http://app.pan.pl/SOM/app68-Croudace_etal_SOM.pdf

SUPPLEMENTARY ONLINE MATERIAL FOR

Iridescent plumage in a juvenile dromaeosaurid theropod dinosaur

Angus D. Croudace, Caizhi Shen, Junchang Lü†, Stephen L. Brusatte,
and Jakob Vinther

Published in *Acta Palaeontologica Polonica* 2023 68 (2): 213-225.
https://doi.org/10.4202/app.01004.2022

Supplementary Online Material

SOM source data available at
http://app.pan.pl/SOM/app68-Croudace_etal_SOM/SOM_data_2023.xlsx

Supplementary methods

Existing changes to sample size bias

Nordén et al., (2019) showed that small sample sizes have considerable variation in length and diameter measurements, calculated from coefficient of variation (CV) for each variable. This is significant in their study because the bias resulting from the variation would be inconsistent and only affect the black, brown and grey categories which consist exclusively of data from Li et al. (2012), but not the iridescent category which Nordén et al. (2019) contributed most of the data for. Samples with a sample size of less than 10 were thus excluded by Nordén et al. (2019) to reduce bias in the datasets; for sample sizes of 10 or greater, CV is within 6% and considered acceptable.

Nordén et al. (2019) also noted that variables of CV and skew are sensitive to sample size and could introduce bias into an analysis; at a sample size of 10, for which variability in length and diameter measurements has been deemed acceptable, length CV has a CV of 23.6%. To include length CV and skew without a large variability, the minimum sample size would need to be much higher than 10, resulting in the exclusion of a substantial proportion of the dataset. Instead, only length, diameter and aspect ratio measurements were included in their study.

Changes to Norden dataset

Four samples originally from the Li et al. (2012) project were mislabelled with the wrong colour category in the online data available from Nordén et al. (2019). These have been corrected in this study. The altered data included Anas_platyrhynchos_a (brown in Nordén et al. *2019*, should be iridescent), Falco_sparverius_c (grey in Nordén et al. *2019*, should be brown), Anas_platyrhynchos_b (grey in Nordén et al. *2019*, should be iridescent) and Anas_platyrhynchos_c (iridescent in Nordén et al. *2019*, should be brown).

QDA

There are some challenges inherent in the previous applications of QDA in this field. Some of the assumptions of QDA (e.g. that the independent variables are continuous and multivariate normally distributed) may not always be met. In addition, the inclusion of some variables has already been criticised (Nordén et al. 2019), as well as the stepwise method in which they are entered in an analysis.

QDA variable selection

Li et al. (2010, 2012) applied a model comprised of eight continuous variables: length, length CV, length skew, diameter, diameter CV, diameter skew, aspect ratio, and aspect ratio skew. Their analysis allowed these eight variables to enter a model in a forward stepwise manner based on significance values. This method includes variables in the model in an order determined by level of a statistical significance (p-values). This results in those where p was not < 0.05 being excluded, in this case length skew and diameter skew. Hu et al. (2018) ran a similar analysis but in a backward stepwise manner. This starts the analysis by including all variables, and then removing those not significant i.e., $p > 0.05$ first. Like Li et al. (2012), those dropped because they were not significant were length skew and diameter skew. Caution should be exercised with any method based on p-values alone. Entering variables based on statistical significance may include nuisance variables that are coincidentally significant, while other important explanatory variables are excluded because they may not happen to be statistically significant (Smith 2018). Stepwise methods are hence known to be prone to overfitting, performing well on training data but poorly on new unknown data (Foster and Stine 2006; Whittingham et al. 2006). Accordingly, stepwise methods are not recommended. In this study, the variables were entered together into the model, rather than entered after stepwise selection based on significance.

