Use of morphometric characters of a fish species to predict its location; a statistical approach

[1]Thilan, A.W.L.P., [2]De Silva, M.P.K.S.K., and [3]Jayasekara, L.A.L.W.

[1,3]Department of Mathematics, [2]Department of Zoology,

University of Ruhuna,

Matara,

Sri Lanka

02 November, 2010

Abstract

Precise taxonomic identification is the preliminary requirement in a study of an organism/specimen. Correct identification however gives only the identity of the specimen. The value of the correctly identified specimen as a study material becomes low when the habitat/location of its collection is unknown. Knowing the exact place of collection, enables to gather information on the distribution of the organism, possible environmental conditions that the organisms encounter and to describe the variations found in morphological and genetic features of the organism.

Present study therefore, aimed on to develop a statistical rule to predict place of collection (river which is unknown) of a given *Puntius dorsalis* (a freshwater fish species) specimen using its morphometric characters.

Fifty two individuals were collected from four major rivers (*Mahaweli, Kelani, Kalu, Nilwala*) in Sri Lanka and 23 morphometric characters were measured from each specimen. Those individuals were categorized into 4 groups according to the river from which they were collected. Measured morphometric characters were used as independent variables of the model to predict unknown group membership (river) of a given *Puntius dorsalis* specimen.

In the case of re-substitution, 82.7% of the *Puntius dorsalis* specimens were successfully classified or predicted with respect to the place of collection (river) using their posterior probabilities. The process had a hit ratio of 69.2% when generalized, as a valid tool to classify fresh *Puntius dorsalis* specimen of unknown group membership. It was also discovered that linear classification function could be used to predict unknown place of collection of a fish. The paper concludes with some suggestions to move into nonparametric approach like *Classification and Regression Trees (CART)* and *Neural Networks*.

Keywords: Linear classification functions, Prediction, *Puntius dorsalis*, Discriminant analysis, Morphometric.

1 Introduction

There are two aspects in discriminant analysis. They are Predictive Discriminant Analysis (PDA) and Descriptive Discriminant Analysis (DDA). Dissimilarities between these two analyzes are not well understood by most researchers [10]. Predictive discriminant analysis focuses on classifying subjects into one of several groups (or to predicate group membership), whereas in descriptive discriminant analysis, the focus is on revealing major differences among the groups. Hence, PDA is appropriate when the researcher is interested in assigning units (individuals) to groups based on composite scores on several predictor variables [28, 25, 11, 15]. The accuracy of such prediction can be assessed by examining "hit rates" as against chance [12].

Discriminant functions (DFs) are linear combinations of independent variables [15] and the first DF is that which maximally separates the groups and the second DF, orthogonal to the first, maximally separates the groups on variance not yet explained by the first DF [20, 30]. Classification functions (CFs) are again linear combinations where coefficients are different from coefficients of DFs [15]. In fact, there will be k classification functions and $s = min(p, k-1)$ discriminant functions, where k is the number of groups and p is the number of variables. DFs are used for both PDA and DDA aspects [1] and CFs are used for PDA. In many cases we do not need all DFs to effectively describe group differences (DDA aspects), whereas all k classification functions must be used in assigning observations to groups [25, 30].

In biological experiments and research, knowledge on place of collection of an organism is one of the preliminary requirements to study about that organism. Place of collection could provide information on the habitat, environmental conditions, adaptations, morphological and genetic variability of the specimen. In addition, place of collection is important for the taxonomic studies too. In museum collections, there are many specimens with unknown identity of place of collections, have lowered their value as a biological specimen.

Morphometric characters are measurable linear measurements of a fish, and are known to vary with the factors like river, altitude range, environmental conditions of the habitats [19, 16, 21, 7, 9]. *Puntius dorsalis* is a fish species commonly found in freshwater bodies of Sri Lanka. *Puntius dorsalis* have also shown distinct morphometric heterogeneity with respect to rivers, altitude and

environmental factors [8]. This morphological variability present in *Puntius dorsalis* enabled us to use it as a test model for the present study. Objective of the present study is to develop a statistical rule to predict the place of collection (river which is unknown) of a *Puntius dorsalis* specimen using its morphometric characters.

2 Materials and Methods

2.1 Data collection

A sample of 52 *Puntius dorsalis* specimens were collected from seventeen sites of four major rivers *Nilwala, Kalu, Kelani and Mahaweli*, basins in Sri Lanka such that 10, 8, 17 and 17 specimens representing those rivers respectively (**Table. 1**). Fish were caught using gape nets, cast nets and scoop nets. Twenty three morphometric characters were recorded from each specimen (**Figure. 1**).

