2017

Beyond Zar: The Use and Abuse of Classification Statistics for Otolith Chemistry

C. M. Jones
Old Dominion University

M. Palmers
Old Dominion University

J. J. Schaffler
Old Dominion University

Original Publication Citation

Jones, C. M., Palmer, M., & Schaffler, J. J. (2016). Beyond Zar: The use and abuse of classification statistics for otolith chemistry. *Journal of Fish Biology, 90*(2), 492-504. doi:10.1111/jfb.13051

Journal of FISH BIOLOGY

fsbi An International Society for Fish Biology

Journal of Fish Biology (2017) **90**, 492–504
doi:10.1111/jfb.13051, available online at wileyonlinelibrary.com

Beyond Zar: the use and abuse of classification statistics for otolith chemistry

C. M. Jones* † ‡, M. Palmer§† and J. J. Schaffler‖†

**Center for Quantitative Fisheries Ecology, Old Dominion University, Norfolk, VA 23529, U.S.A., §Mediterranean Institute for Advanced Studies (IMEDEA), Ecology and Marine Resources, C/Miquel Marques, 21-07190 Esporales, Illes Balears, Spainand ‖Muckleshoot Indian Tribe, 39015 172nd Ave SE, Auburn, WA 98092, U.S.A.*

Classification method performance was evaluated using otolith chemistry of juvenile Atlantic menhaden *Brevoortia tyrannus* when assumptions of data normality were met and were violated. Four methods were tested [linear discriminant function analysis (LDFA), quadratic discriminant function analysis (QDFA), random forest (RF) and artificial neural networks (ANN)] using computer simulation to determine their performance when variable-group means ranged from small to large and their performance under conditions of typical skewness to double the amount of skewness typically observed. Using the kappa index, the parametric methods performed best after applying appropriate data transformation, gaining 2% better performance with LDFA performing slightly better than QDFA. RF performed as well as QDFA and showed no difference in performance between raw and transformed data while the performance of ANN was the poorest and worse with raw data. All methods performed well when group differences were large, but parametric methods outperformed machine-learning methods. When data were skewed the performance of all methods declined and worsened with greater skewness, but RF performed consistently as well or better than the other methods in the presence of skewness. The parametric methods were found to be more powerful when assumptions of normality can be met and can be used confidently when skewness and kurtosis are minimized. When these assumptions cannot be minimized, then machine-algorithm methods should also be tried.

Key words: ANN; chemistry; classification; LDFA; otolith; QDFA; RF.

INTRODUCTION

Connectivity at the subpopulation level through fish movements across a heterogeneous landscape is assumed to improve population viability, metapopulation persistence and resilience to disturbance (Jones, 2006; Gaines *et al.*, 2007). Accordingly, a number of methodological approaches for assessing connectivity have flourished in recent years. Otolith chemistry is one of the most widely used methods. Provided that water chemistry is reflected in the chemistry of the otolith layer being deposited and remains unchanged with fish growth (Campana, 1999), the rationale for estimating adult natal origin is to compare the microchemistry of the inner part of the adult otolith with those of juveniles from all putative sources (Elsdon *et al.*, 2008; Anstead *et al.*, 2015). This,

‡Author to whom correspondence should be addressed. Tel.: +1 757 683 4497; email: cjones@odu.edu
†These authors made an equal contribution to this work.

however, is often a challenging task (Catalan *et al.*, 2014; Morales-Nin *et al.*, 2014) and, among other technical problems, scientists face the dilemma of which classification methods for assignment to choose. Many procedures have been used in the published literature [discriminant function analyses, neural networks and random forest (RF), among others], with each method attesting to its putative superiority (Cappo *et al.*, 2005; Mercier *et al.*, 2011). Although there is guidance for the use of these procedures in the statistical literature, much of this literature is technically complex and not easily accessible to non-statisticians. Moreover, there is a tendency for new approaches to gain more attention upon publication and for their use to spread subsequently even when other traditional methods will yield better results (Hastie *et al.*, 2001). When applied to appropriate data, these newer approaches can lead to improved insights (Hand, 1981; Olden *et al.*, 2008; Armitage & Ober, 2010). When applied inappropriately, however, they can result in a loss of statistical power and the ability to correctly classify. The use of statistics to quantify group memberships has a long history in the statistics literature (Fisher, 1936; Hand, 1981; Choi, 1986). Much of this historic work relies on assumptions about parametric distributions and likelihood, such as in linear discriminant function analysis (LDFA) and quadratic discriminant function analysis (QDFA). Although these methods were among the first applied to classify data, the field of statistical analysis has advanced quickly with the availability of personal computers and programming languages. Whereas traditional parametric methods relied on analytic solutions and assumptions about normality, increased computing power has led to numeric solutions for complex nonlinear functions and to the advancement of machine algorithms that use Monte-Carlo approaches, *e.g.* artificial neural networks (ANN) and RF (Recknagel, 2001; Olden *et al.*, 2008; Suryanarayana *et al.*, 2008). In basic terms, all of these classification methods provide the ability to ascertain differences between g groups and then to allocate new observations to each group, typically by dividing data into training and testing sets (Venables & Ripley, 2002).