Avoiding stepwise selection in MLR using AIC values

Nordén et al. (2019) avoided stepwise variable selection in MLR by adopting the approach of Grueber et al., (2011). This is based on variable selection depending on Akaike information criterion (AIC) values. In practice, this involves examining and comparing AIC values for all possible model combinations of predictor variables. Each AIC value estimates the out-of-sample prediction error for that model. The lowest value identifies the model that loses the least information and is therefore of a higher quality in terms of fit and accuracy. This lowest AIC model is the one that is then used for predictive classification. AIC includes a component in its calculation which discourages overfitting by penalising increasing complexity (greater number of predictor variables included), which makes it a better justified choice for model building than stepwise variable selection.

Variable selection was based on AIC values of length, diameter and aspect ratio predictor variables using the *gvselect* function (Lindsey and Sheather 2015) in Stata-16. This constructed all possible models using the available variables in the dataset and ranked them by AIC value. The resultant predictor variables selected by lowest AIC, and subsequently included were diameter and aspect ratio. Hollowness and flatness were excluded from *gvselec*t-based model building and were also not used by analyses **M1** and **M2** because every sample was solid and cylindrical - no variation would have been present for these two binary categorical predictor variables.

For analysis **M3** categorical predictor variables of hollowness and flatness were added to the previous AIC analysis and lowest AIC here selected diameter, aspect ratio, hollowness, and flatness as predictor variables for inclusion.

Confirmation that LogisticDA is equivalent to MLR

The maximum likelihood values at the convergence of the estimation algorithms for both MLR and LogisticDA were shown to be identical when compared using the same data and equivalent commands. This confirms the exact computational equivalence of the methods to several decimal places. LogisticDA is available for applications with two or more categories and was hence selected for this study, applied using the *discrim logistic* command in Stata-16 (StataCorp 2019).

Model performance, LOOCV and k-fold cross validation

LOOCV - as its name suggests - omits one data point at a time and trains the model on the rest of the dataset. It then tests the model on the left-out data point. This is repeated for all data points and the percentage of results correctly predicted is reported. Although Stata has a function for this, it can only be applied to linear and quadratic discriminant analysis. Fortunately, utilising R version 3.6.3 (R Core Team 2020) allowed computation of a LOOCV value for both QDA and MLR. However, a possible issue with LOOCV is that, because all data points are treated individually, results can exaggerate any error from outlying data points or sample, resulting in higher prediction variation (James et al. 2013).

Repeated *k*-fold cross-validation was the second method used to complement the LOOCV. This randomly partitions the data into *k* equally sized subsets (or ‘folds’), for which *k*-1 are used as a training dataset and the remaining subset is then used to test the trained model. This is repeated *k* times so that each subset is tested once. Typically, *k* is recommended to be 5 or 10 because these values are observed to produce error rate estimates that are not excessively affected by high bias or variance (Kassambara 2017). This entire process is then repeated several times and the prediction scores averaged. This method gives more accurate estimates of test error rate than LOOCV because the data are in small subsets.

This reduces potential prediction error from an outlier or biased data point, making it more robust than LOOCV (James et al. 2013).

The R *caret* library (Kuhn 2008) was used to run repeated *k*-fold cross-validation analyses with 5 repeats. After setting a seed for reproducibility of analyses results, the data were split into random training and test datasets. This is to prevent the test data being 'seen' by the algorithm, to allow for a true test on unseen data. This was carried out by the *createDataPartition* function, which not only randomly selects from rows in the data but also stratifies the data first, dividing it within each colour category (random within colour category), preserving overall class distribution for the training model (Kuhn 2008).
The output provided a value for Cohen's Kappa, which measures the agreement of the predictions considering the possibility that they might be due to chance. Cohen's Kappa ranges from 0 (predictions no better than expected from chance) to 1 (perfect agreement). Tests were run using both $k = 5$ and $k = 10$, i.e. five-fold and ten-fold cross-validation for each model to check for sensitivity, but the effect on the Kappa was negligible (0-0.02% difference).

