**Title:** Stock delineation of Striped Snakehead, *Channa striata* using multivariate generalised linear models with otolith shape and chemistry data

**Authors:** Salman Khan1#, Hayden T. Schilling2,3#, Mohammad Afzal Khan1, Devendra Kumar Patel4, Ben Maslen, Kaish Miyan1

1. Section of Fishery Science and Aquaculture, Department of Zoology, Aligarh Muslim University, Aligarh -202 002, India
2. Centre for Marine Science & Innovation, UNSW Australia, Sydney 2052, Australia
3. Sydney Institute of Marine Science, Chowder Bay Road, Mosman 2088, Australia
4. Indian Institute of Toxicology Research, Council of Scientific and Industrial Research, Lucknow-226 001, India
5. Mark Wainwright Analytical Centre, UNSW Australia, Sydney 2052, Australia

# SK and HTS are joint first authors.

Corresponding author: Hayden T. Schilling, [h.schilling@unsw.edu.au](mailto:h.schilling@unsw.edu.au) , +614 02 9435 4600 (no fax)

**Abstract:**

Otoliths are commonly used to discriminate between fish stocks, through both elemental composition and otolith shape. Typical studies also have a large number of elemental compositions and shape measures relative the number of otolith samples, with these measures exhibiting strong mean-variance relationships. These properties make otolith composition and shape data highly suitable for use within a multivariate generalised linear model (MGLM) framework, yet MGLMs have historically never been applied to otolith data. Here we apply both a traditional distance based PERMANOVA and MGLMs to a case study of *Channa striata* (Striped Snakehead) in India. We also introduce the Tweedie distribution as a suitable error structure for the MGLM’s, drawing similarities in the properties of Biomass data. We demonstrate that otolith elemental data and combined otolith elemental and shape data violate the assumption of homogeneity of variance of PERMANOVA and may give misleading results, while the assumptions of the MGLM with Tweedie and gamma distribution are shown to be satisfied. and are appropriate for both otolith shape and elemental composition data. Consistent differences between 3 groups of *C. striata* were identified using both otolith shape, otolith chemistry and a combined otolith shape and chemistry dataset. This suggests that t future research should be conducted into whether there are demographic differences between these groups which may influence management considerations. The MGLM method is widely applicable and could be applied to any multivariate otolith shape or elemental composition dataset.

**Keywords:**

Otolith shape, otolith chemistry, multivariate generalised linear models, India, stock discrimination

1. **Introduction:**

Natural markers such as genetic, elemental or morphological markers can be used as tools to delineate populations or stocks, providing important information for fisheries management (Turner et al. 2015). Otoliths are a common tool used for stock discrimination and numerous studies have shown the potential of otoliths in addressing research problems related to successful fishery resource management (Tracey et al. 2006, Ferguson et al. 2011). Both otolith shape and elemental composition have become popular and successful tools in discriminating fish stocks (Campana and Casselman 1993; Begg et al. 2001; Miyan et al. 2016a; Nazir and Khan 2019).

Differences in the shape of otoliths can help to discriminate between groups of fish that are at least partly separated, inhabiting different environments (Bird et al., 1986; Campana and Casselman, 1993; Begg et al., 2001; Smith et al., 2002) Variations in otolith shape increase with the extent of genetic discreteness or geographic separation (Castonguay et al. 1991; Friedland and Reddin 1994), although disentangling the physiological and environmental influences is often complicated (Vignon and Morat 2010). Similarly, the elemental composition of otoliths can also be used to distinguish between fish populations (Campana et al. 2000; Longmore et al, 2011). Minor and trace elements laid down within the protein matrix become a permanent record of the chemical characteristics of the environment experienced by the fish (Elsdon and Gillanders 2003; Stransky et al. 2005). While both physiological and environmental factors influence the elemental composition of otoliths (Grammer et al 2017, Izzo et al 2018), if fish inhabit different water masses or environments for a certain period of time they can be differentiated via the elemental composition of their otoliths (Elsdon and Gillanders 2004; Khan et al. 2012; Miyan et al., 2014; 2016a; 2016b). By combining both otolith shape and chemistry data in the same analysis the ability to differentiate groups of fish can sometimes be improved (Fowler et al., 2015).

Both otolith shape and otolith chemistry data are usually multivariate with hypothesis testing traditionally conducted using distance-based methods (eg. PERMANOVA; Ferguson et al. 2011, Schilling et al. 2018, Biolé et al. 2019) or model-based methods which assume a gaussian error distribution (eg. MANOVA or LDA; Maguffee et al 2019). Ecologists typically also use these distance-based methods to form ordination plots to visualise the multivariate groupings in a low-dimensional plot (e.g. nMDS plots or Canonical Discriminant Aanalysis; Tanner et al. 2012, Andrade et al. 2020). The issue however with taking these approaches is that they assume homogeneity, with no mean-variance relationship being taken into account in both the hypothesis testing or visualisation techniques. This is concerning for the otolith shape and chemistry data which have strong mean variance relationships, where the variance increases with the mean concentration and shape parameter value (e.g. Figure 2). The otolith data has a natural boundary at zero which creates a mean-variance relationship as observations found away from this boundary become more variable. Particularly concerning is that both the otolith chemistry and shape data have very small values particularly close to this boundary, with majority of observations being less than 1 making this mean-variance relationship quite strong. A recent study by Warton et al. 2018 found that abundances with means less than 1 cannot reasonably be expected to have their variances stabilised, even with a well-chosen transformation due to the strength of this mean-variance trend. Instead, this trend should be explicitly modelled in the testing and visualisation procedure.

Otolith shape data is positive and continuous and as such can be appropriately modelled using the Gamma distribution (traditionally with a log link) which assumes that the variance increases proportionally to the mean squared. If is Gamma distributed then , where and are shape and scale parameters respectively. Otolith chemistry data however is often more nuanced, with a proportion of null observations where the detection of a chemical is below the limit of detection and therefore unable to be quantified in the otoliths, as well as a distribution of positive continuous observations for the otoliths which do have the chemical present. The positive continuous observations will have a mean variance relationship similar to the shape data (e.g. Figure 2), however a Gamma model will not suffice here as it assumes positive continuous data and therefore will not model the null counts. Ecologists have also used transforms for similar data to avoid the logging of null counts, however this has the same issue outlined in Warton et al. 2018 where the transformation isn’t handling the mean-variance relationship properly and it also isn’t modelling the null counts in a meaningful way, just lumping them all in as log(1). So, a model is required that takes into account both the large number of null observations as well as the mean variance relationship exhibited in the present observations.

A solution to this problem lies with the methods currently used to deal with Biomass data. Biomass data has very similar properties to the otolith chemistry data, having a large number of null observations where the species was not found to be present and a distribution of positive continuous weight samples for the species that are found to be present. The solution to modelling the Biomass data and consequently the otolith Chemistry data is the Tweedie Distribution. The Tweedie distributions suitability to biomass data is explained in detail in Foster et. al 2013, however is largely due it’s equivalence to summing a Poisson number of gamma random variables. This allows the null observations to be modelled with the Poisson component and the positive continuous observations with the gamma component. The Tweedie distribution also has a flexible mean-variance relationship. If are Tweedie distributed then , where is a power parameter that controls the shape of the distribution and (in the context of our study) is a chemical specific dispersion parameter. The mean-variance relationship is therefore defined by Taylor’s power law (Taylor 1961), which has been shown to arise under a variety of ecological processes (Kendal et al. 2004). The use of the Tweedie distribution in an MGLM setting is also discussed in detail in Niku et. al. 2017.

