart de ces problèmes nuisent directement à l’évaluation du pouvoir de discrimination, un paramètre crucial en pratique puisqu’il correspond à une mesure de la qualité de l’équation destinée à être utilisée dans des études de terrain ultérieures. En nous appuyant sur 141 publications et des simulations basées sur un large jeu de données collectées sur des Tourterelles à queue carrée Zenaida aurita adultes, nous avons évalué l’effet du dimorphisme sexuel, de la taille de l’échantillon et de la méthode de validation sur le taux de discrimination. Nous avons comparé les trois méthodes les plus utilisées pour estimer la proportion de mâles et de femelles par AFD: resubstitution, jackknife, ou échantillonnage. Les résultats issus de simulations indiquent que ces procédures peuvent conduire à des conclusions opposées, surtout en cas de petites tailles d’échantillons. Plus précisément, les techniques de resubstitution semblent beaucoup trop optimistes. Nous recommandons que la précision d’une AFD soit évaluée au moyen d’un jackknife. De plus, nous montrons que la plupart des études antérieures indiquent cette précision sans l’assortir d’un intervalle de confiance à 95%, ce qui limite la possibilité de comparaisons entre études. Enfin, nos résultats suggèrent que de grandes tailles d’échantillons devraient être préférées à des mesures répétées sur les mêmes individus puisqu’il est probable que l’erreur de mesure n’ait qu’un effet limité sur la précision du taux de discrimination.

E-mail: fx.dechaume@u-bourgogne.fr

The Auk, Vol. 128, Number 1, pages 78–86. ISSN 0004-8038, electronic ISSN 1938-4254. © 2011 by The American Ornithologists’ Union. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press’s Rights and Permissions website, http://www.ucpressjournals.com/reprintInfo.asp. DOI: 10.1525/auk.2011.10129
Sex-related differences in the behavior and ecology of bird species are crucial to our understanding of sexual selection and mating systems (Andersson 1994) and may also have important consequences for management and conservation (Zavalaña and Paredes 1997, Fernandez-Juricic et al. 2009). Therefore, the ability to identify the sex of an individual is of paramount importance in avian studies. Although sexual dimorphism can be conspicuous (e.g., plumage-based sexual dichromatism), especially in the case of polygynous species (Darwin 1871, Andersson 1994), differences between male and female birds are often subtle or escape the human eye (Cuthill et al. 1999). Several techniques have been developed to alleviate this difficulty, including anatomical examination (Petrides 1950, Miller and Wagner 1955), vocalization analyses (Bourgeois et al. 2007), sex-specific behavior analyses (Castoro and Guhl 1958, Flux and Innes 2001, Fletcher and Hamer 2003), and, more recently, molecular techniques (Dubiec and Zagalska-Neubauer 2006). Although the latter methods are by far the most popular techniques and are thought to be very reliable (but see Robertson and Gemmell 2006), they require training and a license to collect blood or tissues, which raises ethical issues, and imply financial costs and delays for processing of samples.

Sexing based on morphometrics is often a reasonable choice for quick and inexpensive but efficient sex identification in field studies on cryptically monomorphic bird species. Significant, albeit small, differences in biometric measurements often exist between females and males, offering the possibility of discriminating between sexes (Murphy 2007, Cardoni et al. 2009). To that end, various statistical analyses have been used, such as linear models (Jeffrey et al. 1993, Iko et al. 2004, Ura et al. 2005, Gill and Vonhof 2006, Hallgrimsson et al. 2008) or multivariate methods, including principal component analysis (Rubega 1996, McCracken et al. 2000, Remisiewicz and Wennerberg 2006, Urﬁ and Kalam 2006, Schroeder et al. 2008) and discriminant function analysis (DFA). The latter is the most popular of these statistical methods, and its use has increased steadily in recent decades (Fig. 1). The principle of DFA is to provide equations based on morphological measurements in order to predict the sex of birds. This equation is primarily calibrated on individuals of known sex. Every discriminant equation comes with its estimated proportion of correctly sexed individuals.