References

Foster, D.P. and Stine, R.A. 2006. Honest confidence intervals for the error variance in stepwise regression. *Journal of Economic and Social Measurement* 31: 89–102.

Grueber, C.E., Nakagawa, S., Laws, R.J. and Jamieson, I.G. 2011. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology* 24 (4): 699–711.

Hu, D., Clarke, J.A., Eliason, C.M., Qiu, R., Li, Q., Shawkey, M.D., Zhao, C., D'Alba, L., Jiang, J. and Xu, X. 2018. A bony-crested Jurassic dinosaur with evidence of iridescent plumage highlights complexity in early paravian evolution. *Nature Communications* 9 (1): 217.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer, .

Kassambara, A. 2017. *Machine Learning Essentials: Practical Guide in R*. STHDA, .

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28 (5): 1–26.

Li, Q., Gao, K.-Q., Meng, Q., Clarke, J.A., Shawkey, M.D., D'Alba, L., Pei, R., Ellison, M., Norell, M.A. and Vinther, J. 2012. Reconstruction of Microraptor and the Evolution of Iridescent Plumage. *Science* 335 (6073): 1215 LP – 1219.

Li, Q., Gao, K.-Q., Vinther, J., Shawkey, M.D., Clarke, J.A., D'Alba, L., Meng, Q., Briggs, D.E.G. and Prum, R.O. 2010. Plumage Color Patterns of an Extinct Dinosaur. *Science* 327 (5971): 1369 LP – 1372.

Lindsey, C. and Sheather, S. 2015. Best subsets variable selection in nonnormal regression models. *Stata Journal* 15 (4): 1046–1059.

Nordén, K.K., Faber, J.W., Babarović, F., Stubbs, T.L., Selly, T., Schiffbauer, J.D., Peharec Štefanić, P., Mayr, G., Smithwick, F.M. and Vinther, J. 2019. Melanosome diversity and convergence in the evolution of iridescent avian feathers—Implications for paleocolor reconstruction. *Evolution* 73 (1): 15–27.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. .

Ripley, B. and Venables, B. 2002. Modern Applied Statistics with S. Fourth edition. Spring, New York. .

Smith, G. 2018. Step away from stepwise. *Journal of Big Data* 5 (1): 32.

StataCorp. 2019. Stata Statistical Software. .

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. and Frecklenton, R.P. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* 75 (5): 1182–1189.

Table S1: Experimenter bias test: percentage differences in measurements from additional individuals on images from two samples.

Sample	Length	Width	Aspect Ratio
6	0.1%	0.5%	0.1%
	11.8%	11.3%	11%
7	1.9%	9.9%	9.5%
	2.1%	1.5%	4.4%

Table S2: Mean melanosome measurements for Wulong DNHM D2933 used to predict corresponding colour classification. Samples 1 through 4 were not available to study. Samples 5, 8, 10 and 12 did not preserve any melanosomes.

***Wulong DNHM D2933* Sample**	**Sample Location**	**Mean melanosome measurement**			**n =**
		Length (nm)	**Width (nm)**	**Aspect ratio**	
6	Distal portion of metatarsus leg feather	1027.14	175.08	6.05	107
7	Distal portion of tibia leg feather	908.92	139.23	6.73	13
9	Proximal portion of dorsal ilium feather	1444.75	396.31	3.77	68
11	Proximal portion of dorsal ilium feather	1524.72	450.87	3.53	156
13	Unidentifiable portion of a forelimb feather	1475.29	436.55	3.48	115
14	Proximal portion of forelimb ulna feather (right)	1216.78	251.73	5.15	38
15a	Not identifiable. Possibly Abdomen or forelimb humerus feather	1493.25	452.95	3.45	122
15b	Not identifiable. Possibly Abdomen or forelimb humerus feather	833.07	183.86	4.66	14
15c	Not identifiable. Possibly Abdomen or forelimb humerus feather	892.10	622.15	1.31	52
16	Proximal portion of forelimb ulna feather (left)	887.21	171.77	5.45	48

Table S3. Average melanosome length, width and aspect ratio for different colour categories including iridescent sub-categories, from the complete Nordén et al. dataset (2019). Note the conflation of black and grey with some sub-categories of iridescence.