For ecological studies using multivariate abundance data such as species abundances, multivariate generalised linear models (MGLMs) are becoming more popular as they allow increased certainty and interpretability of the results, flexibility, and efficiency compared to distance-based methods (Warton et al 2015). While MGLMs are now common for abundance data (Wang et al 2012), they are rarely used for other datasets despite the flexibility of the method which allows users to specify model parameters to fit a dataset. Otolith Chemistry data and Shape data can be easily applied in an MGLM setting (Niku et. al. 2017, Wang et al. 2012) for instance by specifying appropriate mean-variance relationships and error distributions for the data (Wang et al. 2012). Appropriately modelling the mean-variance relationship of data avoids misleading results that can that arise in the traditional approaches when their homogeneity assumptions are not met (Warton et al. 2012). Assumptions for these models can also be readily checked by plotting Dunn-Smyth residuals (Dunn and Smyth 1996) which are randomised quantile residuals that have been shown to be effective at detecting many forms of model misspecification for generalised linear models (Dunn and Smyth 2018). MGLM’s have also been found to have higher power then distance-based methods such as PERMANOVA (Warton et al. 2012). Visualisations can also be performed in an MGLM setting using a latent variable factor analysis (Hui 2016) or by taking a copula approach (Popovic et al. 2019) that do not mislead users into mistaking differences in variance with differences in the mean.

Another recent technique used to analyse otolith data are machine learning classification methods, that have become prominent as they are robust to many assumptions that are often hard for traditional methods to satisfy (Jones et al., 2017). These methods allow the data to be grouped into different classes, which users can align as their ‘population’ markers. This method however fails to provide a means for hypothesis testing nor to easily visualise the differences among the groups.

*Channa striata*, locally known in India as “Dharidar-Sol” or “striped snakehead”, is commercially important in food, ornamental and sport fisheries along with other species of the family Channidae. *C. striata* is one of the main food fishes in Asian countries including India. In the last few years, due to increasing anthropogenic activities, unstrained harvesting and habitat alterations, the natural stocks of the fish have decreased severely (Rahman and Awal, 2016). Consequently, feeding and natural breeding grounds of this economically important fish species have been reduced, which has caused a shrinkage in wild populations (Rahman and Awal, 2016). Recent work has shown variation in body morphometrics of *C. striata* between 3 sites within India which suggests the potential for sub-population level variation in demographics which should be further investigated (Khan et al 2019). The present study was carried out with the dual aim of firstly, assessing variation in otolith chemistry and shape between the same groups *C. striata* in India to test for further evidence of regional separation, and secondly, demonstrating the use of MGLMs with otolith chemistry and otolith shape data.

1. **Methods and Materials**

*2.1 Study species, region, and sample collection*

The striped snakehead, *Channa striata* is native to east and southeast Asia. It is found in India, Pakistan, southern Nepal, Sri Lanka, Bhutan, southern China, Bangladesh, and all the countries of southeast Asia. It is also native to the major western islands of the Malay Archipelago, including Sumatra, Borneo and Java. The species has been introduced to the Philippines, eastern islands of Indonesia, New Caledonia, New Guinea, Fiji, south-eastern Russia and South Korea (Phen et al., 2004; War et al., 2011). *C. striata* can be found in many types of slow-moving freshwater habitat, including rivers, ponds, lakes, creeks, canals, flooded rice paddies, swamps, and irrigation reservoirs (Cagauan 2007).

Eighteen *C. striata* were collected from each of three locations. Each site was located on a different major river in northern India with fish collected regularly from each site between October 2017 and November 2018 using cast nets (25mm mesh) and drag nets (28mm mesh). The three locations were Narora (27° 30' N; 78° 25' E) on the river Ganga, Agra (27.1767° N; 78.0081°E) on the river Yamuna and Lucknow (26° 55' N; 80° 59' E) on the river Gomti (Figure 1). Identification of the fish was based on the descriptions of Jayaram (1999); Talwar and Jhingaran (1991). Total length was measured to the nearest mm. Otoliths were extracted using forceps, cleaned in fresh water and stored dry before subsequent shape and chemical analysis. Full details of fish used in this study can be seen in Table S1.

All methods were carried out in accordance with the relevant guidelines and regulations. The target fish species is a commercially exploited common food fish in India; therefore the Committee for the Purpose of Control and Supervision of Experiments on Animals (CPCSEA) 2018, Ministry of Environment, Forests and Climate Change, Government of India, does not require ethical approval to be given for this study.

*2.2 Otolith shape*

The shape of the otoliths was quantified using wavelet coefficients using R v3.6.0 (R Core Team, 2019). The R package *‘shapeR*’ was used to calculate both Normalized Elliptic Fourier and the discrete wavelet coefficients using photographs of each otolith which create mathematical representations of the otolith outlines. All otoliths were photographed using a light microscope and reflected light with the otolith placed with distal surface up on a black background. The procedure followed is fully detailed in Libungan and Pálsson (2015) although some photos of otoliths needed manual editing to accurately capture the otolith outlines. Once the photos were captured the outlines of the otoliths were smoothed to remove high frequency pixel noise around the otolith outlines using the *smoothout()* function with 100 iterations. The wavelet method then fitted a series of approximating functions within restricted domains to quantify the outline shapes (Graps 1995). The elliptical Fourier method by contrast fitted a number of harmonic functions to capture crenulations and lobes on the edges of the otoliths (Tracey et al. 2006). Both methods result in coefficients which can be used to quantify the shape. Using 10 wavelets (63 wavelet coefficients), >99 % of otolith shape was explained as opposed to the elliptical Fourier transformed coefficients which were only able to reproduce 95% of the shape (Fourier transformed results not shown) and we therefore proceeded only with the wavelet analysis.

To visualise the difference in mean shape between the three sites, the mean shape was reconstructed using the mean wavelets for each site. Wavelet coefficients were standardised for fish length as per Libungan and Pálsson (2015) before analysis to test for differences between the three sites.

*2.3 Otolith chemistry*

To remove any surface contamination, otoliths were soaked in 3 % hydrogen peroxide for 5 min and immersed for 5 min in 1 % HNO3. Otoliths were then flooded with ultra-pure water for 5 min to remove the acid. After decontamination, the otoliths were dried under a laminar flow hood and weighed to the nearest 0.1 mg (Turan 2006; Khan et al. 2012). For analysis, the decontaminated otoliths were dissolved in 10 ml of 37 % HNO3 and the volume was brought up to 25 ml with Milli Q water. Elemental composition of whole otoliths were analysed using inductively coupled plasma atomic emission spectrometry (ICP–AES; Thermo Electron IRIS Intrepid II XSP DUO). Blank samples were used to correct for background noise in readings. The elements (and detection limits in ppm) measured from the otoliths included: Ca (0.005), Na (0.05), Mg (0.0005), Sr (0.0005), Ba (0.0005), Mn (0.001), Fe (0.005), Pb (0.05), Ni (0.005), Zn (0.005), Cd (0.005), Cr (0.005) and K (0.1). All elements were above minimum detection levels except for 4 Zn samples from the Agra site and 7 Cd samples from the Lucknow site. Internal standards Indium (In) and Gallium (Ga) were added in samples and blanks, which were used to correct for the remaining matrix effect and to compensate for instrument drift. Multi elemental standards were prepared with high purity ICP multi-element standard solution IV certiPUR (NIST SRM) obtained from Merck (Germany) using Milli-Q water and analytical grade 2% v/v HNO3 for external calibration. Standards were run every 10 samples. A calibration blank was also prepared in the same procedure. The calibration curve was obtained for five points. The concentration of elements in the sample and blank were calculated and expressed as µg g-1 (ppm) on dry weight basis (Turan, 2006; Miyan et al., 2016b). All elemental concentrations were converted from ppm to ratios of element:Calcium (mmol:mol) to control for the size of each analysed otolith.

*2.4 Statistics*

All analysis was performed in R v3.6.0 (R Core Team, 2019). As the initial goal of the paper was to demonstrate the applicability of Multivariate generalised linear models to otolith chemistry and shape data it is necessary to also show the results of a standard analysis method as a baseline. We choose to use a common distance-based analysis, PERMANOVA with a nMDS ordination plot to visualise the differences. Using the vegan R package, we created a dissimilarity matrix using Euclidean distances as is common for otolith datasets using the ‘*vegdist()*’ function (Schilling et al. 2018). Distance based analyses such as PERMANOVA have an often untested assumption of homogeneity of variance between groups. If this assumption is violated, then results can become misleading with inflated standard errors and confidence intervals leading to a possible increase to the Type 1 error rate. In fact, a google scholar search of “otolith PERMANOVA” for 2018 and 2019 revealed less than 10% of papers check this assumption.