Linear measurements were made using vernier calipers to the nearest 0.01 millimeter. All morphometric variables were standardized to remove the effect of individual size. In that case, *eye diameter (ED)* and *post orbital length (POL)* were divided by *head length (HL)* and all other variable shown in (**Figure. 1**) were divided by *standard length (SL)* to remove the effect of individual size. "Log" and "Arcsine square-root" transformations were carried out separately for skewed and proportional morphometric characters respectively [4, 22].

Method of Stepwise Discriminant Analysis was used to find variables made significant independent and combined contributions [13, 29]. These were the *caudal fin length (CFL), standard length (SL), head length (HL), length of the caudal peduncle (LCPD), anal fin length (AFL), prevental length (PVL)* and *eye diameter (ED).*

2.2 Data analysis

The aim of the study was to build up a statistical rule to predict place of collection (river which is unknown) of a given *Puntius dorsalis* specimen. In this case, given *Puntius dorsalis* specimen whose group membership (river) unknown is to be classified into one of four groups formed based on their place of collection. Unknown group membership can be predicted using posterior probability [5, 2],

River	Location	Altitude range
Mahaweli R- 4	Habarana	0-150 m
	Parakrama Samudra	0-150 m
	Ginnoruwa	0-150 m
	Arawa	151-300 m
	Pallegama	301-600 m
Kelani R-3	Biyagama	0-150 m
	Bulathkohupitiya	0-150 m
	Pinnawala	0-150 m
	Kitulgala	151-300m
Kalu R-2	Agalawatta	0-150 m
	Handurukanda	151-300 m
	Athweltota	301-600 m
	Kalawana	601-1200 m
Nilwala R-1	Godapitiya	0-150 m
	Deiyandara	151-300m
	Dediyagala	151-300 m
	Mawarala	301-600 m

Table 1. River, name of the location and altitude range of *P. dorsalis* collected .

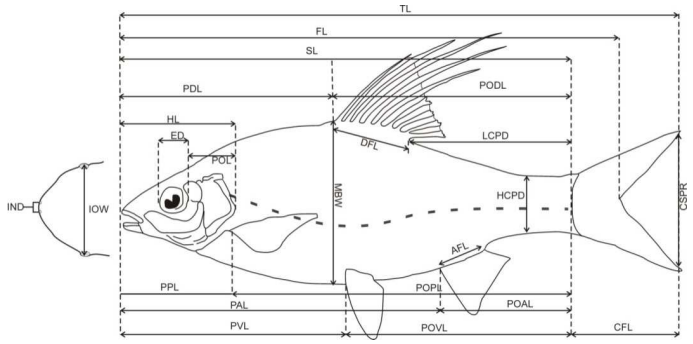

Figure 1: *Total length, TL; Standard length, SL; Fork length, FL; Maximum body depth, MBW; Head length, HL; Eye diameter, ED; Distance between pair of nostrils, IND; Inter orbital distance, IOW; Post orbital length, POL; Dorsal fin length, DFL; Pre dorsal length, PDL; Post dorsal length, PODL; Anal fin length, AFL; Pre anal length, PAL; Post anal length, POAL; Pre ventral length, PVL; Post ventral length, POVL; Pre pelvic length, PPL; Post pelvic length, POPL; Caudal fin length, CFL; Width of the caudal fin when fully spread, CSPR; Caudal peduncle height, HCPD; end of the dorsal fin to end of the caudal peduncle length, LCPD.*

CFs [24], CART [3] and Neural networks [17, 20] etc. The data were processed using SPSS statistical software package [13, 22], and it classifies subjects into predicted groups using posterior probability. Bayes' rule is used for posterior probability method.

2.3 Bayes' rule

If there are k groups, the Bayes' rule is to assign the object to group G_i, where

$$p(G_i|D) > p(G_j|D) \qquad \text{for all } j \neq i \text{ and } i, j = 1, 2, ..., k. \tag{1}$$

We want to know the probability $p(Gi|D)$ that an object belongs to group G_i, given the values D on each of the DFs. The subject is then classified (predicted) to be in the group with the higher posterior probability [13]. There is a relationship between the two conditional probabilities that well known as Bayes Theorem:

$$p(G_i|D) = \frac{p(D|G_i)p(G_i)}{\sum_i^k p(D|G_i)p(G_i)} \tag{2}$$

Prior probability $p(G_i)$ is an estimate that belongs to a particular group when no information about it is available. However, to use the Bayes rule directly, it is hard [2] to calculate $p(D|G_i)$.