The fundamental differences between the traditional (LDFA and QDFA) and machine-algorithm (ANN and RF) methods lie in distributional assumptions and in the derivation of the classification function itself. When Fisher (1936) developed LDFA, he used analytic methods to derive a linear combination of variables whose class means were maximally separated relative to the within class variance (Venables & Ripley, 2002). Traditional classification methods, such as LDFA and QDFA, derive their classification functions from the estimates of the variance–covariance matrix (Σ) and the group means (μ_i) (Hand, 1981). In LDFA, data should be multivariate normal with equal within-group variance-covariance, while in QDFA data should be multivariate normal but covariances do not have to be equal (Legendre & Legendre, 2012). In contrast to the traditional methods, ANN and RF use recursive algorithms to obtain parameter estimates (Hand, 1981; Shiffman, 2012). At its simplest, ANN determine a weighted sum of variables that are compared with group identity threshold values (T), $w_k x_k > T$ (Hand, 1981), where weights are determined recursively from the training set. Similarly, for tree classifiers such as RF, the training set is used to designate group variables with the most clear-cut differences to topmost nodes and then to undertake variable weightings into ongoing nodal partitions.

The choice of which classification method to use depends of the characteristics of the available data. In otolith chemistry studies, these data are often non-normally distributed and can be highly skewed (Thorrold *et al.*, 2001; Dorval *et al.*, 2005; Ashford *et al.*, 2012; Anstead *et al.*, 2015). In the literature, three basic approaches have been

taken. Traditional classification approaches (LDFA and QDFA) are applied to raw data without addressing the underlying assumptions or applied to transformed data that approximate the underlying assumptions (Schaffler *et al.*, 2009, 2014). The third approach is to apply machine-learning approaches to raw (or sometimes transformed) data (Thorrold *et al.*, 2001; Mercier *et al.*, 2011).

This article is written to help clarify the assumptions and limitations of methods used with otolith chemistry to classify fishes to habitats and to measure connectivity. Four methods (LDFA, QDFA, ANN and RF) were used to analyse a typical data set of otolith chemistry for the Atlantic menhaden *Brevoortia tyrannus* (Latrobe 1802) that was originally used to determine the habitat use of regions of Chesapeake Bay by juveniles (Schaffler *et al.*, 2014). Classification method performance was evaluated when assumptions were met and were violated. Moreover, these four methods were also tested with simulated data where differences between group means for each variable were small and large. Finally, simulated data were used to test for the performance of methods under conditions of small and large amounts of skewness. Altogether these tests reveal important issues of classification performance.

MATERIALS AND METHODS

COLLECTIONS AND OTOLITH ANALYSES

Juveniles of *B. tyrannus* ($n = 84$) were collected in three major nursery areas, upper mid and lower Chesapeake Bay (Fig. 1) during 2005 and 2006 as described in Schaffler *et al.* (2014). Sagittal otoliths were removed and processed using clean techniques. One of each pair was randomly selected and prepared for solution-based inductively-coupled plasma mass spectroscopy (ICP-MS) analysis of minor and trace elements. The solution was analysed on a thermo Finnegan Element 2 magnetic sector ICP-MS (Thermo Scientific; www.thermoscientific.com) at Woods Hole Oceanographic Institution. The remaining otolith of each pair was crushed to a homogeneous powder and prepared for stable-isotope analysis. The powder was analysed with an automated Isoprime micromass carbonate analyser (Isoprime; www.isoprime.co.uk) at the University of Washington Stable Isotope Laboratory.

OTOLITH CHEMISTRY VARIABLES

A suite of minor and trace elements and stable isotopes were measured and MANOVA was used to determine that six variables (Mg:Ca, Mn:Ca, Ba:Ca, Sr:Ca, $\delta^{13}C$ and $\delta^{18}O$) could be used to discriminate between areas (Fig. 2). These variables were transformed beforehand using the Box & Cox (1964) formulae to ensure normality. For this article, raw data were Box–Cox transformed using the function boxcox from the MASS library (Venables & Ripley, 2002) of the R package (www.r-project.org). This function estimates the likelihood profile for a range of λ values. The value showing the maximum likelihood was used for transforming the data using: $\text{transformed} = \lambda^{-1}(\text{raw}^{\lambda} - 1)$. Normality was tested with the Shapiro–Wilk's W-test and equality of variance with Levene's test. Both the raw and transformed data were concatenated into a data set that was used for subsequent analyses.

CHARACTERISTICS OF SELECTED VARIABLES

Raw trace element and stable-isotope data were non-normally distributed with skewed distributions (Fig. 2). As reported in Schaffler *et al.* (2014), the λ-values of variables following transformation ranged from −0·9064 for Mg to +1·4932 for O (Schaffler *et al.*, 2014), indicating that raw data distributions varied widely. After transformation, all trace element and