To check this assumption in a PERMANOVA setting we can perform a dispersion test using the ‘*betadisper()*’ function where a significant result (P < 0.05) indicates an unequal variance between groups and therefore a violation of the assumptions of PERMANOVA. If this assumption is satisfied, the typical approach will be to proceed with the PERMANOVA for multivariate differences between our three sites using the ‘*adonis()*’ function. We can then also create an nMDS ordination plot using the ’*isoMDS()*’ function from the MASS R package(Venables & Ripley 2002) based upon the earlier created distance matrix. However, if however the homogeneity assumption is not satisfied, then PERMANOVA would not be recommended to be used for the analysis and we would be unable to do hypothesis testing without getting misleading results.

For the model-based MGLM analysis we followed the analysis guidelines provided in Warton et al. (2015), following a defined modelling process. We first identified our question: Are there differences in otolith chemistry or otolith shape between the three groups of *C. striata*? Secondly, we considered our data. We had only one predictor variable, Site (a categorical variable) and many response variables (all the elemental concentrations and shape coefficients). Thirdly, we conducted exploratory data analysis but as we only had a single categorical predictor variable this was limited. Next, we selected an appropriate model for the question. Our goal was to compare means between three groups using multivariate data and our *a priori* hypothesis was that there will be multivariate differences between the three sites. Both the otolith chemistry and shape data are positive continuous data but otolith chemistry can contain zeros when elements are below the levels of detection, therefore, Tweedie error distributions were considered as the most appropriate fit for our otolith data while gamma error distributions were considered most appropriate for the shape data. We therefore used multivariate generalised linear models (MGLMs) with a Tweedie error distributions (variance power 1.75, log-link) to test for our hypothesis with chemistry data and a gamma error distribution with log-link for the otolith shape data (coded as a Tweedie distribution with variance power 2 which is equivalent to a gamma distribution). For the combined shape and chemistry data we individually specified error distributions for each variable (Tweedie for the chemistry data and gamma for the shape data). When using multivariate models it is important to understand the relationship between the mean of each response variable and the observed variance (Warton, 2008, Warton et al., 2012). To investigate this relationship in our data, we created mean variance plots which show how the variance changes with the mean of each variable. The mean variance plots identified that for both chemistry and shape data, as the mean increased, the variance also increased (Figure 2). As a final step prior to inspecting the results, we assessed our models. To assess if the MGLMs accurately captured the properties of our data, Dunn-Smyth residual plots were inspected for each model. No strong pattern were visible and the models were deemed to be accurately representing our data (Figure 3), allowing the use of these models to address our hypothesis. All MGLM models were run using the *‘manyany()’* function in the mvabund R package (Wang et al., 2012).

To compare the effectiveness of otolith chemistry and otolith shape data in discriminating the three sites, three MGLMs were run. One only used otolith chemistry data, one only used otolith shape data and one combined both chemistry and shape data. For the two MGLMs involving the elemental data, univariate generalised linear models (GLMs) were also ran for each variable to identify which variables were driving the differences. This was conducted within the ‘*manyany()*’ function. The influence of each variable in driving the differences (similar objective to a distance-based SIMPER analysis) was quantified using the individual contribution to the Sum-of-LR (Warton et al. 2012), whereby variables with a larger likelihood ratio value are more influential. For the GLMs which included shape data there is no meaningful interpretation of the univariate GLMs as the wavelet shape coefficients cannot be interpreted individually but it does allow the relative contributions of otolith chemistry and shape to be assessed in the combined model. Posthoc tests to identify which sites had showed evidence of differences in specific otolith elemental concentrations were run manually using two sites at a time using the same *‘manyany()’* GLM process and adjusting the *P*-values using the Bonferroni method with the ‘*padjust()*’ funtion. To visualise the multivariate differences between the 3 fish groups (as an alternative to the commonly applied distance-based ordinations), two factor model-based latent variable ordinations were produced using the *boral* R package (Hui 2016), again using Tweedie error distributions for chemistry data and gamma error distributions for the shape data with the assumptions being visually assessed (Hui et al. 2015). The *boral* package takes generalised linear models for each response variable, using Bayesian Markov chain Monte Carlo methods to estimate latent variables that account for between response correlation, which can then be used to visualise multivariate differences on a low-dimension plot (Hui 2016). By using generalised linear models, this approach can align the visualisation model with the testing model, check assumptions and specify mean-variance relationships. The code and data used in these analysis is available at: <https://github.com/HaydenSchilling/MGLMs-Otoliths>

1. **Results**

*3.1 Distance-based Assumptions and Analysis*

The dispersion test of equal variance between the samples from the three sites showed that there were significant differences in variance between sites for both the otolith chemistry data (*F*2,51 = 9.409, P < 0.001) and the combined chemistry and shape data (*F*2,51 = 9.277, P < 0.001). The assumption of equal variance was therefore only satisfied for the shape only dataset (*F*2,51 = 0.420, P = 0.659).

For the purpose of our demonstration we proceeded with all three sets of analyses (elemental data, shape data and combined elemental and shape data) but due to the assumption violations caused by the unequal variance between sites, only the shape data analysis should be considered reliable.

Using the otolith shape data, the PERMANOVA showed strong evidence of differences between the three sites (*F*2,53 = 6.06, P < 0.001, Figure 4). Visualising the multivariate differences in the NMDS ordination reveals some separation between sites, driven by the Narora site while the Agra and Lucknow samples have considerable overlap (Figure 5, stress = 0.12).

Using the otolith chemistry data, PERMANOVA showed strong differences between sites (*F*2,53 = 513.15, P < 0.001, Figure 6). Visualisation using the NMDS ordination showed large separation between the Agra site and the other sites along the NMDS 1 axis (Figure 5, stress = 0.01). The Narora site was heavily dispersed along the NMDS 2 axis with some samples overlapping the Lucknow site.

When combined, the PERMANOVA again showed clear differences between sites (*F*2,53 = 511.76, P < 0.001). The separation in the NMDS ordinations was clearly driven by the differences in otolith chemistry with an almost identical pattern observed (Figure 5, stress = 0.01).

An important point to note in the NMDS plots for the Chemistry only and combined visualisations (Figure 5) is that the Lucknow group is seen to have points much closer together than the other sites. One could interpret this observation by stating that samples from the Lucknow group are ‘less variable’ then samples from other sites and thus should be easier to distinguish if they didn’t overlap with more variable samples from Norora. This observation however is incredibly misleading and is a prime example of the dangers in ordination techniques that do not take into account mean-variance relationships (being explained in detail in Warton et. al. (2012)). Samples from Lucknow had the lowest mean concentrations of all chemicals besides Magnesium (Figure 6). This difference in ‘variance’ we are seeing in the NMDS plots for the Lucknow samples is in reality just a difference in mean concentration, with the difference in variance arising from the mean-variance relationship this data has (where the variance increases with the mean). Without properly accounting for this relationship users can inflate differences in mean concentrations for differences in concentration variability (confounding location and dispersion effects). Conversely, if we look at the model based latent variable ordinations (Figure 7) that do take into account a mean-variance relationship we not only get ‘greater power’ to pick apart the different populations, but we also have samples from the Lucknow site no longer being depicted with small variability, instead being similar in variability to samples from Norora, removing this previously misleading result.

*3.2 MGLM Analyses*

Using the wavelet coefficients, the MGLM analysis showed clear difference in otolith shape between all three sites (*LR*: 22.368, *P* < 0.001; Figure 4).

Otolith chemistry was also clearly different between the three sites (*LR* = 1147.9, *P* < 0.001; Figure 6; Table 1). Large differences in mean concentration were observed for many elements with the Agra site having the highest concentrations of 10 of the 12 tested elements (Figure 6). The Narora site had the highest concentrations of the other two elements (Zinc & Magnesium; Figure 6). The Lucknow site showed the lowest concentrations of all elements (Figure 6).

The combined analysis of otolith chemistry and shape also revealed clear differences between all three sites (*LR* = 1166.2, *P* < 0.001). Within this combined analysis most of the differences were driven by the chemistry data (98.4% of the *LR* ratio was made up by the element data).