If we assume that all groups have the same covariance matrix, then we get another classification technique which is called Fisher's Linear Classification [25, 11].

2.4 Multivariate test of significance

Box's M test [22] was used to evaluate homogeneity of covariance matrix and test statistic is not significant under 0.01 significance level (**Table. 2**). Therefore, covariance matrix homogeneity is acceptable. The *mshapiro-wilk* test [26] was used to evaluate multivariate normality and test is significant for each groups under 0.01 significance level (**Table. 3**). So, as a statistical test multivariate normality is violated for each of the four groups and p values indicate that violation is higher in groups one and two than other two groups. That is also confirmed from visual inspection (**Figure. 2**), based on the fact that straight lines indicate multivariate normally distributed data and deviation from the straight line is a measure of deviation from multivariate normality [12].

Box's M	121.498
F Approx.	0.974
df1	84
df2	2509.387
Sig.	0.548

Table 2. Box's M test.

2.5 Linear classification analysis

We use samples from each of the k groups to find the sample mean vectors $\overline{\mathbf{y}}_\mathbf{1}, \overline{\mathbf{y}}_\mathbf{2}, ..., \overline{\mathbf{y}}_\mathbf{k}$. As a univariate approach can not address any joint effect (interactions) of variables, each individual was considered to be a single multivariate observation in the analysis [32, 6]. Specimen whose group membership is unknown, one approach is to use a distance function [12] to find the mean vector that $\mathbf{y}$ (independent variables measured for a fresh *Puntius dorsalis* specimen) is closest to and assign specimen to the corresponding group. We can estimate the common population covariance matrix [25] by a pooled sample covariance matrix,

$$\mathbf{S_{pl}} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S_i} = \frac{\mathbf{E}}{N-k}, \tag{3}$$

where n_i and $\mathbf{S_i}$ are the sample size and covariance matrix of the i^{th} group, $\mathbf{E}$ is the error matrix from one-way MANOVA, and $N = \sum_i n_i$. We compare $\mathbf{y}$ to each $\overline{\mathbf{y}}_\mathbf{i}, i = 1, 2, ..., k$ by the distance function,

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \overline{\mathbf{y}}_\mathbf{i})'\mathbf{S_{pl}^{-1}}(\mathbf{y} - \overline{\mathbf{y}}_\mathbf{i}) \tag{4}$$

and assign $\mathbf{y}$ to the group for which $D_i^2(\mathbf{y})$ is smallest.

Group	Test statistic and p-value
1	w=0.4292, p-value=5.432e-07
2	w=0.4184, p-value=1.047e-06
3	w=0.7085, p-value=0.0001447
4	w=0.7119, p-value=0.0001582

Table 3. mshapiro test for normality.

We can obtain simplified form as a linear classification rule by expanding (4),

$$\begin{aligned} D_i^2(\mathbf{y}) &= \mathbf{y}'\mathbf{S_{pl}^{-1}}\mathbf{y} - \mathbf{y}'\mathbf{S_{pl}^{-1}}\overline{\mathbf{y}}_\mathbf{i} - \overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\mathbf{y} + \overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\overline{\mathbf{y}}_\mathbf{i} \\ &= \mathbf{y}'\mathbf{S_{pl}^{-1}}\mathbf{y} - \mathbf{2}\overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\mathbf{y} + \overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\overline{\mathbf{y}}_\mathbf{i}. \end{aligned} \tag{5}$$

The first term $\mathbf{y}'\mathbf{S_{pl}^{-1}}\mathbf{y}$ can be neglected because it does not change from group to group. But second term is a linear function of $\mathbf{y}$, and the third does not involve $\mathbf{y}$. So by multiplying $-\frac{1}{2}$ and deleting first term we get linear classification function and denote by $L_i(\mathbf{y})$,

$$L_i(\mathbf{y}) = \overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\mathbf{y} - \frac{1}{2}\overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\overline{\mathbf{y}}_\mathbf{i}, \qquad\qquad i = 1, 2, ..., k. \tag{6}$$