Biologists justifiably pay great attention to the proportion of misclassified birds, because a discriminant equation is not published as an end in itself but as a tool for reliably and rapidly sexing birds in later field studies. Various methods can be used to estimate the proportion of correctly classified males and females by DFA, and yet the choice of a method is rarely justified in the avian literature. In addition, the estimated rate of correct discrimination has been reported to be sensitive to sample size (Morrison 1984). Among or within subjects, variance in measurements may also have an effect on the ability to discriminate males from females. Such variance has two components: the true variance among individuals and the measurement error (ME) that arises from random or systematic errors. Large MEs increase the risk of Type II error (Loughheed et al. 1991, Yezerinac et al. 1992). Francis and Mattlin (1986) showed that discriminant power could fall from 89% to <50% when a small amount of bias exists in morphometrics measurement. The need for ME assessment has been emphasized (Bailey and Byrnes 1990, Arnqvist and Mårtensson 1998), yet few studies that used DFA have reported ME (Flux and Innes 2001, Devlin et al. 2004, Kenward et al. 2004), and its potential effect on discrimination rate has rarely been investigated (but see Mallory and Forbes 2005).

Here, we address the problem using a twofold approach. First, using simulations based on a large data set, we assess the choice of validation method and the effect of sample size or measurement errors on the estimated accuracy of the DFA. Second, we review the literature to quantify the methodologies used, assess the effects of the extent of sexual dimorphism, sample size, or number of variables on discriminant rate, and test predictions from our simulations. Finally, we provide recommendations for future studies.

**Methods**

*Simulations.*—A total of 525 adult Zenaida Doves (Zenaida aurita; 294 females and 231 males) were captured on Barbados from March to May 2007 (Monceau et al. 2011). This sexually monomorphic species is widely distributed throughout the Caribbean. Each individual was molecularly sexed (Fridolfsson and Ellegren 1999). The reliability of molecular sexing techniques has been reported to vary according to which PCR amplification procedure is used (Dubiec and Zagalska-Neubauer 2006, Robertson and Gemmel 2006, Daniel et al. 2007). In a preliminary study (Monceau 2009), we first assessed the reliability of our procedure by comparing results from molecular sexing with behavioral observations based on 48 pairs.

For each bird, we measured three dimensions of the bill at nostrils (length, depth, and width), head plus bill length, left and right tarsus length, left and right wing chord, and tail length. All measurements were made by the same person (K.M.) with a digital caliper (precision: ±0.2 mm), except for wing chord and tail length, which were measured with a ruler (precision: ±1 mm). Each character was measured twice, and the caliper was removed between measurements in order to assess MEs (Bailey and Byrnes 1990).
Because MEs were low for all traits except bill depth (Table 1), we used the mean of first and second measurements of each variable in subsequent analyses. Birds were weighed with a Pesola digital pocket scale MS500 (precision: ±0.1 g). Males differed significantly from females in all morphological traits, and the mean difference index (MDI = 100 × mean female/mean male; Delestrade 2001, Helfenstein et al. 2004) was >95% for all characters (Table 1). We performed DFA on the original data set using the eight morphological variables. Because of a violation of the assumption of homogeneity of the variance–covariance matrices (box’s M test, $\chi^2 = 267.1, df = 190, P = 0.0002$), we performed quadratic DFA (Stevens 1992) using the qda function from the MASS package (Venables and Ripley 2002) for R, version 2.10.1 (R Development Core Team 2010). We searched for potential multivariate outliers by measuring robust Mahalanobis distances (Rasmussen 1988, Jarrell 1994, Team 2010). We searched for potential multivariate outliers by measuring robust Mahalanobis distances (Rasmussen 1988, Jarrell 1994, Team 2010). We assessed the leverage of outliers by removing them from the data set and estimating the new discriminant rate. The effect of automated variable selection was assessed using the stepclass function and minimization of Wilks’s lambda criterion (Mardia et al. 1979) as implemented in the klaR package for R (Wehns et al. 2005).