The differences identified by the MGLMs between sites were visible in the latent variable ordinations (Figure 7). Similar patterns were visible to those identified in the multivariate generalised linear models with larger differences evident in the otolith chemistry data (Figure 7a) than the otolith shape data (Figure 7b). When both datasets were combined, the sites were the most tightly grouped and distinct (Figure 7c).

1. **Discussion**

This study demonstrated how multivariate generalised linear models (MGLMs) can be applied to otolith chemistry and otolith shape data to test for differences between groups of samples. We showed that distance-based analyses including PERMANOVA are not appropriate for our otolith chemistry data due to violations of the assumption of homogeneity of variance stemming from a non-linear mean variance relationship in the data. This mean-variance relationship can be directly modelled with MGLMs which we then use to show that *Channa striata* from three sites have clear differences in both otolith chemistry and shape. The MGLM method applied here to a simple test between three groups could be easily adapted and expanded to answer other ecological questions requiring more complex model frameworks as is currently done in the broader field of ecology.

*4.1 The MGLM method for otolith data*

This study has demonstrated the potential for MGLMs to be used as an analysis tool for otolith chemistry and/or otolith shape data, for example in fisheries stock discrimination. We successfully applied a model-based multivariate analysis method to a case study in India and identified differences in otolith chemistry data and otolith shape data for *C. striata* collected from three sites. The MGLM framework which we have used can be considered a robust alternate to the more widely used distance-based analyses including permutational ANOVAs (PERMANOVAs). We demonstrated that in some instances such as our example, distance-based analyses are not appropriate due to violations of the assumption of homogeneity of variance. The advantages for using GLMs over distance based methods are well documented in Warton et al. (2012), but briefly we describe the biggest advantages of applying MGLMs to otolith data as well as a potential disadvantage below.

A major advantage of this method is that MGLMs are flexible, being able to specify mean variance assumptions and error distribution that are appropriate to the data, avoiding misleading results from models that do not properly take these relations into account. These assumptions can be easily checked (and models altered if required) and the appropriateness of the models assessed before any inference is made from the results. We demonstrated this using mean-variance and Dunn-Smyth residual plots in our case study where we demonstrated that the MGLM with a Tweedie or gamma error distributions were an appropriate fit to the otolith chemistry and otolith shape data, thus accounting for the non-linear mean-variance relationship (Figures 2 & 3). Not only do MGLM’s help to avoid misleading results but they have also been shown to have greater power at detecting effects when compared to traditional distance based approaches (Warton et al. 2012). Warton et al. 2012 also found that mean-variance misspecification can lead to the confounding of dispersion and location effects in ordination plots (which we have verified in this study), misleading results when attempting to identify which response the effect is expressed in and a failure to detect multivariate effects unless it expressed in a high variance response. We have also demonstrated the flexibility of MGLMs with our combined shape and chemistry analysis which used different error distributions for the two datasets which ensures both datasets are treated appropriately in the same analysis.

The main downside of using this approach is that computational time can be longer when there are a large number of variables with a Tweedie error distribution. This could be a potential problem for shape data as there are often many coefficients which are used as variables but with the gamma distribution this is not a concern. Our examples with shape data (gamma error distribution) took only 4 min while the chemistry data (Tweedie error distribution) took 40min and the combined analysis (combining both Tweedie and gamma error distributions) took 44 min using a single core (8 gb RAM). While these calculations can be run on regular computers the time factor is a trade-off which individual researchers will need to consider, particularly if they do not have access to large computing resources, although with advances in computing software and technology this is likely to become faster and more accessible.

The latent variable model-based ordinations successfully visualised the multivariate differences identified in the MGLMs. The ordinations visually matched the model results with the elemental data clearly driving the separation but the overall separation improving when shape data was combined with the elemental data. While the current study used a Bayesian model based latent variable method (Hui et al. 2015), an alternative ordination method directly based upon the MGLM model could be produced using Gaussian copula graphical models which can be run using the ‘ecoCopula’ R package (Popovic et al., 2019). Both these ordination methods provide an alternative to traditional based distance-based ordination methods which we have shown to be misleading by failing to account for mean-variance relations. By following the code provided with this paper, the MGLMs and model-based ordination methods can easily be applied in future studies.

*4.2 Implications for C. striata in India*

Both otolith elemental composition and shape data showed differences between the three sampling sites. Otolith chemistry showed the largest differences while the differences in shape were significant but less clear. The distinct otolith chemistry and shapes suggests that *C. striata* in these three rivers are not regularly mixing. This confirms recent research which used truss morphometry based upon body shape of *C. striata* to suggest that the same three groups analysed in the current study may be distinct sub-populations (Khan et al. 2019). Further research should next examine key demographic dynamics at each of these three sites including growth rates and age of maturity. If the demographics at each site also differ then management changes may be required (Begg et al. 1999).

The unusually high concentrations of some elements in the otoliths likely reflects a heavily polluted environment as it is known that in India there continues to be concerns around pollution of waterways (Sengupta 2006). The Yamuna river is very polluted due to many cities lying on its bank and pouring sewage and other industrial effluents directly into the river. For this reason, the Yamuna river is recognized as one of the most polluted in the world (Bhardwaj et al., 2017). Our fish from the Agra site were located on the Yamuna river and their otoliths are reflective of the heavily polluted state with high concentrations of many elements, particularly heavy metals. It should be noted that fish at the Agra site were also bigger than the other sites (Table S1) but as we used whole otolith elemental composition and controlled for length in the shape analysis, the comparison of differences remains valid as there were very large differences between all three sites, particularly in the elemental composition of the otoliths. There were variations many elements which contributed to the multivariate differences discussed in the current paper and the drivers behind the specific elemental differences, whether natural or potential pollution present the opportunity for future study.

*4.3 Conclusion*

This study has successfully demonstrated the use of the Tweedie and gamma error distributions and by extension multivariate generalised linear models with otolith data by identifying differences between three sites in India based upon *C. striata* otolith chemistry and otolith shape data. These results suggest that further research into potential demographic differences in now necessary which then call for the recognition of different stocks of *C. striata*. The MGLM method (and code provided with this paper) is highly flexible and has the potential to be applied to many ecological questions using multivariate otolith data.

**Author Contributions**

SK, HS & KM conceived the idea, SK, MK, DP & KM collected the data, HS analysed the data, SK, HS & BM wrote the manuscript, MK, DP & KM critically reviewed the manuscript. BM provided statistical expertise for the paper. All authors approved publication.

**Funding**

The stipend for HS was partially funded by a UNSW Faculty of Science writing scholarship, a NSW Research Attraction and Acceleration Program grant to the Sydney Institute of Marine Science, the UNSW Research Infrastructure Scheme Network Lab for Ocean Collaboration and an Australian Research Council Linkage Project (LP150100923). SK is grateful to the Council of Scientific and Industrial Research, New Delhi, for funding the study and providing financial assistance to the first author as Senior Research Fellowship (ACK No. 113724/2K18/1).

**Acknowledgments**

This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney. This is contribution no. XXX of the Sydney Institute of Marine Science.

**References**

Andrade, H., van der Sleen, P., Black, B. A., Godiksen, J. A., Locke, W. L., Carroll, M. L., ... & Geffen, A. (2020). Ontogenetic movements of cod in Arctic fjords and the Barents Sea as revealed by otolith microchemistry. *Polar Biology*, *43*(5), 409-421.

Begg, G.A., Friedland, K.D., Pearce, J.B., 1999. Stock identification and its role in stock assessment and fisheries management: an overview, *Fish. Res.* 43, 1–8.

Begg, G.A., Overholtz, W.J., Munroe, N. J., 2001. The use of internal otolith morphometrics for identification of haddock (*Melanogrammus aeglefinus*) stocks on Georges Bank. *Fish. Bull.* 99, 1–14.

Bhardwaj, R., Gupta, A., & Garg, J. K. 2017. Evaluation of heavy metal contamination using environmetrics and indexing approach for River Yamuna, Delhi stretch, India. *Water Sscience*, *31*(1), 52-66.