Assign $\mathbf{y}$ to the group for which $L_i(\mathbf{y})$ is a maximum [15] (sign reversed when multiplying by -1/2). To highlight the linearity of (6) as a function of $\mathbf{y}$, we can express it as,

$$L_i(\mathbf{y}) = \mathbf{c}_\mathbf{i}'\mathbf{y} + c_{i0} = c_{i1}y_1 + c_{i2}y_2 + ... + c_{ip}y_p + c_{i0}, \tag{7}$$

where $\mathbf{c}_\mathbf{i}' = \overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}$ and $c_{i0} = -\frac{1}{2}\overline{\mathbf{y}}_\mathbf{i}'\mathbf{S_{pl}^{-1}}\overline{\mathbf{y}}_\mathbf{i}$. To assign $\mathbf{y}$ to a group using this procedure, we calculate $\mathbf{c_i}$ and c_{i0} for each of the k groups, evaluate $L_i(\mathbf{y})$, $i = 1, 2, ..., k$, and allocate $\mathbf{y}$ to the group for which $L_i(\mathbf{y})$ is largest. This will be the same group for which $D_i^2(\mathbf{y})$ in (4) is smallest, that is, the group whose mean vector $\overline{\mathbf{y}}_\mathbf{i}$ is closest to $\mathbf{y}$. In order to use the prior probabilities, the density functions for the two populations, $f(\mathbf{y}|G_1)$ and $f(\mathbf{y}|G_2)$, must also be known. Then the optimal classification

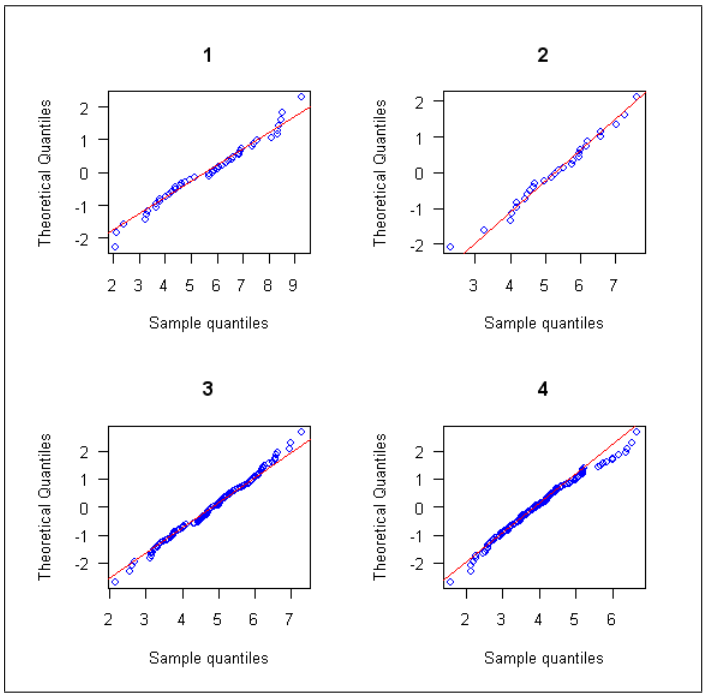

Figure 2: Visual inspection of multivariate normality for four groups.

rule [31] that minimizes the probability of misclassification is: Assign $\mathbf{y}$ to G_1 if,

$$p_1 f(\mathbf{y}|G_1) > p_2 f(\mathbf{y}|G_2) \tag{8}$$

and to G_2 otherwise. Note that $f(\mathbf{y}|G_1)$ is a convenient notation for the density when sampling from the population represented by G_1. It does not represent a conditional distribution in the usual sense. For the case of several groups, the optimal rule in (8) extend to,

$$\text{Assign } \mathbf{y} \text{ to the group for which } p_i f(\mathbf{y}|G_i) \text{ is maximum.} \tag{9}$$

The probability of misclassification is minimized with this rule. If we assume normality with equal covariance matrices and with prior probabilities of group membership $p_1, p_2, ..., p_k$, then $f(\mathbf{y}|Gi) = N_p(\mu_\mathbf{i}, \mathbf{\Sigma})$, and the rule in (9) becomes (with estimates in place of parameters) [24]: Calculate

$$L_i'(\mathbf{y}) = \ln p_i + \overline{\mathbf{y}}_\mathbf{i}' \mathbf{S}_\mathbf{pl}^{-\mathbf{1}} \mathbf{y} - \frac{1}{2} \overline{\mathbf{y}}_\mathbf{i}' \mathbf{S}_\mathbf{pl}^{-\mathbf{1}} \overline{\mathbf{y}}_\mathbf{i}, \qquad i = 1, 2, ..., k \tag{10}$$