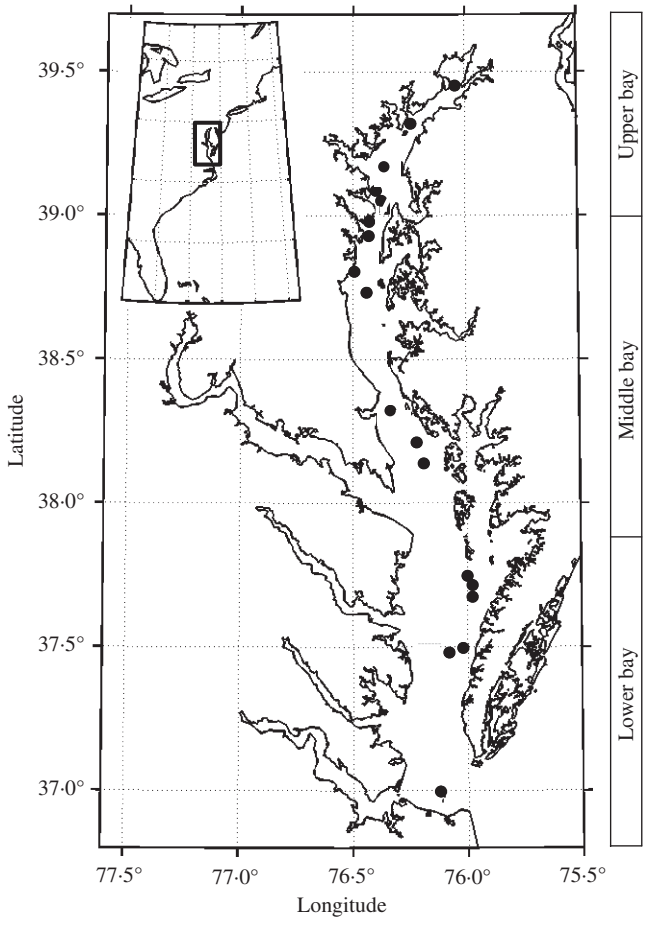

FIG. 1. Location (●) within Chesapeake Bay of otolith samples of juvenile *Brevoortia tyrannus* used in the analyses.

stable-isotope data had equal variances among areas but oxygen isotope ratios did not meet normality assumptions based on the Shapiro–Wilk's test (Fig. 2).

CLASSIFICATION TECHNIQUES

Four classification methods were applied to both the Box–Cox-transformed variables and the raw data variables to test the performance when assumptions of the methods were violated and were met. The two parametric techniques were LDFA and QDFA and the two machine algorithms were ANN and RF. The R implementation of the four classification functions was tested using the well-known Iris data set (Fisher, 1936). LDFA and QDFA were completed using the functions lda and qda from the MASS library in R. Equal prior probabilities were used for both lda and qda. ANN was implemented using the function nnet from the nnet library in R (Venables & Ripley, 2002). This function fits an ANN with a single intermediate layer. To select the values of the parameters of the function, a sensibility analysis was completed. This analysis consisted of comparing the kappa index after changing the values of the parameters one at a time. Fifty replicated sets of simulated data were used for each combination of values of the parameters. Details of the simulation procedure are provided below.

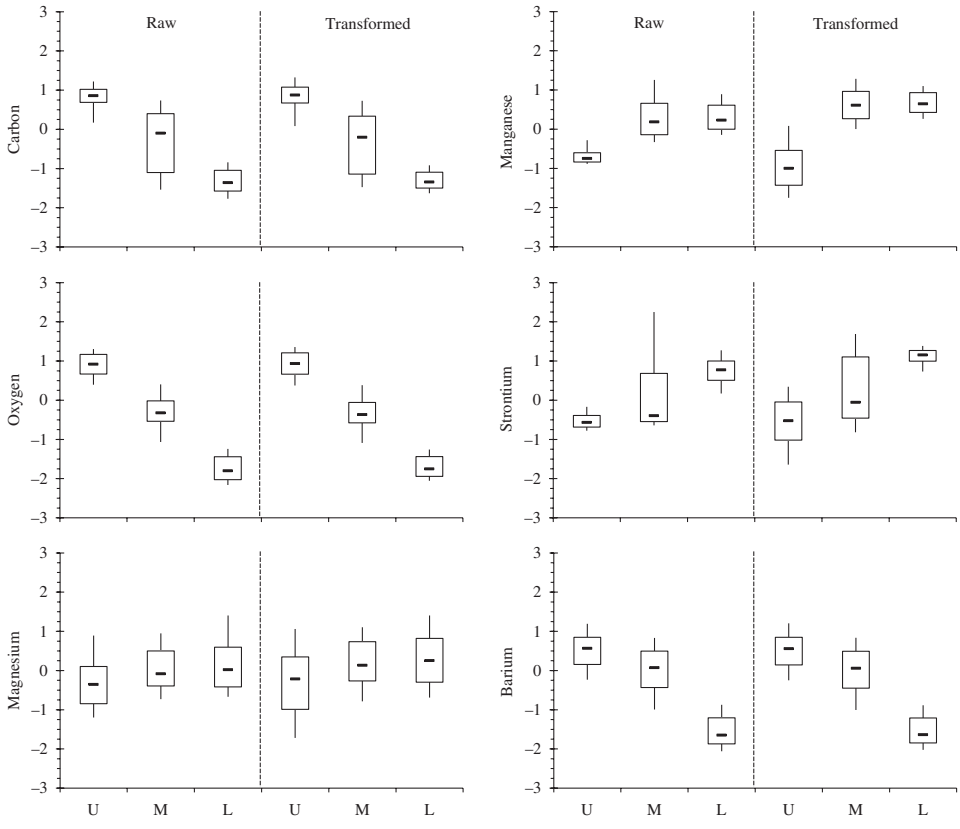

FIG. 2. Boxplots showing raw and Box–Cox transformed otolith chemistry variables used in classification comparisons. U, upper; M, middle; L, lower Chesapeake Bay. ▭, limits of the first and third quartiles; ▬, median; |, the expected range.

After the results of the sensitivity analysis, the following settings have been used: number of neurons at the intermediate layer = 30; logistic output (default), initial random weights = 0·5; weight decay = $5e^{-4}$; maximum number of iterations = 100 (default); maximum allowable number of weights = 1000 (default). RF was implemented using the randomForest function from the randomForest library in R (Liaw & Wiener, 2002). After the results of the sensitivity analysis, these settings have been used: number of trees to grow = 5000; number of variables randomly sampled as candidates at each split = square root of the number of variables (default); cutoff (the winning class for an observation is the one with the maximum ratio of proportion of votes to cutoff) = (number of populations)$^{-1}$ (default); minimum size of terminal nodes = 1 (default); size of samples to draw = 2/3 of the number of fish per population (default).