We compared three validation methods commonly used in sexing birds to estimate the proportion of correctly classified individuals by a DFA: resubstitution, jackknife (Manly 1994), and sample splitting (Picard and Berk 1990). In the first method, the sex of each individual is predicted using the discriminant function calculated from the complete data set. Using the jackknife (or leave-one-out) method, the sex of an individual is predicted from the discriminant equation calculated after that individual has been removed from the data set. This procedure is repeated until a sex is assigned to each individual (Tabachnick and Fidell 2000). With sample splitting, the data set is randomly divided into two subsamples, and the training set (two-thirds of the individuals, as recommended by Picard and Berk 1990) is used to compute the discriminant function that is then used to assess the accuracy of the discriminant function by classifying the remaining third of the individuals.

We simulated both smaller data sets and data sets with larger MEs in order to assess the effect of sample size, the choice of validation method, and MEs on the discriminant power. We defined 100 different sample sizes ranging from 25 to 520 individuals, regularly spaced every five individuals. For each sample size, we simulated 500 different data sets by randomly sampling individuals from the complete data set. For each of these 50,000 simulated data sets, we performed DFA and assessed the discriminant rates using resubstitution, jackknife, and sample-splitting methods.

We performed a second series of simulations using the complete original data set ($n = 525$ individuals) to assess the effect of ME. Each variable used in the DFA was the mean of two repeated measures recorded in order to assess ME (Table 1). We simulated higher ME for a given variable by adding random noise to both repeated measures. This additional noise followed a normal distribution with mean $\mu = 0$ and standard deviation $\sigma$. We adjusted $\sigma$ up to obtain the chosen ME ($\pm0.01$). Note that it was not possible to artificially decrease ME, because we could not improve the accuracy of the data. Variables with ME already higher than the ones requested for the simulation were not modified. For example, for a minimal ME of 10%, we increased ME for every variable (including mass), except for bill depth, which already showed a larger ME (Table 1). For each value of minimal ME, we simulated 500 data sets and performed a DFA on each data set. For each DFA, the proportion of correctly sexed individuals was estimated using the jackknife procedure.

**Literature survey.**—We searched the literature and found 141 studies (from 1961 to 2010) on 132 species from 43 families (online supplement; see Acknowledgments) that used DFA to identify an individual’s sex or describe sexual dimorphism in birds. For a given group (same population and same age class), some authors reported more than one discriminant rate using several discriminant equations or several validation methods or both. In order to avoid giving too much weight to studies that reported multiple DFAs, we retained only the equation that led to the best discrimination between sexes. If the resubstitution method or sample-splitting was used along with the jackknife, we report only the discriminant rate estimated for the jackknife validation procedure. We collected a total of 186 DFAs on different populations, age classes, and species. For each of these analyses, we recorded the sample size used to calculate the discriminant function, the validation method, and the discriminant rate. Some authors used a stepwise procedure to reduce the number of variables used in the discriminant equation. We reported the stepwise procedure used (if available) and the number of variables retained in the discriminant equation. Finally, the mean tarsus length of males and females was also collected, if available, in order to estimate the relative sexual dimorphism (Storer’s index: absolute value of the difference between male and female divided by the mean tarsus length divided by the mean female tarsus length; $\Delta$).

**Table 1.** Comparisons between male and female Zenaida Doves on Barbados for all characters with measurement error (ME) and mean difference index (MDI) calculated as (mean female/mean male) × 100.