			Predicted Group Membership				
		INDEX	1.00	2.00	3.00	4.00	Total
Original	Count	1.00	8	0	1	1	10
		2.00	0	6	1	1	8
		3.00	2	0	15	0	17
		4.00	1	1	1	14	17
	%	1.00	80.0	.0	10.0	10.0	100.0
		2.00	.0	75.0	12.5	12.5	100.0
		3.00	11.8	.0	88.2	.0	100.0
		4.00	5.9	5.9	5.9	82.4	100.0
Cross-validated [a]	Count	1.00	5	1	2	2	10
		2.00	0	5	1	2	8
		3.00	3	0	14	0	17
		4.00	1	2	2	12	17
	%	1.00	50.0	10.0	20.0	20.0	100.0
		2.00	.0	62.5	12.5	25.0	100.0
		3.00	17.6	.0	82.4	.0	100.0
		4.00	5.9	11.8	11.8	70.6	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 82.7% of original grouped cases correctly classified.

c. 69.2% of cross-validated grouped cases correctly classified.

Table 4. Classification result for analysis and cross validated samples.

and $\mathbf{y}$ is assigned to the group with maximum value of $L_i'(\mathbf{y})$. If $p_1 = p_2 = ... = p_k$, then (10), which optimizes the classification rate for the normal distribution, reduces to (6), which was based on the heuristic approach of minimizing the distance of $\mathbf{y}$ to $\overline{\mathbf{y}}_\mathbf{i}$.

3 Results and Discussion

SPSS statistical software package uses posterior probability method for classification. In that case, calculates the posterior probabilities of being in each of the four groups and a subject is then classified (predicted) to be in the group with the higher posterior probability (**Table. 4**). The misclassification rate for each group is the proportion of sample observations in that group that are misclassified.

In the case of re-substitution [23, 1], 82.7% of *Puntius dorsalis* specimens were correctly classified into their respective rivers. But, it is not an unbiased estimator for actual correct classification rate. Because, the data set used to compute the DFs are also used to evaluate them [25, 14, 18, 12]. But,

cross validation [23, 1] treats $n-1$ out of n training observations as a training set. It determines the DFs based on these $n-1$ observations and then applies them to classify the one observation left out. This is done for each of the n training observations. For this study, cross validated correct classification rate is 69.2%. It is nearly an unbiased estimate but with a relatively large variance [25, 14, 18, 12]. Both analysis and cross validated correct classification rates are not so good. One reason for that is, violation of multivariate normality and it can be avoided by increasing the sample sizes and removing unnecessary variables from the study [5, 14, 25].

After doing transformation for data, we should reassess model assumptions to see the effect of it to the distribution of data. In this study, we did "Log" and "Arcsine square-root" transformations for skewed and proportional variable respectively. But, reassessment confirmed that those transformations have no any acceptable effect to increase distributional characteristics and it merely changed our measured variables. Therefore, "Log" and "Arcsine square-root" transformations are not necessary for these data.

The combined groups plot [15] (**Figure. 3**) can be used to see where each specimen (in training data set) falls in the space defined by the first two DFs and it emphasizes good separation of groups formed based on place of collection (rivers). But our intention is to classify future observation, that is a fresh (not a member of training data) *Puntius dorsalis* specimen caught from any of four rivers (*Mahaweli, Kalu, Kelani and Nilwala*) and exactly from which river is unknown.

In its original form proposed by Fisher, the method assumes equality of population covariance matrices, but does not explicitly require multivariate normality. However, optimal classification performance of Fisher's discriminant function can only be expected when multivariate normality is present as well, since only good discrimination can ensure good allocation [27]. That means, when a parametric classification criterion (linear) is derived from a non normal (critical violation) population, the probability of misclassification is high [25].