CROSS VALIDATION ANALYSIS WITH THREE GROUPS AND SIX VARIABLES

All four methods allow predicting the most probable group membership for fishes (subjects) from unknown source. Thus, it is possible to test the classification reliability of a method algorithm by comparing the predicted group membership with the true source. It is well known, however, that the rate of classification success is too optimistic when the subjects for which predictions are made are the same subjects used for parameterizing the classifications functions. Therefore, the available subjects are typically split into a training data set (used for building the

TABLE I. The kappa index (k) is defined by: $k = \langle (a + d)\{[(a + b)(a + c) + (d + b)(d + c)]N\}^{-1} \rangle \langle N - \{[(a + b)(a + c) + (d + b)(d + c)]N^{-1}\} \rangle^{-1}$, where N is the total number of subjects, given the agreement table for correct assignments, a is full positive agreement, b is negative to positive agreement, c is positive to negative agreement and d is full negative agreement

	+	−
+	a	b
−	c	d

classification functions) and a testing data set (used for predicting and comparing true and predicted group membership). Subjects were randomly split in a way that only one half of the fish from a population is used as training data set and the other half is used as testing data set. This splitting procedure was randomly repeated 1000 times. At each time, the rate of classification success was estimated for the four classification methods used. Instead of using the percentage of correct assignments as a measure of classification success, the kappa index (Fielding & Bell, 1997) was used because it is unbiased when the number of samples differs between groups (Table I).

SIMULATION EXPERIMENTS: MEAN-DIFFERENCE EFFECTS

The first simulation experiment was to compare the classification success in two well contrasted situations: (1) when between-group differences are small (large between-population overlap) and (2) when between-group differences are large (clear-cut groups). In order to emulate the real data, three populations were simulated. The differences between the three population means were achieved (1) by adding a fixed value (α) to the mean of one population ($M_2 = M_1 + \alpha$); (2) by subtracting the same quantity to the second population ($M_3 = M_1 - \alpha$); (3) by keeping unchanged the mean of the remaining population ($M_1 = M_1$). Each variable was modified independently; thus, the same population may experience a subtraction for one variable and an addition for another variable.

The value of α is specific for each variable and it is determined by the amount of the desired overlap between the population distributions (grey area in Fig. 3). The amount of overlap is defined by the probability that a given random sample drawn from the normal distribution on the left (Fig. 3) will be given by the reciprocal $(1 - P)$ of the cumulative distribution function: $(1 - P) = \left(\sqrt{2\pi\sigma^2}\right)^{-1} \int_{-\infty}^{M_1 + 0{\cdot}5\alpha} e^{-(x - M_1)(2\sigma^2)^{-1}}$, where M_1 is the mean of the left population (known) and σ is its S.D. (known and assumed to be the same for the three simulated populations). In the case of large between-population overlap (equals small discrimination), $P = 0{\cdot}45$ (45% overlap), thus the value of $(M_1 + 0{\cdot}5\alpha)$ is given by the built-in R function qnorm, a function that calculates normal quantiles for a particular P-value, $(0{\cdot}55, M_1, \sigma)$, where α is the only unknown quantity. Equivalently, different degrees of between-population overlap were simulated by setting $P = 0{\cdot}40$, $0{\cdot}35$ and $0{\cdot}30$.

To focus on the differences between means rather than on any differences in the magnitude of the S.D., one of the observed variance–covariance matrix (V) of one of the three real populations was selected and used throughout. Then, the function mvrnorm of the MASS library, which simulates a multivariate normal distribution, was used for generating random subjects from a multivariate normal distribution having the same variance–covariance matrix (V).

For a single iteration, the values of six variables were simulated for 100 fish (50 for the training data set and 50 for the testing data set). This data set was submitted to the same classification functions described for the real data (see cross validation analysis above), thus allowing comparison of the classification success of the four methods when the data strictly meet the assumptions of multivariate normality and common variance–covariance. Finally, the mean and the variability of the classification success were estimated from 1000 simulated data sets of 100 fish each.

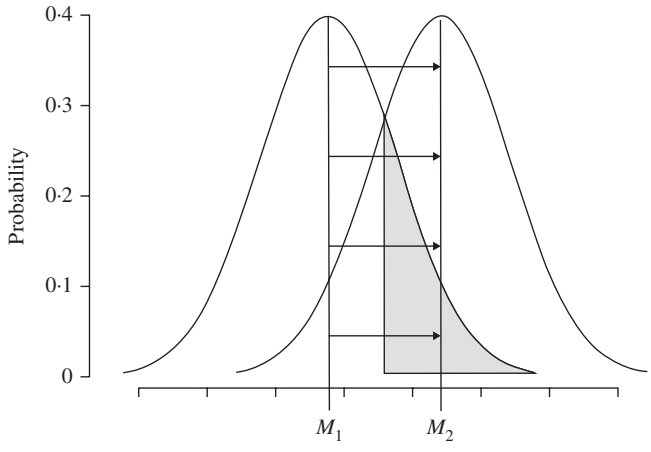

FIG. 3. Schematic representation of how simulations of population mean overlap were created, where M_1 = mean 1 and M_2 = mean 2.