<table>
<thead>
<tr>
<th>Characters</th>
<th>Males (mean ± SD)</th>
<th>Females (mean ± SD)</th>
<th>Cohen’s d (95% CI)</th>
<th>t-test</th>
<th>P</th>
<th>ME (%)</th>
<th>MDI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill length (mm)</td>
<td>10.67 ± 0.48</td>
<td>10.45 ± 0.51</td>
<td>0.45 (0.28–0.62)</td>
<td>−5.11</td>
<td>&lt;10−5</td>
<td>7.92</td>
<td>97.94</td>
</tr>
<tr>
<td>Bill width (mm)</td>
<td>4.07 ± 0.22</td>
<td>3.93 ± 0.26</td>
<td>0.58 (0.41–0.77)</td>
<td>−6.73</td>
<td>&lt;10−5</td>
<td>9.63</td>
<td>96.56</td>
</tr>
<tr>
<td>Bill depth (mm)</td>
<td>4.15 ± 0.21</td>
<td>3.98 ± 0.21</td>
<td>0.78 (0.61–0.96)</td>
<td>−8.83</td>
<td>&lt;10−5</td>
<td>21.86</td>
<td>95.90</td>
</tr>
<tr>
<td>Head plus bill length (mm)</td>
<td>48.59 ± 0.99</td>
<td>47.40 ± 1.02</td>
<td>1.18 (1.01–1.36)</td>
<td>−13.47</td>
<td>&lt;10−5</td>
<td>3.65</td>
<td>97.55</td>
</tr>
<tr>
<td>Mean tarsus length (mm)</td>
<td>26.76 ± 0.86</td>
<td>25.94 ± 0.77</td>
<td>1.01 (0.92–1.21)</td>
<td>−11.30</td>
<td>&lt;10−5</td>
<td>6.83</td>
<td>96.94</td>
</tr>
<tr>
<td>Mean wing chord (cm)</td>
<td>15.31 ± 0.54</td>
<td>14.74 ± 0.46</td>
<td>1.61 (0.98–1.36)</td>
<td>−12.93</td>
<td>&lt;10−5</td>
<td>2.22</td>
<td>96.28</td>
</tr>
<tr>
<td>Tail length (cm)</td>
<td>10.38 ± 0.36</td>
<td>9.94 ± 0.56</td>
<td>0.79 (0.60–0.97)</td>
<td>−8.94</td>
<td>&lt;10−5</td>
<td>4.07</td>
<td>95.76</td>
</tr>
<tr>
<td>Body mass (g)</td>
<td>149.19 ± 15.21</td>
<td>141.81 ± 13.95</td>
<td>0.51 (0.33–0.70)</td>
<td>−5.73</td>
<td>&lt;10−5</td>
<td>—</td>
<td>95.05</td>
</tr>
</tbody>
</table>
length; Smith 1999). We assessed the effect of relative sexual dimorphism on the number of birds sampled using analysis of variance and on the number of variables selected in the discriminant equation using generalized linear models (Poisson regression). We used linear models to assess the effect on the discriminant rate of sample size, relative sexual dimorphism, validation method, number of variables in the discriminant equation, variable-selection methods (automated or not), and year of publication. Percentages were inspected for normality and normalized using square-root arcsine transformation if needed (Zar 1999).

**RESULTS**

*Effect of sample size and validation methods.*—The differences between the three validation methods were far from negligible, especially for small sample sizes (Fig. 2). The proportions of correctly sexed birds (discriminant rate) were larger when estimated with the resubstitution method, and decreased with increasing sample size (Fig. 2A). For small data sets (n < 60 birds), the mean estimate of the discriminant rate was >90%. By contrast, the mean estimates with the jackknife (Fig. 2B) or sample-splitting (Fig. 2C) methods increased with increasing sample size and were consistent, although the variance of the estimates was smaller with the jackknife than with the sample-splitting method.

Differences between methods were less pronounced with large data sets. Considering the DFA based on the complete original data set (n = 525 adults), the resubstitution method estimated a proportion of correctly sexed adults of 81.0% (95% confidence interval [CI]: 77.5–84.8%), whereas the same value was 80.4% (95% CI: 76.8–84.4%), and 79.1% (95% CI: 75.6–82.6%) for the jackknife and the sample-splitting methods, respectively. Because their bootstrapped 95% CI largely overlapped, the estimated proportion of correctly classified birds did not significantly differ according to the validation method. By measuring robust Mahalanobis distances, we identified 39 potential multivariate outliers (19 males and 20 females). However, these outliers had a limited effect in our analyses, because removing these extreme individuals from the original data set only slightly increased the proportion of correctly sexed birds estimated with a given validation method: resubstitution, 84.0% (95% CI: 80.0–87.7%); jackknife, 82.0% (95% CI: 77.7–86.0%); and sample splitting, 80.1% (95% CI: 73.7–85.7%).