Biolé, F. G., Thompson, G. A., Vargas, C. V., Leisen, M., Barra, F., Volpedo, A. V., & Avigliano, E. (2019). Fish stocks of Urophycis brasiliensis revealed by otolith fingerprint and shape in the Southwestern Atlantic Ocean. *Estuarine, Coastal and Shelf Science*, *229*, 106406.

Bird, J. L.; Eppler, D. T.; and Checkley, D. M., Jr., 1986: Comparisons of herring otoliths using Fourier series shape analysis. *Can. J. Fish. Aquat. Sci.* 43, 1228– 1234.

Cagauan, A. 2007. Exotic aquatic species introduction in the Philippines for aquaculture- A threat to biodiversity or a boon to the economy?. *Journal of Environmental Science and Management*, 10/1: 48-62.

Campana, S.E., and Casselman, J.M., 1993. Stock discrimination using otolith shape analysis. *Can. J. Fish. Aquat. Sci.* 50, 1062-1083.

Campana, S. E., Chouinard, G. A., Hanson, J. M., Frechet, A., & Brattey, J. (2000). Otolith elemental fingerprints as biological tracers of fish stocks. *Fisheries Research*, *46*(1-3), 343-357.

Castonguay, M., Simard, P., Gagnon, P., 1991. Usefulness of Fourier analysis of otolith shape for Atlantic mackerel (*Scomber scombrus*) stock discrimination. *Can. J. Fish. Aquat. Sci.* 48, 296–302.

Dunn, P. K., & Smyth, G. K. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236-244.

Dunn, P. K., & Smyth, G. K. 2018. Generalized Linear Models: Diagnostics. In *Generalized Linear Models With Examples in R* (pp. 297-331). Springer, New York, NY.

Elsdon, T. S., and B. M. Gillanders. 2003. Reconstructing migratory patterns of fish based on environmental influences on otolith chemistry. *Rev. Fish Biol. Fish.* 13: 219– 235. doi:10.1023/B:RFBF.0000033071.73952.40

Elsdon TS, Gillanders BM 2004. Fish otolith chemistry influenced by exposure to multiple environmental variables. *J Exp Mar Biol Ecol* 313:269–284

Foster, S.D. and Bravington, M.V., 2013. A Poisson–Gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics*, *20*(4), pp.533-552.

Fowler, A. M., P. I. Macreadie, D. P. Bishop, and D. J. Booth. 2015. Using otolith microchemistry and shape to assess the habitat value of oil structures for reef fish. *Marine Environmental Research* 106:103-113.

Friedland, K. D., and D. G. Reddin. 1994. Use of otolith morphology in stock discriminations of Atlantic salmon (*Salmo salar*). *Can. J. Fish. Aquat. Sci.* 51:91–98.

Ferguson, G. J., Ward, T. M., & Gillanders, B. M. (2011). Otolith shape and elemental composition: Complementary tools for stock discrimination of mulloway (*Argyrosomus japonicus*) in southern Australia. *Fisheries Research*, *110*(1), 75-83.

Grammer, G. L., J. R. Morrongiello, C. Izzo, P. J. Hawthorne, J. F. Middleton, and B. M. Gillanders. 2017. Coupling biogeochemical tracers with fish growth reveals physiological and environmental controls on otolith chemistry. *Ecological Monographs* 87: 487– 507.

Graps, A. 1995. An introduction to wavelets. *IEEE computational science and engineering*, 2(2), 50-61.

Hui, F. K. C. 2016. boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution* 7:744-750.

Hui, F. K. C., S. Taskinen, S. Pledger, S. D. Foster, and D. I. Warton. 2015. Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* 6:399-411.

Izzo, C., Reis-Santos, P., and Gillanders, B. M. 2018. Otolith chemistry does not just reflect environmental conditions: a meta-analytic evaluation. *Fish and Fisheries* 19, 441–454. doi:10.1111/FAF.12264

Jayaram, K.C., 1999. The Freshwater Fishes of the Indian Region. Narendra Publishing House, Delhi, India.

Jones, C. M., Palmer, M., & Schaffler, J. J. 2017. Beyond Z ar: the use and abuse of classification statistics for otolith chemistry. *Journal of Fish Biology*, *90*(2), 492-504.

Kendal, R. L., Coolen, I., & Laland, K. N. (2004). The role of conformity in foraging when personal and social information conflict. *Behavioral Ecology*, *15*(2), 269-277.

Khan, M.A., Miyan, K., Khan, S., Patel, D.K., Ansari, N.G., 2012. Studies on elemental profile of otoliths and truss network analysis for stock discrimination of the threatened stinging catfish, *Heteropneustes fossilis* (Bloch 1794), from the river Ganga and its tributaries. *Zool. Stud.* 51, 1195‒1206.

Khan, M.A., Khan, S., Miyan, K., 2019. Stock identification of the *Channa striata* inhabiting the Gangetic river system using truss morphometry. *Russ. J. Ecol.* 50, 391–396.

Libungan, L. A., and S. Pálsson. 2015. ShapeR: an R package to study otolith shape variation among fish populations. *PLOS ONE* 10:e0121102.

Maguffee, A. C., Reilly, R., Clark, R., & Jones, M. L. 2019. Examining the potential of otolith chemistry to determine natal origins of wild Lake Michigan Chinook salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(11), 2035-2044.

Miyan, K., Khan, M.A., Khan, S., 2014. Stock structure delineation using variation in otolith chemistry of snakehead, *Channa punctata* (Bloch, 1793), from three Indian rivers. *J. Appl. Ichthyol.* 30, 881‒886.

Miyan, K., Khan, M.A., Patel, D.K., Khan, S., Ansari, N.G., 2016a. Truss morphometry and otolith microchemistry reveal stock discrimination in *Clarias batrachus* (Linnaeus, 1758) inhabiting the Gangetic river system*. Fish. Res.* 173, 294‒302.

Miyan, K., Khan, M.A., Patel, D.K., Khan, S., Prasad, S., 2016b. Otolith fingerprints reveal stock discrimination of *Sperata seenghala* inhabiting the Gangetic river system. *Ichthyol. Res.* 63, 294‒301.

Nazir, A., Khan, M.A., 2019. Spatial and temporal variation in otolith chemistry and its relationship with water chemistry: Stock discrimination of *Sperata aor*. *Ecol. Freshw. Fish.* 28, 499–511.

Niku, J., Warton, D.I., Hui, F.K. and Taskinen, S., 2017. Generalized linear latent variable models for multivariate count and biomass data in ecology.*Journal of Agricultural, Biological and Environmental Statistics*, *22*(4), pp.498-522.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & Wagner, H. 2019. vegan: Community Ecology Package. R package version 2.5-5.

Phen, C., T. Thang, E. Baran, L. Vann. 2004. Biological Reviews of Important Cambodian Fish Species, Based on Fishbase. Penang, Malaysia: WorldFish Center.

Popovic, G. C., D. I. Warton, F. J. Thomson, F. K. C. Hui, and A. T. Moles. 2019. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution* 10:1571-1583.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M.A., Awal, S., 2016. Development of Captive Breeding, Seed Production and Culture Techniques of Snakehead Fish for Species Conservation and Sustainable Aquaculture. *Int. J. Adv. Agric. Environ. Eng*. 3, 117-120.

Schilling, H. T., P. Reis-Santos, J. M. Hughes, J. A. Smith, J. D. Everett, J. Stewart, B. M. Gillanders, and I. M. Suthers. 2018. Evaluating estuarine nursery use and life history patterns of *Pomatomus saltatrix* in eastern Australia. *Marine Ecology Progress Series* 598:187-199.

Sengupta, B., 2006. Water quality status of Yamuna River (1999–2005). Assessment and Development of River Basin Series: ADSORBS/41/2006-07. CPCB Ministry of Environment and Forests, Delhi. 136 pp.

Stransky, C., 2005. Geographic variation of golden redfish (Sebastes marinus) and deep-sea redfish (S. mentella) in the North Atlantic based on otolith shape analysis. *ICES J. Mar. Sci.* 62, 1691–1698.

Talwar, P. K., Jhingran, A.G., 1991. Insland Fishes of India and Adjacent Countries. Vol. 1-2. Oxford and IBH Publishing Co. Pvt. Ltd., New Delhi, Calcutta, India.