If the covariance matrices are unequal but the joint distribution of the variables is normal, then the quadratic classification rule is the optimum one. However, if the covariance matrices are not too dissimilar, the linear classification performs quite well, especially if the sample sizes are small [13]. In this study sample sizes are small and main assumption of linear classification analysis that is

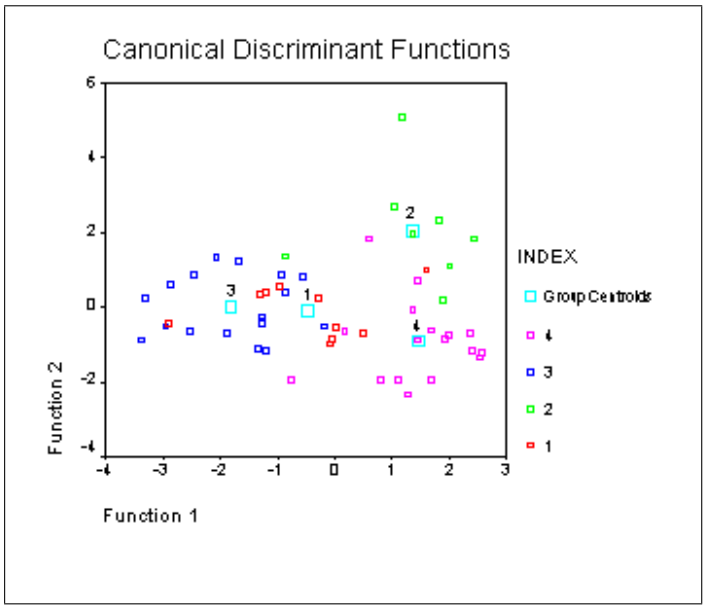

Figure 3: Combined groups plot.

homogeneity of covariance is satisfied. Because of that, for this particular study, linear classification is the appropriate parametric criterion.

SPSS statistical software package outputs the coefficient [15] for linear classification functions and they can be used to construct following four CFs.

$$L_1'(\mathbf{y}) = 59.50\ SL + 4218.35\ HL + ... - 766.59\ CFL + 1985.98\ LCPD - 1776.20 \quad (11)$$

$$L_2'(\mathbf{y}) = 62.90\ SL + 4502.39\ HL + ... - 794.79\ CFL + 2174.09\ LCPD - 1939.43 \quad (12)$$

$$L_3'(\mathbf{y}) = 57.97\ SL + 4220.49\ HL + ... - 749.61\ CFL + 2128.59\ LCPD - 1781.26 \quad (13)$$

$$L_4'(\mathbf{y}) = 61.10\ SL + 4354.19\ HL + ... - 842.97\ CFL + 2086.35\ LCPD - 1880.29 \quad (14)$$

Classification functions can directly be used to predict group membership. The groups (R-3) and (R-4), represent the fish collected from *Kelani* and *Mahaweli* rivers respectively (**Table. 2**). Suppose two fresh (not used to calculate CFs) *Puntius dorsalis* specimens, one came from *Mahaweli* (R-4) and the other one from *Kelani* (R-3) rivers were considered and following are the two observation vectors corresponding to them respectively. We need only to measure and do relevant standardization for

morphometric variables which are appeared in CFs.

$$\begin{aligned}\mathbf{y_4} &= (6.54, 0.281346, 0.288043, 0.09633, 0.480122, 0.214067, 0.183486)\\ \mathbf{y_3} &= (6.14, 0.267101, 0.304878, 0.081433, 0.477199, 0.179153, 0.184039)\end{aligned}$$

Linear CFs can be used to predict place of collection of above two specimens and since we knew real place of collections (two rivers namely *Mahaweli* and *Kelaniya*), it is a measure of the accuracy of prediction.

Substitute observation vector $\mathbf{y}$ in each of four CFs and then specimen is assigned to group i for which;

$$L_i'(\mathbf{y}) > L_j'(\mathbf{y}) \qquad\qquad j \neq i \text{ where } i, j = 1, 2, 3, 4.$$

Let's first consider observation vector $\mathbf{y_4}$:

$$\begin{aligned}L_1'(\mathbf{y_4}) &= 59.50(6.54) + 4218.35(0.281346) + 1773.50(0.288043) - 180.00(0.09633)\\ &\quad +2809.01(0.480122) - 766.59(0.214067) + 1985.98(0.183486) - 1776.20\\ &= 1842.216164\\ L_2'(\mathbf{y_4}) &= 62.90(6.54) + 4502.39(0.281346) + 1875.84(0.288043) - 240.36(0.09633)\\ &\quad +2826.39(0.480122) - 794.79(0.214067) + 2174.09(0.183486) - 1939.43\\ &= 1841.622906\end{aligned}$$

By similar substitution of $\mathbf{y_4}$ into $L_3'(\mathbf{y})$ and $L_4'(\mathbf{y})$, we get $L_3'(\mathbf{y_4}) = 1841.846301$ and $L_4'(\mathbf{y_4}) = 1843.042745$. Maximum one is $L_4'(\mathbf{y_4})$ and that is the classification function corresponding to *Mahaweli* (R-4) river. Therefore, a fish of *Mahaweli* river is correctly classified into its real river by linear classification functions.