SIMULATION EXPERIMENTS: SKEWNESS EFFECTS

The aim of the second simulation experiment was to compare the success of the four classification methods when distributions are skewed to a greater to lesser degree. Skewness is a measure of the asymmetry of a distribution. A normal distribution is not skewed and a distribution becomes more skewed when one of the two tails becomes larger than the other. Skewed data are commonly found in ecological studies and especially in otolith chemistry studies. The skewness of a sample is defined by: $g = \left[n^{-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^3\right]\left\{\left[n^{-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]^{-3/2}\right\}$. An unbiased estimator of g from a sample is given by: $g_{\text{sample}} = \sqrt{n\,(n-1)}g\,(n-2)^{-1}$.

In contrast to the first simulation experiment described above, where the subjects strictly meet very specific criteria, there are many ways a distribution can became skewed. The gamma distribution is a very convenient way for simulating asymmetric distributions because depending on only two parameters, it obtains a continuous range of distributions of progressive skewness. Gamma distributions are always right skewed (positive skewness) and are positive definite, as in most of otolith chemistry data sets (Ashford *et al.*, 2012; Catalan *et al.*, 2014; Morales-Nin *et al.*, 2014; Schaffler *et al.*, 2014). The skewness of a gamma distribution is given by: $g_{\text{gamma}} = 2\sigma m^{-1}$, where σ and m are the S.D. and the mean of the distribution. The consequence of this relationship is that it is not possible to simulate a gamma distribution with exactly the same g, σ and m than those estimated for the observed data because any given two parameters determine the third. Accordingly, the following strategy was adopted for simulating skewed but realistic samples for a given variable: (1) to estimate the mean and g_{sample} for the three groups of real data (upper, mid and lower Chesapeake Bay), (2) to compute the averaged g_{sample} for the three groups, (3) to compute the value of σ_{gamma} that gives a gamma distribution with the estimated (from the real data samples) group means and the averaged g_{sample} by: $\sigma_{\text{gamma}} = 0\cdot5 g_{\text{sample}} m_{\text{sample}}$ and (4) to randomly drawn new subjects from a gamma distribution using the build-in R function to simulate a gamma distribution, rgamma (n, shape, scale), where n is the number of random subjects to be generated, shape (S) is given by: $S_{ij} = m_{ij}^2\left(F\sigma^2\right)^{-1}$, where m is the sample mean of the i group and the j variable, σ is σ_{gamma} of the i group and the j variable and F is the factor that controls the skewness. Note that by multiplying σ for a factor F, the skewness is increased by the same factor F. Finally, scale (s) in the rgamma function from the R package is given by: $s_{ij} = m_{ij}S_{ij}^{-1}$. In this way, gamma-distributed, random subjects were produced with the same mean as the observed data and with the desired amount of skewness. Preliminary trials have shown when $F = 2$, the averaged skewness across all variables and groups is very close to the averaged skewness from

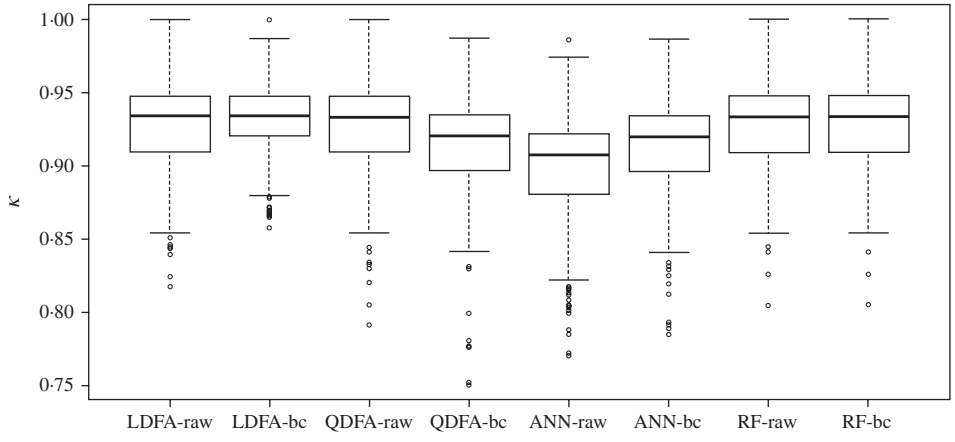

FIG. 4. Cross validation results comparing the classification success of four methods (LDFA, linear discriminant function analysis; QDFA, quadratic discriminant function analysis; ANN, artificial neural network; RF, random forest; raw, untransformed data; bc, data after Box–Cox transformation) measuring using the kappa index (k). Data were randomly split between training and testing data for 1000 times. ▭, limits of the first and third quartiles; ——, median; ⁞, the expected range; ○, outliers.

the real data. Therefore, two simulations were completed using $F = 2$ (*i.e.* simulated and data have similar skewness to the real data) and $F = 4$ (*i.e.* simulated data have skewness twice as large as the real data).

For a single iteration, values of six variables were simulated for 100 fish (50 for the training data set and 50 for the testing data set). This data set was submitted to the same classification functions described above (see cross validation analysis above), thus allowing the classification success of the four methods to be compared when the data have known skewness. Finally, the mean and the variability of the classification success were estimated from 1000 simulated data sets of 100 fish each.