Tanner, S. E., Vasconcelos, R. P., Cabral, H. N., & Thorrold, S. R. (2012). Testing an otolith geochemistry approach to determine population structure and movements of European hake in the northeast Atlantic Ocean and Mediterranean Sea. *Fisheries Research*, *125*, 198-205.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*(4766), 732-735.

Tracey, S. R., Lyle, J. M., & Duhamel, G. 2006. Application of elliptical Fourier analysis of otolith form as a tool for stock identification. *Fisheries Research*, *77*(2), 138-147.

Turan, C., 2006. The use of otolith shape and chemistry to determine stock structure of Mediterranean horse mackerel *Trachurus mediterraneus* (Steindachner). J. Fish. Biol. 69(Suppl. C), 165-180.

Turner, S. M., Manderson, J. P., Richardson, D. E., Hoey, J. J., and Hare, J. A. 2015. Using habitat association models to predict Alewife and Blueback Herring marine distributions and overlap with Atlantic Herring and Atlantic Mackerel: can incidental catches be avoided? *ICES J. Mar. Sci.* 73, 1912–1924. doi: 10.1093/icesjms/fsv166

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Vignon M, Morat F (2010) Environmental and genetic determinant of otolith shape revealed by a non-indigenous tropical fish. *Mar Ecol Prog Ser* 411:231-241. <https://doi.org/10.3354/meps08651>

Wang, Y., U. Naumann, S. T. Wright, and D. I. Warton. 2012. mvabund– an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3:471-474.

War, M., K. Altaff, M. Haniffa. 2011. Growth and survival of larval snakehead Channa striates (Bloch, 1793) fed different live feed organisms. *Turkish Journal of Fisheries and Aquatic Sciences*, 11: 523-528.

Warton, D. I. 2008. Raw data graphing: an informative but under-utilized tool for the analysis of multivariate abundances. *Austral Ecology* 33: 290-300.

Warton, D. I., S. D. Foster, G. De’ath, J. Stoklosa, and P. K. Dunstan. 2015. Model-based thinking for community ecology. *Plant Ecology* 216:669-682.

Warton, D. I., S. T. Wright, and Y. Wang. 2012. Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution* 3:89-101.

Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74(1), 362-368.

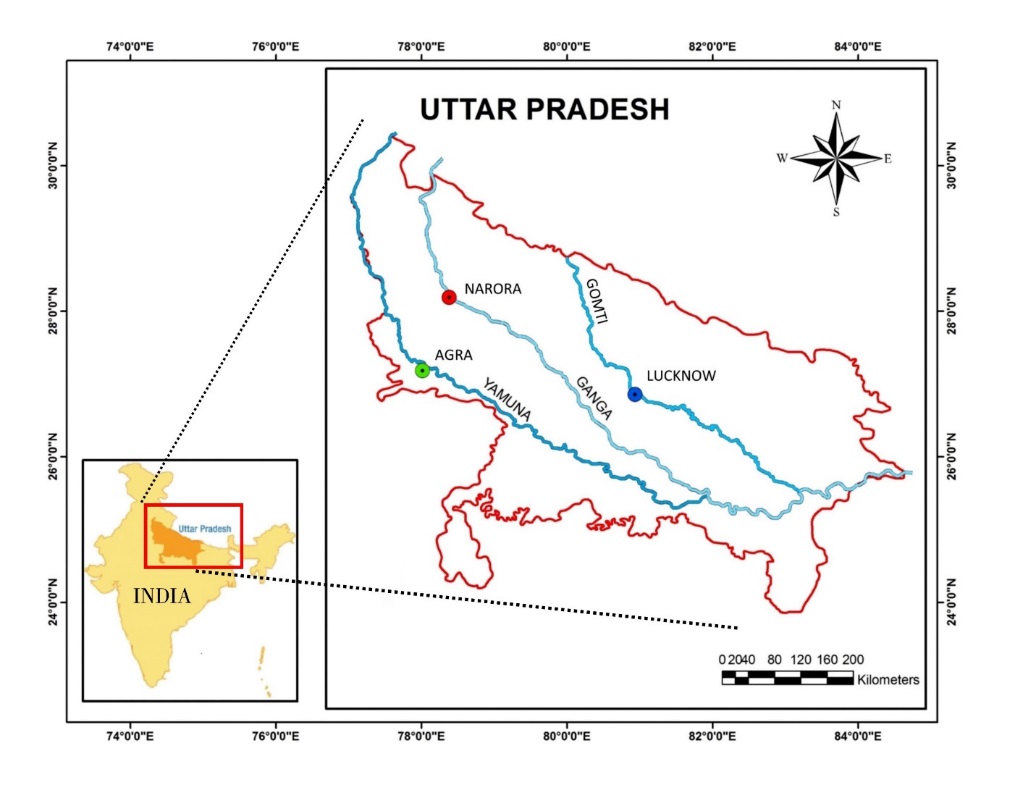
**Tables**

**Table 1** Univariate GLM results for the otolith chemistry analysis. For each element, the Likelihood Ratio test statistic (LR; 2 dp) and *P*-value (3 dp) are shown as well as the % contribution to the multivariate differences (2 dp).

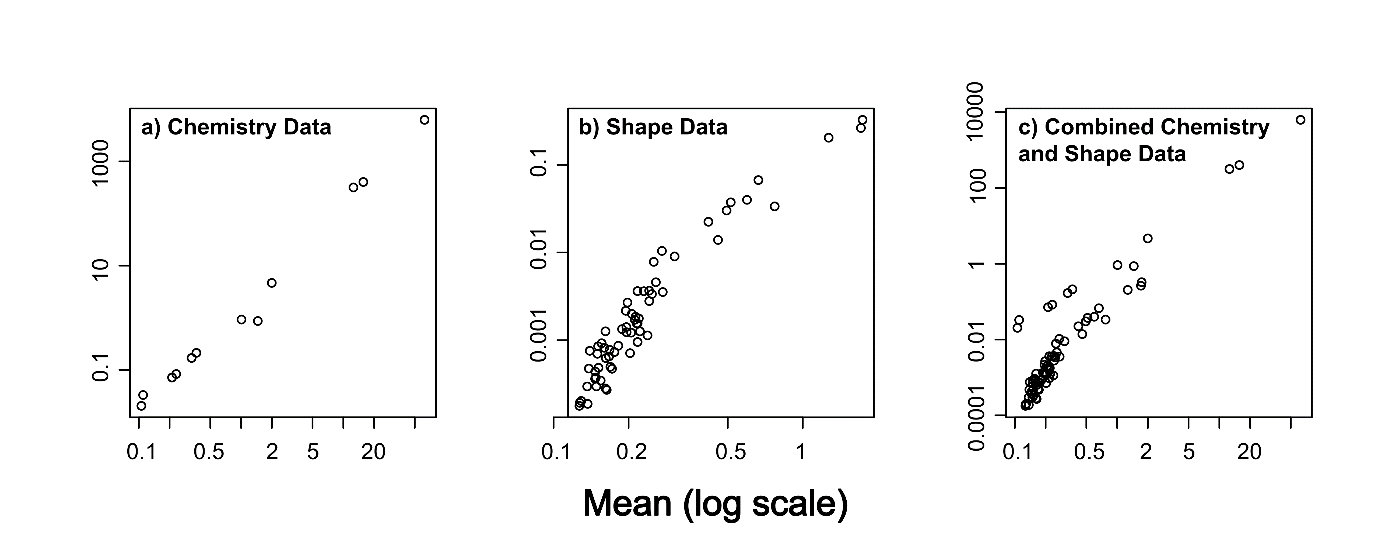
|  |  |  |  |
| --- | --- | --- | --- |
| **Element** | **LR** | ***P*-value** | **% Contribution** |
| Na | 296.57 | <0.001 | 25.84 |
| Sr | 216.61 | <0.001 | 18.87 |
| Mg | 113.60 | <0.001 | 9.90 |
| Ni | 91.59 | <0.001 | 7.98 |
| Cd | 90.53 | <0.001 | 7.89 |
| Ba | 88.85 | <0.001 | 7.74 |
| Mn | 74.82 | <0.001 | 6.52 |
| Cr | 66.78 | <0.001 | 5.82 |
| Pb | 66.60 | <0.001 | 5.80 |
| Zn | 30.30 | <0.001 | 2.64 |
| Fe | 8.23 | 0.012 | 0.72 |
| K | 3.44 | 0.016 | 0.30 |
| Total | 1147.92 | NA | 100 |