Let's consider observation vector $\mathbf{y_3}$:

$$\begin{aligned}
L_1'(\mathbf{y_3}) &= 59.50(6.14) + 4218.35(0.267101) + 1773.50(0.304878) - 180.00(0.081433) \\
&\quad +2809.01(0.477199) - 766.59(0.179153) + 1985.98(0.184039) - 1776.20 \\
&= 1810.516334 \\
L_2'(\mathbf{y_3}) &= 62.90(6.14) + 4502.39(0.267101) + 1875.84(0.304878) - 240.36(0.081433) \\
&\quad +2826.39(0.477199) - 794.79(0.179153) + 2174.09(0.184039) - 1939.43 \\
&= 1808.097322
\end{aligned}$$

By similar substitution of $\mathbf{y_3}$ into $L_3'(\mathbf{y})$ and $L_4'(\mathbf{y})$, we get $L_3'(\mathbf{y_3}) = 1808.687386$ and $L_4'(\mathbf{y_3}) = 1814.801304$. Maximum one is $L_4'(\mathbf{y_3})$ and that is the classification function corresponding to *Mahaweli* (R-4) river. Therefore, a fish of *Kelani* is misclassified into *Mahaweli* river. Using the above procedure, unknown place of collection (river) of fresh *Puntisu dorsalis* specimen can be found easily. Only four rivers were considered in developing the statistical rule and therefore a *P. dorsalis* caught from any of these four rivers (*Mahaweli, Kalu, Kelani* and *Nilwala*) could only be assigned correctly to its place of collection.

4 Conclusions

The overall percentage of correct classifications for both the analysis and hold-out samples, which are measure of predictive ability, shows that discriminant analysis can be used to predict unknown place of collection (river) of fish specimens. We can also conclude that morphometric characters of a fish can be used to predict its unknown place of collection. The prediction of place of collection of a fresh specimen using its posterior probability is a difficult task. Because, it is hard to calculate posterior probability of specimen belong to a particular group, given each of the discriminant scores. But, linear classification functions can be used to do such a prediction easily.

In non parametric classification, distributional assumptions are not necessary. Therefore, it motivates us to try techniques like *CART* and *Neural Networks* for further research on this area.

References

[1] Ackerman, J.T., Takekawa, J.Y., Bluso, J.D., Yee, J.L., and Eagles, C.A. (2008). Gender Identification of Caspian Terns using External Morphology and Discriminant Function Analysis, *The Wilson Journal of Ornithology*, **120(2)** : 378383.

[2] Angelopoulos, N., and Cussens, J. (2002). Exploiting Informative Priors for Bayesian Classification and Regression Trees. Department of Computer Science University of York Heslington, York YO10 5DD, UK.

[3] Atabati, M., Zarei, K., and Abdinasab, E. (2009). Classification and Regression Tree Analysis for Molecular Descriptor Selection and Binding Affinities Prediction of Imidazobenzodiazepines in Quantitative Structure-Activity Relationship Studies. Bull. Korean Chem. Soc, Vol. **30**, No. 11.

[4] Austin, C.M., and Knott, B. (1996). Systematics of the freshwater crayfish genus Cherax Erichson (Decapoda: Parastacidae) in south-western Australia: electrophoretic, morphological and habitat variation. *Australian Journal of Zoology*, **44**: 223-258.

[5] Bickel, P.J., and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli **10(6)**, 9891010.

[6] Bowering, W.R. (1988). An analysis of morphometric characters of Greenland halibut (Reinhardtius hippoglossoides) in the northwest Atlantic using a multivariate analysis of covariance. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**: 580-585.

[7] Deraniyagala, P.E.P. (1952). *A coloured atlas of some vertebrates from Ceylon*. Volume 1: *Fishes*. pp. 150 , 34pls. The Ceylon Government Press, Ceylon.

[8] De Silva, M.P.K.S.K., Liyanage, N.P.P., and Hettiarachi, S. (2006). Intra specific morphological plasticity in three *Puntius* species in Sri Lanka. *Ruhuna Journal of Science*, **1**: 82-95.

[9] De Silva, M.P.K.S.K., and Liyanage, N.P.P. (2010). A multivariate approach for developing a dichotomous key for identification and differentiation of Puntius (Osteichthyes: Cyprinidae) species in Sri Lanka, *Journal of National Science Foundation*, **38(1)**: 15-27.