The stepclass function for automated variable selection suggested several subsets of variables. The most frequent set was made up of three variables: wing chord, head length, and tarsus length. Stepwise variable selection based on minimization of Wilks's lambda criterion, a frequently used procedure in the bird literature, led to much larger variable sets, with up to eight variables being included in the complete sample (n = 525). We used the reduced set of three variables obtained from the stepwise procedure (wing chord, head length, and tarsus length) and performed the same simulations detailed above. The shape of the relationship between mean discriminant rate and sample size was unchanged, though it was slightly less steep for small sample sizes. In addition, asymptotic discriminant rates calculated for the complete sample size (n = 525) were slightly smaller, with 79.8% (95% CI: 74.8–82.5%), 78.9% (95% CI: 74.2–81.9%), and 78.1% (95% CI: 74.9–86.1%) of correctly sexed birds for resubstitution, jackknife, and sample-splitting, respectively.

*Effect of measurement errors.*—An increase of ME led to a decrease in the proportion of correctly sexed individuals (Fig. 3). Despite a large drop of accuracy in the simulated measurements, the mean proportion of correctly classified individuals only decreased from 80.38% to 78.09%, which corresponds to ~12 additional adults incorrectly classified. Using the smaller subset of variables (wing chord, head length, and tarsus length) selected by automated stepwise procedure, we found similar results. The mean discriminant rate dropped marginally from 78.9% to 77.7% with a simulated ME of 30%.

*Literature survey.*—Sample size varied greatly, from 10 birds (4 males and 6 females) to 1,891 (918 males and 973 females). The median sample size was 80 birds (95% CI: 72.5–94.0). The median sex ratio (male/female) was 1.04 (95% CI: 0.97–1.09). Discriminant
Fig. 3. Boxplot of the jackknifed estimated discriminant rate (proportion of correctly classified individuals) as a function of the minimal measurement error (ME). ME was increased to a given value by adjusting random noise to the repeated measures except when real ME was already larger. For each value of minimal ME, 500 data sets were simulated from the complete data set. Every discriminant equation relied on the eight variables listed in Table 1. For a given minimal ME, the black circle depicts the mean discriminant rate, the thick line the median, and the box the interquartile range. The horizontal dashed line represents the reference discriminant rate estimated from the original data, with no noise added.

Fig. 4. Results of a literature survey on the effect of relative sexual dimorphism (absolute difference of tarsus length between males and females divided by mean tarsus length) on the discriminant rate (proportion of correctly sexed birds). The line and the gray area depict the predicted trend and its 95% confidence interval fitted by linear regression ($F = 64.7$, df = 1 and 132, $P < 10^{-5}$) after square-root arcsine transformation for normalization of the percentages.

Relative sexual dimorphism had the largest effect on the discriminant rate ($F = 55.69$, df = 1 and 122, $P < 10^{-5}$) and explained 30.4% of the variance. Neither the validation method ($F = 0.25$, df = 2 and 127, $P = 0.78$), the year of publication ($F = 2.81$, df = 1 and 128, $P = 0.096$), nor the variable selection procedure (automated or not) ($F = 1.71$, df = 1 and 128, $P = 0.19$) had an effect on the discriminant rate. However, the interaction between sample size and the number of variables was significant ($F = 9.16$, df = 1 and 129, $P = 0.003$), with the discriminant rate decreasing with an increasing number of variables for small sample sizes. No other interaction had a significant effect on the discriminant rate (all $P > 0.18$). The discriminant rate significantly increased with the relative sexual dimorphism (Fig. 4). One might expect that the authors adjusted their measurement effort (sample size or number of morphological variables) accordingly. However, there was no significant association between relative sexual dimorphism and number of birds sampled ($F = 1.24$, df = 1 and 132, $P = 0.27$) or the number of variables used in the discriminant equation (Poisson regression: $\chi^2 = 0.27$, df = 1, $P = 0.16$).