RESULTS

APPLICATION OF FOUR METHODS DIRECTLY TO THE RAW AND TRANSFORMED DATA SETS

The results of the direct comparisons to show differences in classification success using the kappa index depending on whether the appropriate data are used in analyses (Fig. 4). All methods perform similarly when comparing raw and transformed data because departures from normality of the raw data are small. Nevertheless, the best performance is achieved by LDFA with transformed data ($k = 0{\cdot}93$). RF reaches virtually the same performance and showed no difference between raw and transformed data ($k = 0{\cdot}92$). In contrast, the performance of ANN was the poorest ($k = 0{\cdot}90$ for raw and 0·91 for transformed data). Between-trial variability (*i.e.* variability in performance using different training sets) may be interpreted as uncertainty in the estimated performance. The performance of ANN was the most variable ($\sigma = 0{\cdot}035$) and could occasionally be quite poor with kappa indices as low as 0·76. Conversely, between-trial variability of transformed LDFA is the smallest ($\sigma = 0{\cdot}024$).

TABLE II. Performance of four classification methods (LDFA; QDFA; ANN; RF) when group means overlap by 30-50% as measured by the averaged kappa index (1,000 simulations of 100 fish each) when data group-means have different levels of overlap. Numbers in bold indicate the largest kappa index within each comparison.

	No differences (50%)	Very Small (45%)	Small (40%) (Fig. 5A)	Clear (35%)	Very Clear (30%) (Fig. 5B)
% significant multivariate difference	5.2	96.7	100	100	100
% at least one significant univariate difference	5.8	100	100	100	100
LDFA	0.0	**0.31**	**0.66**	**0.83**	**0.98**
QDFA	0.0	0.25	0.62	0.81	0.98
ANN	0.0	0.14	0.49	0.73	0.95
RF	0.0	0.21	0.49	0.72	0.92

MEAN-DIFFERENCE EFFECTS FOR SIMULATED DATA

Results comparing the performance of the four methods with group-mean differences (Table II) show that all methods perform well when group-mean differences were large (overlap only 30%), ranging from 0·92 for RF to 0·98 for LDFA [Fig. 5(a)]. Performance decreased with increasing overlap. At 40% overlap, kappa indices ranged from 0·49 for RF to 0·66 for LDFA [Fig. 5(b)]. When differences were very small (45% overlap), performance was poor for all methods with the best performance by LDFA (kappa index, $k = 0{\cdot}31$) (Table II). When assumptions are strictly met, LDFA has the best overall performance, and the parametric methods performed better than the machine-learning methods.

SKEWNESS SIMULATIONS

As skewness increases in the data, classification performance declines across all methods (Fig. 6). In the case of raw-level skewness [Fig. 6(a)], the two parametric methods performed better with Box–Cox transformed data, with QDFA having the best performance of all the methods. RF performed almost as well as QDFA and there were no differences between transformed or raw data. In contrast, ANN performed more poorly than the other methods. When skewness was doubled [Fig. 6(b)], the pattern seen previously for parametric methods was again demonstrated between transformed and raw data, with both LDFA and QDFA performing better with transformed data. This pattern was also seen with ANN performing better with transformed data. No difference was noted with RF whether data were transformed or not.

DISCUSSION

When the assumptions of normality were met, the traditional parametric methods of LDFA and QDFA provide the best classification success. This advantage decreased when group-mean differences were large and all methods perform almost equally well,

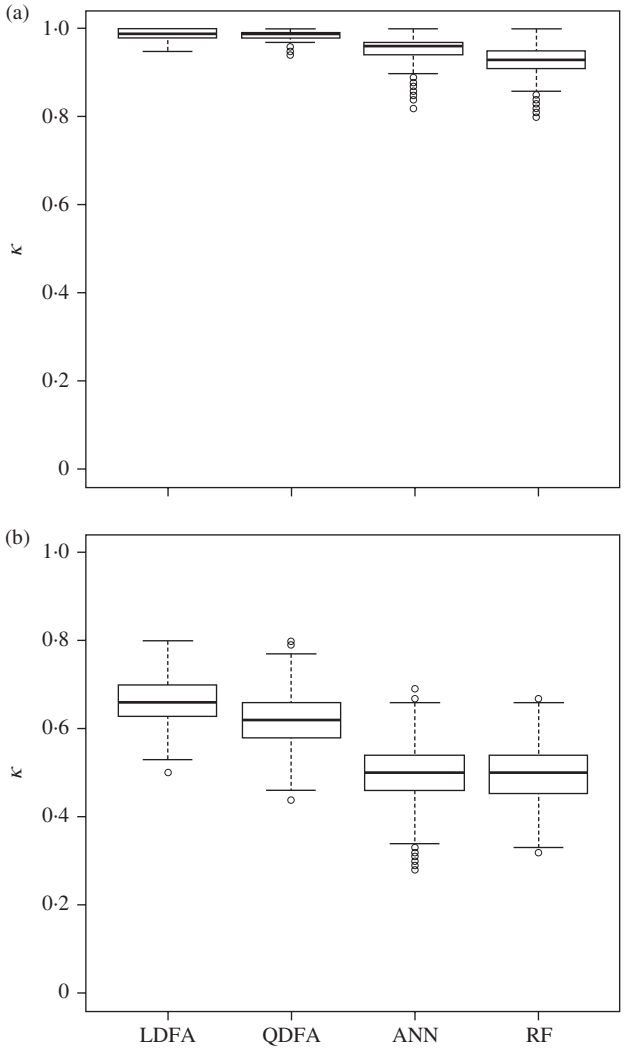

FIG. 5. Classification success for four methods (LDFA, linear discriminant function analysis; QDFA, quadratic discriminant function analysis; ANN, artificial neural network; RF, random forest) measured with the kappa index (k) for (a) smaller overlap of 30% of group means and (b) large overlap of 40%. ▭, limits of the first and third quartiles; ——, median; ⁞, the expected range; ○, outliers.

albeit the parametric methods perform less well when their normal assumptions were not met. Although the advantages in performance that were found for parametric methods were not large, classification errors can propagate when the classification functions are next applied to classify unknown populations, such as using juvenile chemistries to classify adults back to their nursery grounds (Ashford *et al.*, 2012; Schaffler *et al.*, 2014). This error propagation has not been explicitly discussed in the otolith chemistry literature, nor have the potential increases in uncertainty been modelled due to error propagation. Nonetheless, small improvements in classification success are especially important when error is multiplicative.