[10] Erimafa, J.T., Iduseri, A., and Edokpa, I.W. (2009). Application of discriminant analysis to predict the class of degree for graduating students in a university system, *International Journal of Physical Sciences*, Vol. **4(1)**, pp. 016-021.

[11] Fernandez, G.C.J. (1999). Discriminant Analysis, A Powerful Classification Technique in Data Mining. Department of Applied Economics and Statistics / 204, University of Nevada - Reno, Reno NV 89557, Paper 247-27.

[12] Fernandez, G., (2002). Discriminant Analysis, a Powerful Classification Technique in Predictive Modeling. University of Nevada, Reno.

[13] George, D., and Mallery, P. (2007). SPSS for Windows Step- By- Step 13.0. 6^{th} edition, New Delhi : Dorling Kindersley, ISBN 81-317-0394-0.

[14] Hardle, W., and Simar, L. (2007). Applied Multivariate Statistical Analysis: Second Edition. Verlag Berlin Heidelberg.

[15] Hassan, S.M. (2007). Parameters Estimation of a Discriminant Function For Some Higher Institutes Graduates in Egypt. Thebes Higher Institute for Management and Information Technology, Al Haram, Sakkra Rd., Giza, Egypt.

[16] Jayaram, K.C. (1991). Revision of the Genus Puntius Hamilton from the Indian Region (Pisces: Cypriniformes, Cyprinidae, Cyprininae). *Records of Zoological Survey of India*, Occaional. Paper **135**: 1 -178.

[17] Johnson, R.A., and Wichern, D.C. (2002). Applied Multivariate Statistical Analysis. Fifth edition, Pearson Education Pet. Ltd, pp: 581-666, ISBN 81-7808-686-7.

[18] Lachenbruch, P.A., and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis: Technometrics, **10**: 111.

[19] Munro, I.S.R. (1955). *The marine and freshwater fishes of Ceylon*. pp. 349,56pls, Department of External Affairs, Canbara, Australia.

[20] Nogueira, A., De Oliveira, M.R., Salvador, P., Valadas, R., and Pacheco, A. (2009). Classification of Internet Users using Discriminant Analysis and Neural Networks. University of Aveiro/Institute of Telecommunications, Campus de Santiago, 3810-193 Aveiro, Portugal.

[21] Pethiyagoda, R. (1991). *Freshwater fishes of Sri Lanka*. pp 1-362, Wildlife Heritage Trust of Sri Lanka, Colombo.

[22] Raykov, T., and Marcoulides, G.A. (2008). Introduction to applied multivariate analysis. Taylor and Francis Group, LLC, ISBN-13: 978-0-8058-6375-8.

[23] Raudys, S.J., and Jain, A.K. (1991). Small Sample Size Effects in Statistical pattern Recognition: Recommendation for Practitioners. IEEE Transactions on pattern analysis and machine intelligence, Vol. **13**, No 3.

[24] Rencher, A.C. (1998). Multivariate Statistical Inference and Applications, New York: Wiley.

[25] Rencher, A.C. (2002). Methods of Multivariate Analysis. Second Edition. A John Wiley and Sons, Inc.Publication.

[26] Royston, P. (1995). A Remark on Algorithm AS 181: The W Test for Normality. Applied Statistics, **44**, 547-551.

[27] Sever, M., Lajovic, J., and Rajer, B. (2005). Robustness of the Fishers Discriminant Function to Skew-Curved Normal Distribution. Metodoloski zvezki, Vol. **2**, No. 2, 231-242.

[28] Stevens, J. (1996). Applied Multivariate Statistics for the Social Sciences. Third Edition. Lawrence Erlbaum Associate, Inc.

[29] Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Edu. Psychol. Measur. **55(4)**: 525-534.

[30] Tim, J.K., Andrew, A.B., and David, S.H. (2004). Comparison of discriminant function and classification tree analyses for age classification of marmots. OIKOS **105**: 575-587.

[31] Welch, B.L. (1939). Note on Discriminant Functions, Biometrika, **31**, 218220.

[32] Winnas, G.A. (1985). Using morphometric and meristic characters for identifying stocks of fish. In: *Proceedings of stock identification symposium*. (Eds. H. E.Kumf, R. N.Vaught , C. B.Grimes, A. G. Johnson and E. L. Nakamura) pp. 199-223, NOAA Technical Memorandum NMFS-SEFC.