Contrary to our simulations, validation methods had no significant effect on the discriminant rate in published studies, and lower sample sizes had an effect only in interaction with the number of variables. This might be explained by publication bias. We therefore examined the funnel plots of discriminant rate as a function of the log-transformed sample size (Palmer 2000). Because of sampling error, the variance in estimates of the discriminant rate is expected to be higher for studies with smaller sample sizes (Møller and Jennions 2001). For each validation method, there was no significant difference between the variance of discriminant rates of the smallest 50% of sample sizes and the largest 50% (Levene test; resubstitution: $P = 0.34$, jackknife: $P = 0.69$, sample-splitting: $P = 0.45$), which reveals a lack of variance for small sample sizes. Because there was no significant relationship between sexual dimorphism and sample size, the lack of variance for small sample sizes cannot be attributed to studies on strongly dimorphic species (which require fewer individuals to generate a high discriminant rate) and is possibly then, attributable to a publication bias. Small data sets that led to small discriminant rates were probably rejected by editors or underreported by the authors.
**DISCUSSION**

*Effect of the validation method.*—The simulations based on our Ze-naida Dove data set indicate that the choice of validation method may have a strong effect on the estimated discriminant rate. This effect was particularly sensitive for small to intermediate sample sizes (<200 individuals). Estimates of the discriminant rate based on subsampling techniques (such as jackknife cross-validation or splitting of the data into training and test sets) should be preferred over estimates based on resubstitution, which appeared to be overly optimistic for small sample sizes. It has been well illustrated, in both the statistical (Johnson and Wichern 1992, Huberty 1994, Manly 1994, Piraux and Palm 2001, Wehberg and Schumacher 2004) and the applied literature (Eisenbeis 1977, Lance et al. 2000), that reclassification of the original individuals used in constructing the discriminant equation leads to a biased estimated discriminant rate. In one of the first papers to use DFA to sex birds, Ryder (1978) explicitly raised the problem of resubstitution and recommended that it not be used, especially in the case of a comparison among studies. But as the technique became more popular within the ornithologist community, this advice was somehow overlooked. The two subsampling techniques are not equivalent, however. Sample splitting led to a mean estimate of the proportion of correctly classified individuals consistent with the mean estimate using the jackknife procedure, but with a much larger variance. Therefore, jackknife cross-validation should always be preferred over the sample-splitting procedure or resubstitution.

Surprisingly, we found no difference in the discriminant power between these three validation methods in our survey. However, most published discriminant rates were quite high (median over the sample-splitting procedure or resubstitution. Therefore, jackknife cross-validation should always be preferred using the jackknife procedure, but with a much larger variance. Sample splitting led to a mean estimate of the proportion of correctly classified individuals consistent with the mean estimate using the jackknife procedure, but with a much larger variance. Therefore, jackknife cross-validation should always be preferred over the sample-splitting procedure or resubstitution.

**Importance of large sample size.**—Our simulations with small to medium sample sizes (<200 birds) highlighted the wide variance in the estimated proportion of misclassified birds. High discriminant success can be obtained by chance, raising reasonable doubts about the subsequent use of equations constructed from small data sets. In accordance with previous studies (Brennan et al. 1991, Wehberg and Schumacher 2004, Shealer and Cleary 2007, Isaksson et al. 2008), we recommend caution in dealing with discriminant equations computed from small data sets. In addition, the sample size in interaction with the number of variables in the equation had a significant effect on the discriminant rates from our literature survey. If there are too few individuals in relation to the number of variables, the analysis leads to a poor discriminant rate (Marks and Dunn 1974, Burnham and Anderson 2002). This result is known as the “curse of dimensionality” (Bellman 1957). Several authors have advised carefully adjusting the number of birds measured to the number of morphological variables in order to have a sample size ≥3× larger than the number of variables used in the DFA (Williams and Titus 1988, McGarigal et al. 2000), but this criterion is sometimes claimed to be arbitrary (James and McCulloch 1990).

**Variable selection.**—Estimating the optimal number of variables to measure is of obvious importance. Morphological measurements are usually highly correlated, which can lead to unstable parameter estimates. Eliminating multicollinearity among the variables could improve the discriminant rate, although the procedure is highly debated (James and McCulloch 1990). A common procedure is to reduce each subset of highly correlated variables to only one variable. More importantly, morphometric measures are costly in time and can be stressful for the birds. Field biologists have to limit themselves to a narrow set of morphological variables. Two-thirds of the authors relied on automated stepwise techniques in order to identify, from a larger number of potentially relevant measures, a reasonable set of variables. Biologists should be extremely cautious with the output of automated analysis. Several authors strongly recommend that stepwise procedure be avoided entirely (James and McCulloch 1990, Snyder et al. 1991, Thompson 2001). One peer-reviewed journal even made it an editorial policy to summarily reject any article that used the technique (Babyk 2004). Stepwise methods capitalize on sampling error and yield results that are not replicable (Thompson 1995); there is no reason to think that they lead to the best equation for sexing the model species, and no guarantee that the results will be useful for later studies (James and McCulloch 1990, Derksen and Keselman 1992). Most importantly, the greatest danger of the use of stepwise analysis is the temptation to leap directly from routine procedure to straightforward conclusions about ecological relevance. These procedures are unable to select from a set of variables those that are most influential (Burnham and Anderson 2002); therefore, the selected variables are not necessarily more biologically relevant than nonselected ones.