Colour Category	Average Length (nm)	Average Diameter (nm)	Average Aspect Ratio	Number of samples	Notes
Black (Solid Cylindrical)	985.76	279.73	3.67	27	
Brown (Solid Cylindrical)	476.40	286.13	1.70	28	
Grey (Solid Cylindrical)	1202.19	405.14	3.19	25	
Iridescent Solid Cylindrical	1129.4	209.74	5.62	90	Single iridescent morphology used in all previous studies
Iridescent Solid Flat	1386.34	346.32	4.11	22	
Iridescent Hollow Cylindrical	1066.10	256.05	4.36	21	Similar dimensions to black
Iridescent Hollow Flat	1404.27	579.97	2.64	20	Similar dimensions to grey, particularly aspect ratio

Table S4. Colour classification posterior probabilities for *Wulong* using an unmodified ‘Nordén’ dataset which includes hollow and flat iridescent melanosomes.

Sample	Location of sampled feather	Predicted colour	Probability (%) Q3 QDA Nordén	Predicted colour	Probability (%) M3 MLR Nordén
6	Distal portion of metatarsus	Iridescent	99.86	Iridescent	97.63
7	Distal portion of tibia	Iridescent	100	Iridescent	99.37
9	Proximal portion of dorsal ilium	Iridescent	75.64	Grey	54
11	Proximal portion of dorsal ilium	Iridescent	70.83	Grey	68.66
13	Unidentifiable forelimb portion	Iridescent	72.30	Grey	66.07
14	Proximal portion of forelimb ulna (right)	Iridescent	74.65	Iridescent	84.16
15a	Not identifiable. Possibly abdomen or forelimb humerus	Iridescent	70.57	Grey	69.59
15b		Iridescent	83.99	Iridescent	81.26
15c		Grey	53.56	Brown	83.19
16	Proximal portion of forelimb ulna (left)	Iridescent	99.57	Iridescent	94.29

STATA CODE

AIC – for updated variable selection method

```
gvselect <term> length diameter aspectratio : mlogit colourcategory
<term>
```

QDA

```
discrim qda length diameter aspectratio, group(colourcategory)
priors(proportional)
```

MLR

```
discrim logistic diameter aspectratio hollow flat
group(colourcategory) priors(proportional)
```

R CODE

Required libraries (MASS, nnet (Ripley and Venables 2002), caret (Kuhn 2008))

QDA - LOOCV

```
> q2loocv = train(colour~.,data = q2,method = "qda", trControl =
trainControl(method= "loocv"), trace = FALSE)
> q2loocv
```

QDA – repeated *k*-fold cross-validation

```
> q2_idx = createDataPartition(q2$colour, p=0.8, list = FALSE)
> q2_trn = q2[q2_idx, ]
> q2_tst = q2[-q2_idx, ]
> q2kfold = train(colour~.,data = q2,method = "qda", trControl =
trainControl(method= "repeatedcv", repeats = 5, number = 5), trace =
FALSE)
> q2kfold
```

MLR - LOOCV

```
>m2loocv = train(colour~.,data = m2,method = "multinom", trControl =
trainControl(method= "loocv"), trace = FALSE)
> m2loocv
```

MLR – repeated *k*-fold cross-validation

```
> m2_idx = createDataPartition(m2$colour, p=0.8, list = FALSE)
> m2_trn = m2[m2_idx, ]
> m2_tst = m2[-m2_idx, ]
> m2kfold = train(colour~.,data = m2,method = "multinom", trControl
= trainControl(method= "repeatedcv", repeats = 5, number = 5), trace
= FALSE)
> m2kfold
```