**Figure Captions:**

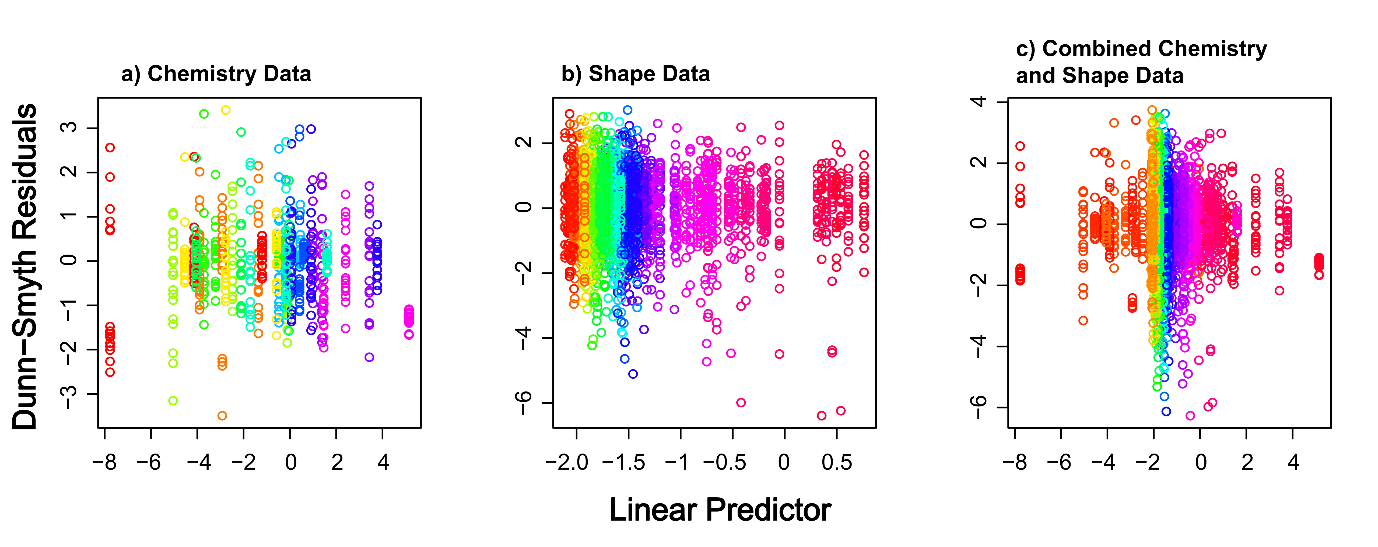
**Figures:**



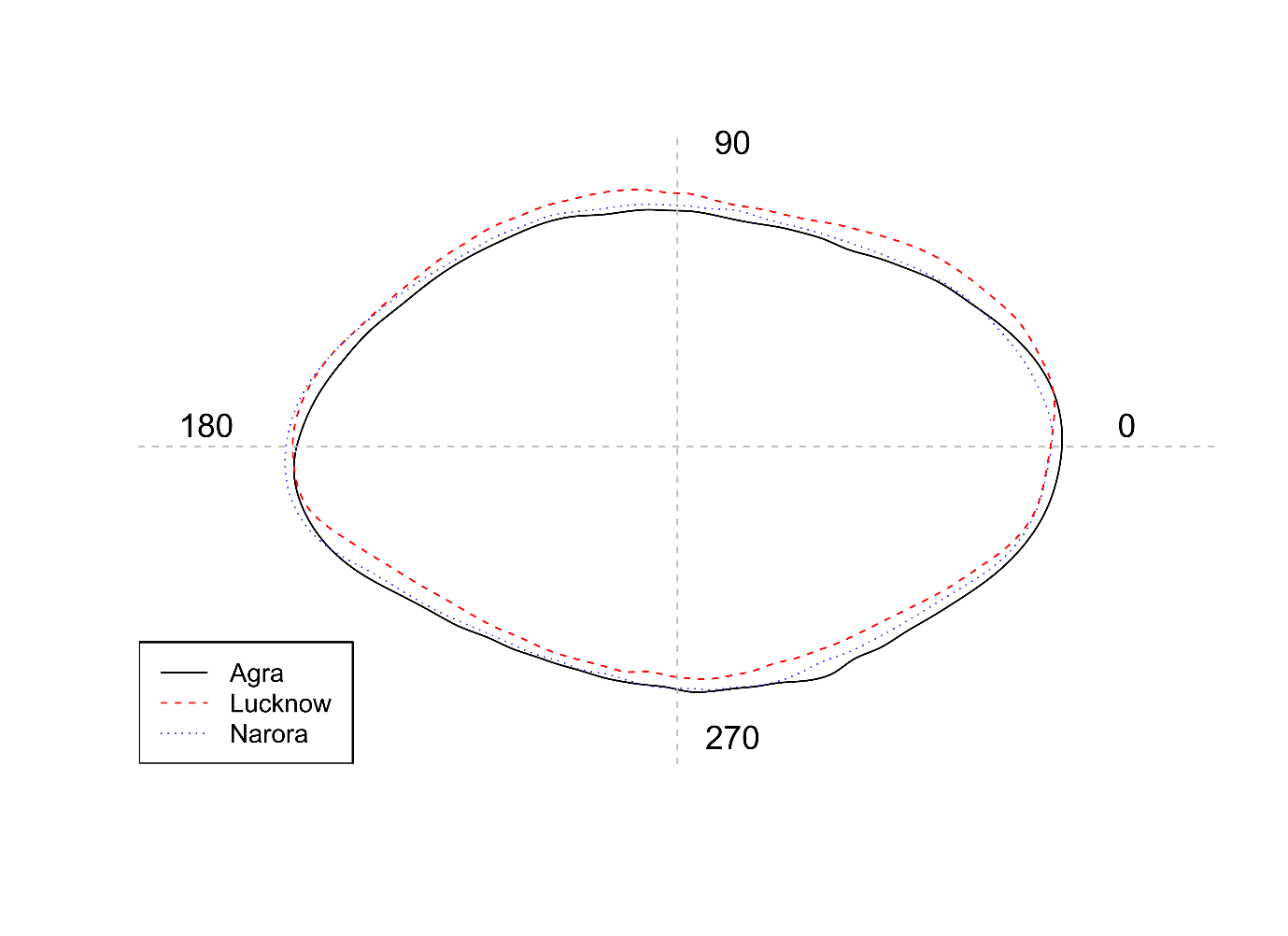
**Figure 1** – Map showing the locations where samples of *C. striata* were collected. The 3 coloured dots represent the sample locations on the 3 different rivers.



**Figure 2** Mean variance plots showing non-linear relationships for a) the otolith chemistry dataset, b) the otolith shape dataset, and c) the combined otolith chemistry and shape dataset. Note the log scale on both axes.



**Figure 3** Dunn-Smyth Residual plots for a) the otolith chemistry dataset, b) the otolith shape dataset, and c) the combined otolith chemistry and shape dataset. No strong patterns are visible in any of the subplots, suggesting that our GLM models were appropriate. Colours show the variables in the analysis.