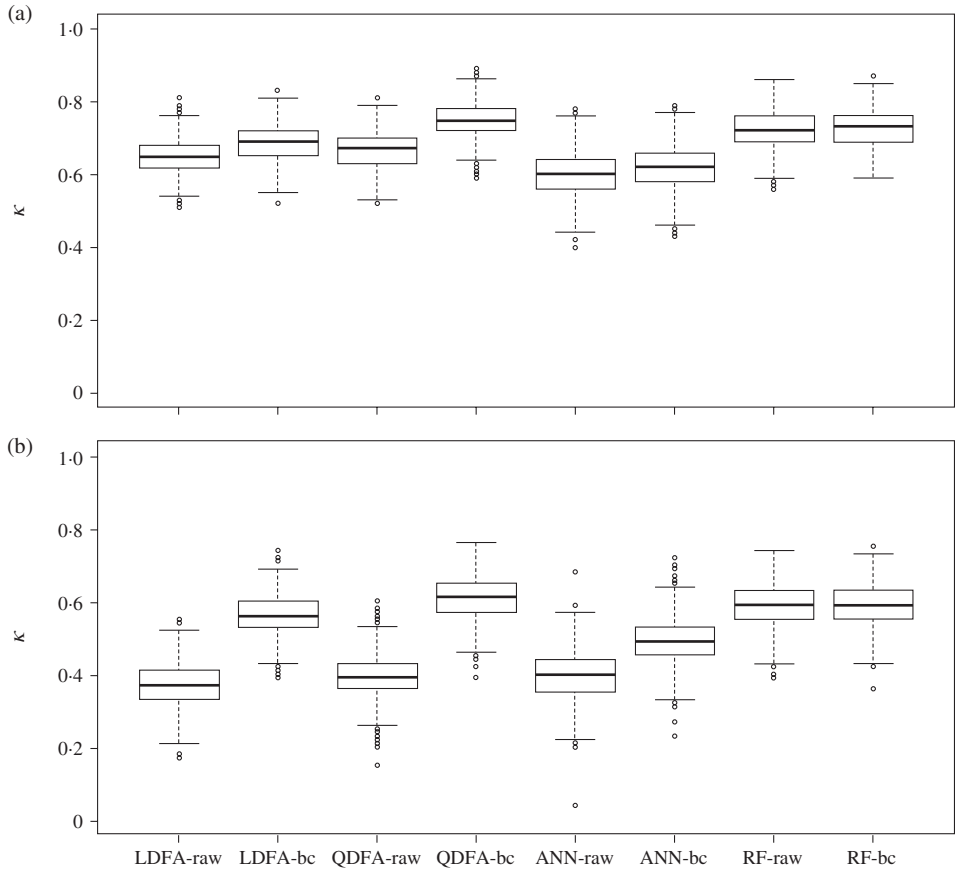

FIG. 6. Classification success for four methods (LDFA, linear discriminant function analysis; QDFA, quadratic discriminant function analysis; ANN, artificial neural network; RF, random forest; raw, untransformed data; bc, data after Box–Cox transformation) measured with the kappa index (k) for (a) low skew and (b) high skew. ▭, limits of the first and third quartiles; ——, median; ⁞, the expected range; ○, outliers.

It is important to understand the effects of meeting underlying assumptions of parametric models and how this affected results. For example, Mercier *et al.* (2011) found that the RF method outperformed discriminant functions for otolith chemistry data. Moreover, they recommended the use of RF over the parametric methods. These comparisons, however, were performed on raw data, thus, parametric assumptions were not met. Shown here, careful analysis of method performance considering normality, mean-group differences and skewness revealed more nuanced results. Note that both real and simulated data here corresponded to cases with relative large numbers of fish per population (83 and 100 fish on average, respectively). Therefore, this setting was particularly favourable to computer-intensive methods. In contrast, most otolith chemistry papers considered far smaller sample sizes for which parametric methods were expected to perform better. In addition, most frequently, the parametric methods can be robust to problems such as skewness and can perform well even when the assumptions of normality are violated. When skewness is severe, however, they underperformed.

An issue that is infrequently addressed is the choice of variables to include in analysis. Statisticians often use analytic methods, such as Rao's (1965) test, that can be used to assign significance to the contribution of variables in the classification function. The value that variables add in building classification functions in machine-algorithm methods has been evaluated through computation-intensive approaches (Mercier *et al.*, 2011). There is still no common practice for choosing variables.

Even given the better performance of parametric methods, some otolith chemistry data can be sufficiently non-normal and skewed that they cannot be normalized even with Box–Cox transformations. Although parametric classification methods can still perform well in the face of some non-normality, they do poorly when non-normality is large and uncorrected. With such data, the choice of a machine-algorithm method is recommended.

It is not always clear in the methods sections of otolith chemistry papers that use parametric methods if normality has actually been achieved. Many papers state that data were log transformed, when that alone may not normalize data. Results of this article confirm that parametric methods are somewhat robust to non-normality, but regardless, normality should be checked by, for example, the Shapiro–Wilk's W-test. It is inadvisable to use a method without testing that assumptions are met.