We offer the reminder to authors that stepwise procedures are exploratory and only suggest combinations of variables that do an adequate job, not necessarily the most discriminating or the simplest. Our survey of the literature showed that studies that relied...
on automated stepwise procedures did not differ from the others in the number of variables used in the discriminant equation or on their discriminant power. Scientific judgment and choice of biologically meaningful combination of variables must play a leading role in selecting variable (James and McCulloch 1990, Burnham and Anderson 2002). If stepwise procedures are used, Manly (1994) advised that the analyses be rerun several times with randomized subsamples of individuals to check the validity of the results. Measurement errors.—We also found that the effect of ME on discriminant rates is generally moderate but nonetheless significant. Simulating large ME in our measures only marginally decreased the estimated proportion of correct classification. This result contrasts sharply with that obtained by Francis and Matlin (1986), who reported high sensitivity of DFA to small amounts of bias in measurement. They simulated systematic bias by adding a constant to each of the dimension of their data set, which was later analyzed with a discriminant function constructed from the unmanipulated data set. We, on the other hand, investigated the effect of random bias. As long as these MEs equally affected males and females in the same proportion and with no systematic bias, they had a limited effect on the probability of misclassification. This suggests that increasing the number of sampled birds might be preferable to additional measurements of the same bird. Yet, because measurement errors were rarely reported in the papers we reviewed, we could not compare our results from the simulation with actual field data. More studies are needed before firm conclusions can be made regarding the importance of ME for the accuracy of DFA.

Finally, several authors warned that multivariate outliers can lead to violation of multinormality assumption and can severely impair the accuracy of DFA (Manly 1994, Tabachnick and Fidell 2000, Osborne and Overbay 2004). Rather extreme observations can inflate classification error rates and bias parameter estimation. However, search methods for outliers were almost never reported in the ornithological literature that we reviewed (but see Blakesley et al. 1990, Santiago-Alarcon and Parker 2007). We advise that future studies that use DFA to sex birds systematically test for the presence of outliers and, if they are present, decide, on the basis of established criteria (discussed at length in Tabachnick and Fidell 2000), whether or not to remove them. Our purpose was not to criticize the use of DFA, which remains in our opinion a very efficient and straightforward technique. Molecular methods are not foolproof, but they are unquestionably more reliable and popular than DFA. If a person does not have access to a molecular laboratory, proper performance of DFA should follow the guidelines that we have recommended. Most of the common pitfalls addressed here can easily be avoided. Authors should be extremely cautious with the predictions of discriminant equations constructed from small data sets, even when the reported discriminant rate is high. The accuracy of DFA should be estimated with the jackknife cross-validation procedure, and this accuracy should always be presented with 95% CIs to allow interstudy comparisons.

Acknowledgments

Supplementary material (Table S1) is available at dx.doi.org/10.1525/auk.2011.10129. We thank S. Devonish of the Natural Heritage Department for providing us with a permit to capture and ring Zenaida Doves in Barbados. We thank J. Moreau, S. Motreuil, and G. Prato for help in catching birds and W. Worrell for hospitality. K.M. was supported by a doctoral grant from the Ministère de la Recherche et de l’Enseignement Supérieur. Financial support was partly provided by a research grant from the Agence Nationale pour la Recherche (program Monogamix ANR-08-BLAN-0214-02). Calculations were performed using HPC resources from DSI-CCUB (Université de Bourgogne).

Literature Cited


Associate Editor: M. E. Hauber