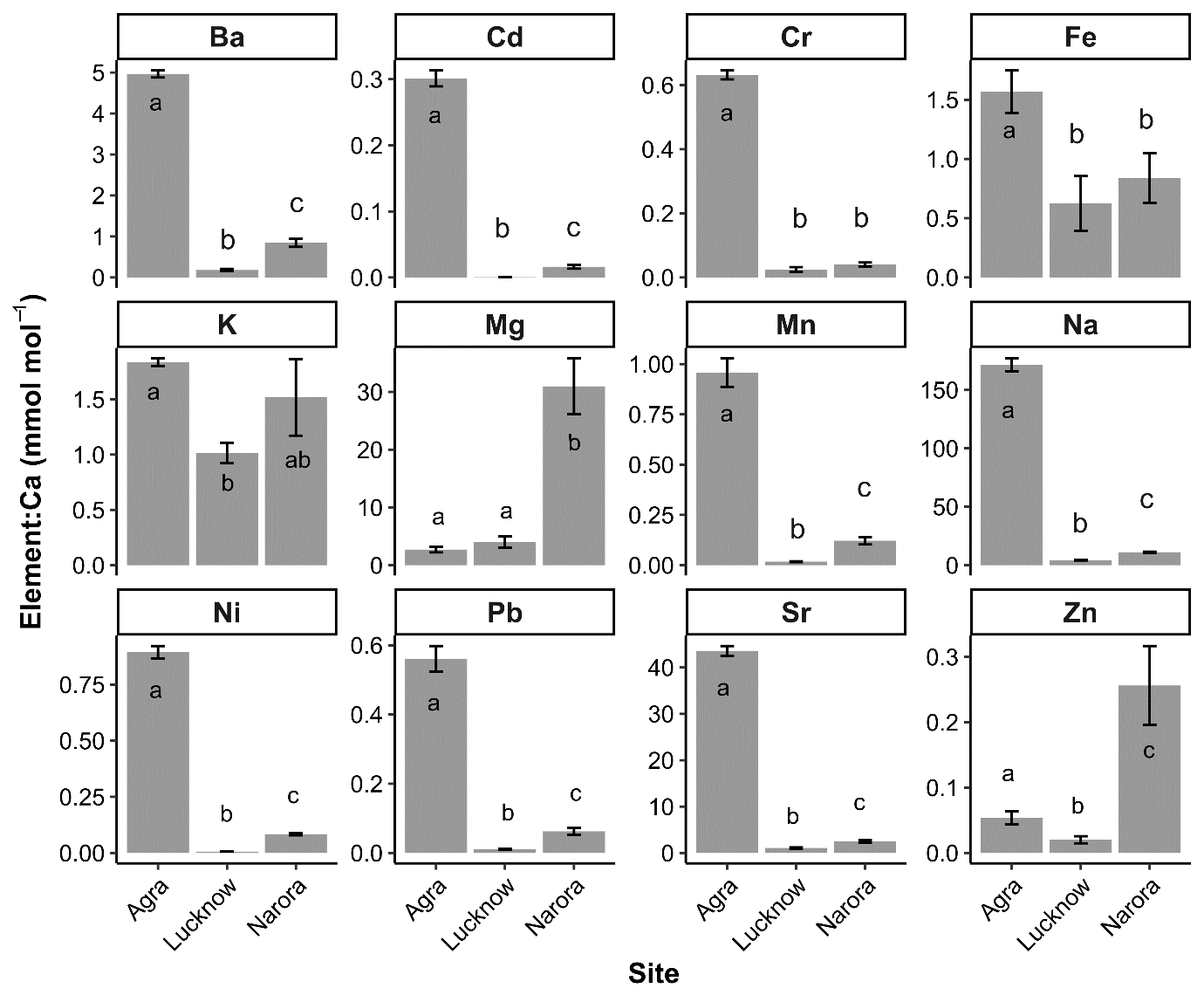


**Figure 4** Mean otolith shape from the three sites. The solid black line represents Agra, dashed red line represents Lucknow and the dotted blue line represents Narora. The wavelet coefficients recreated over 99% of the variance in otolith shape.

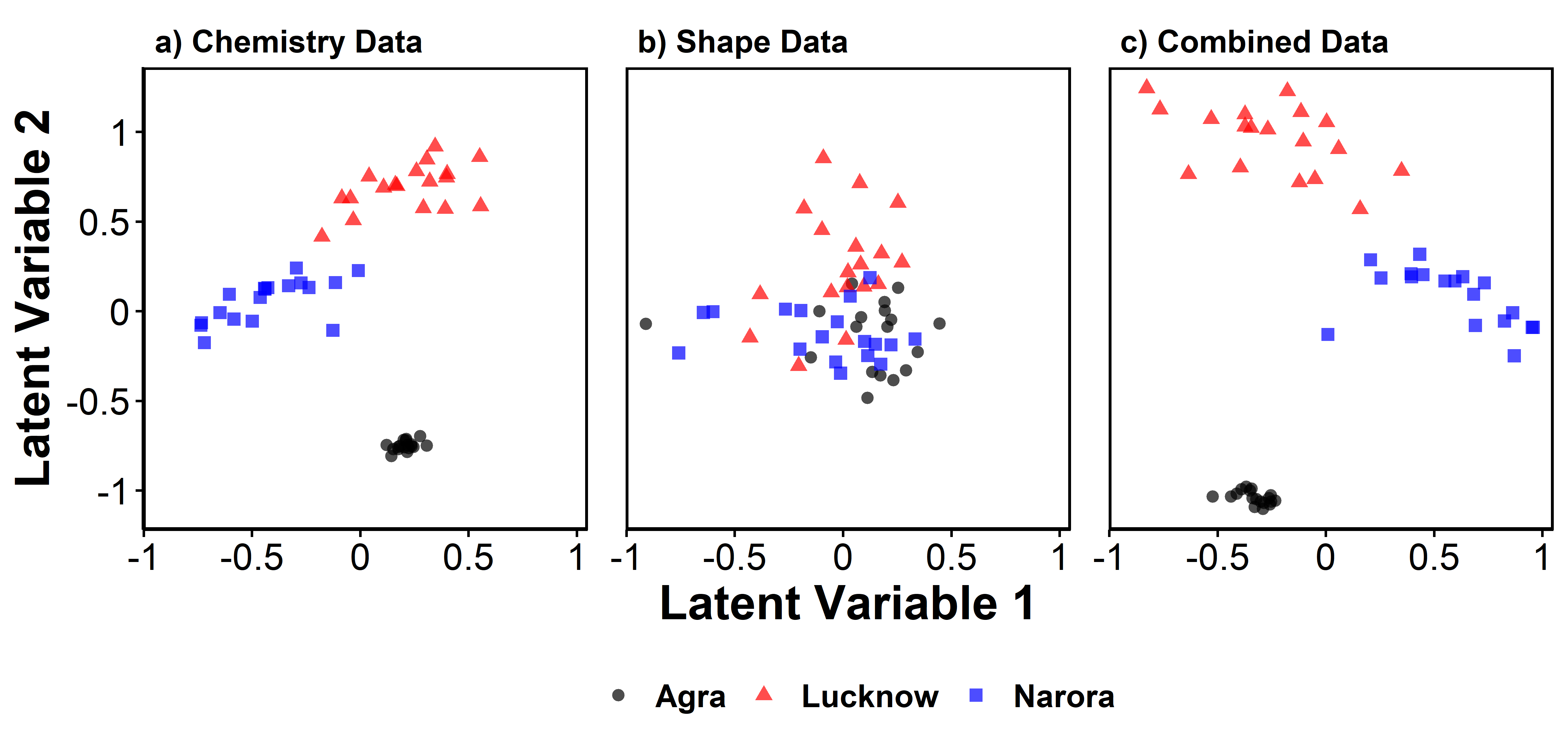
**Chart, scatter chart

Description automatically generated**

**Figure 5** Distance based NMDS ordination based upon a) the otolith chemistry dataset, b) the otolith shape dataset, and c) the combined otolith chemistry and shape dataset. Colours and shapes represent the three groups of *C. striata*. Dissimilarity matrix was based upon Euclidean distances. The assumption of homogeneity of variance was only satisfied for b) Shape Data which would result in the results potentially being unreliable in a) and c) (Warton et al. 2012).



**Figure 6** Mean otolith element concentrations (mmol:mol Ca) for each of the three populations. Error bars show 1 standard error. Within a subplot, bars which do not share a common letter are clearly different (MGLM analysis: *P* < 0.05). For univariate GLM results see Table 1.



**Figure 7** Model-based latent variable ordinations of a) the otolith chemistry dataset, b) the otolith shape dataset, and c) the combined otolith chemistry and shape dataset. Colours and shapes represent the three groups of *C. striata*.