It is shown here using typical otolith-chemistry data that when assumptions are met then parametric methods are more powerful, as has been widely recognized theoretically in the statistical literature and demonstrated here through simulated data sets that emulate those typically obtained when using otolith chemistry in number of variables, magnitude of between population differences, skewness and number of analysed fish per population. It is recommended that when tests show data to be non-normally distributed, then Box–Cox transformations applied, followed by the evaluation for skewness. If an obvious level of skewness is present, then machine-algorithm methods should also be tried.

References

Anstead, K., Schaffler, J. J. & Jones, C. M. (2015). Coast-wide juvenile otolith signatures of the Atlantic menhaden *Brevoortia tyrannus*, 2009–2011. *Transactions of the American Fisheries Society* **144,** 96–106.

Armitage, D. W. & Ober, H. K. (2010). A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics* **5,** 465–473.

Ashford, J. R., Fach, B. A., Arkhipkin, A. I. & Jones, C. M. (2012). Testing early life connectivity supplying a marine fishery around the Falkland Islands. *Fisheries Research* **121–122**, 144–152.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* **26,** 211–252.

Campana, S. E. (1999). Chemistry and composition of fish otoliths: pathways, mechanisms and applications. *Marine Ecology Progress Series* **188,** 263–297.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Perna, C. & Brunskill, G. (2005). Development of a robust classifier of freshwater residence in barramundi (*Lates calcarifer*) life histories using elemental ratios in scales and boosted regression trees. *Marine and Freshwater Research* **56,** 713–723.

Catalan, I. A., Perez-Mayol, S., Alvarez, I., Ruiz, J., Palmer, M., Baldo, F. & Morales-Nin, B. (2014). Daily otolith growth and ontogenetic geochemical signatures of age-0 anchovy (*Engraulis encrasicolus*) in the Gulf of Cádiz (SW Spain). *Mediterranean Marine Science* **15,** 781–789.

Choi, S. C. (1986). Discrimination and classification: overview. *Computers & Mathematics with Applications A* **12,** 173–177.

Dorval, E., Jones, C. M., Hannigan, R. & van Montfrans, J. (2005). Can otolith chemistry be used for identifying essential seagrass habitats for juvenile spotted seatrout, *Cynoscion nebulosus*, in Chesapeake Bay? *Marine and Freshwater Research* **56,** 645–653.
Elsdon, T. S., Wells, B. K., Campana, S. E., Gillanders, B. M., Jones, C. M., Limburg, K. E., Secor, D. H., Thorrold, S. R. & Walther, B. D. (2008). Otolith chemistry to describe movements and life-history parameters of fishes: hypotheses, assumptions, limitations and inferences. *Oceanography and Marine Biology: An Annual Review* **46,** 297–330.
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.
Fisher, R. A. (1936). The use of multiple measures in taxonomic problems. *Annals of Eugenics* **7,** 179–188.
Gaines, S. D., Gaylord, B., Gerber, L. R., Hastings, A. & Kinlan, B. P. (2007). Connecting places: the ecological consequences of dispersal in the sea. *Oceanography* **20,** 90–99. doi: 10.5670/oceanog.2007.32
Hand, D. J. (1981). *Discrimination and Classification*. New York, NY: Wiley.
Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer.
Jones, C. M. (2006). Estuarine and diadromous fish metapopulations. In *Marine Metapopulations* (Kritzer, J. P. & Sale, P. F., eds), pp. 119–156. New York, NY: Academic Press.
Legendre, P. & Legendre, L. (2012). *Numerical Ecology*, 3rd edn. Amsterdam: Elsevier.
Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News* **2,** 18–22.
Mercier, L., Darnaude, A. M., Bruguier, O., Vasconcelos, R. P., Cabral, H. N., Costa, M., Lara, M., Jones, D. L. & Mouillot, D. (2011). Selecting statistical models and variable combinations for optimal classification using otolith chemistry. *Ecological Applications* **21,** 1352–1364.
Morales-Nin, B., Pérez-Mayol, S., Palmer, M. & Geffen, A. J. (2014). Coping with connectivity between populations of *Merluccius merluccius*: an elusive topic. *Journal of Marine Systems* **138,** 211–219.
Olden, J. D., Lawler, J. J. & Poff, N. L. (2008). Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology* **83,** 171–193.
Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York, NY: Wiley & Sons.
Recknagel, F. (2001). Applications of machine learning to ecological modeling. *Ecological Modelling* **146,** 303–310.
Schaffler, J. J., Reiss, C. S. & Jones, C. M. (2009). Spatial variation in otolith chemistry of Atlantic croaker larvae in the Mid-Atlantic Bight. *Marine Ecology Progress Series* **382,** 185–195.
Schaffler, J. J., Miller, T. & Jones, C. M. (2014). Spatial and temporal variation in otolith chemistry of juvenile Atlantic menhaden in Chesapeake Bay. *Transactions of the American Fisheries Society* **143,** 1061–1071.
Suryanarayana, I., Braibanti, A., Rao, R. S., Raman, V. A., Sudarsan, D. & Rao, G. N. (2008). Neural networks in fisheries research. *Fisheries Research* **92,** 115–139.
Thorrold, S. R., Latkcozy, C., Swart, P. K. & Jones, C. M. (2001). Natal homing in a marine fish metapopulation. *Science* **291,** 297–299.
Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edn. New York, NY: Springer.

Electronic Reference

Shiffman, D. (2012). *The Nature of Code: Simulating Natural Systems with Processing*. The Magic Book Project. Available at http://natureofcode.com/book/