Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date:
Application Note

OXFORD

Genome analysis

Clustering Size Measurements of Extinct and Extant Bird Species

Fiel Dimayacyac [1]

[1]University of British Columbia, Bioinformatics, Vancouver, British Columbia

Associate Editor: XXX

Received on XXX; revised on XXX; accepted on XXX

Abstract

Motivation: Identify if extinct and extant birds cluster according to order, extinction status, or flightlessness using size measurements.
Results: This analysis found that bird size data minimized total within sum of squares with a k means of seven, but that bird size measurements did not cluster according to extinction status or flightlessness. Some clustered according to taxonomic relationship.
Availability: All scripts, data, and manuscript files are freely available on the web at the Open Science Foundation website.
Contact:fieldimayacyac@gmail.com
Supplementary information: Supplementary data are available on the Open Science Foundation website.

1 Introduction

In this study I reanalyze the paper by Fromm and Meiri (2021) wherein the authors aggregated body size measurements, flightlessness, order, and habitat of 9000 bird species with a focus on identifying size trends between extant species and species driven to extinction by human intervention. Fromm and Meiri (2021) found a significant difference between extinct and extant birds in overall body size and other factors, concluding that species driven to extinction tended to be larger, flightless, and island-dwelling.

My aim is to see if those trends hold up in the body size data itself; i.e., can an unsupervised learning approach naively discover these same trends. Specifically, I use k-means clustering as implemented in the tidymodels R package (Kuhn and Wickham, 2020).

2 Results

Before performing k-means clustering, I first cleaned the data, and then selected variables to use later. I found that much of the data was unavailable, so I chose two bone measurements with the most overlapping available data points that included both extinct and extant species. These bone measurements being for humerus and tarsometatarsus.

Following data filtering I then visualized the remaining data to see if there were any obvious trends across extinction status, flightlessness, or order. Axes were log-normalized for inter-specific comparisons. No evident trends were visible previous to clustering (Figure 1).

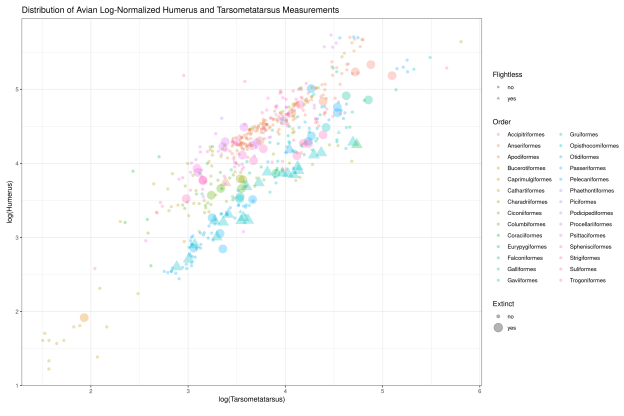

Fig. 1. Distribution of size measurements for extinct and extant bird species with flightlessness, Order, and extinction status. Tarsometatarsus measurements are on the x-axis and Humerus measurements are on the y-axis

To identify the optimal number of k-means to describe the patterns in the data, I then performed k-means analysis, varying the k value from one to ten (Figure 2).

To quantify how optimal each k-means value was, I calculated the total within sum of squares for each value k for the model. The lowest value k that minimized the total within sum of squares was seven (Figure 3).

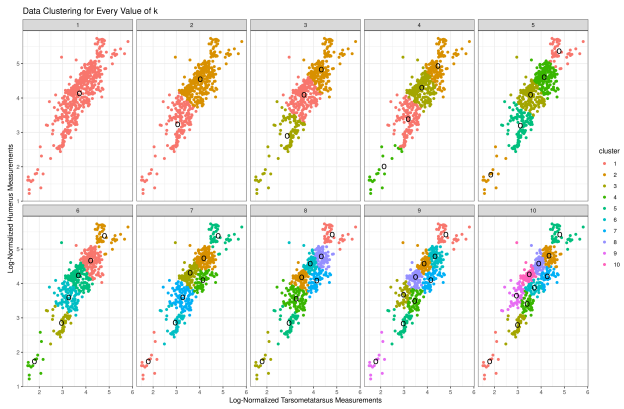

Fig. 2. Clustered size measurement data with varying values of k. Circles represent cluster centers

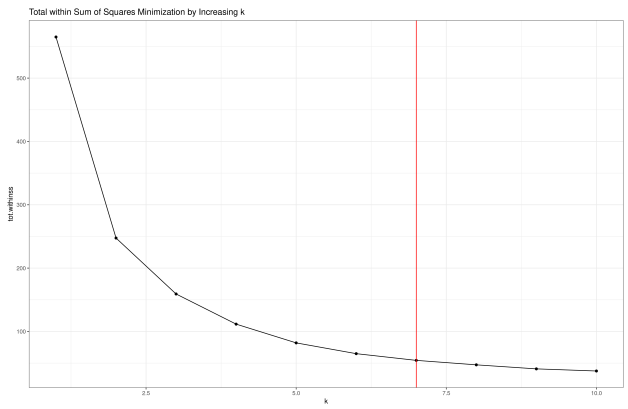

Fig. 3. Total within sum of squares for varying k values. The red line intersects at a k of seven

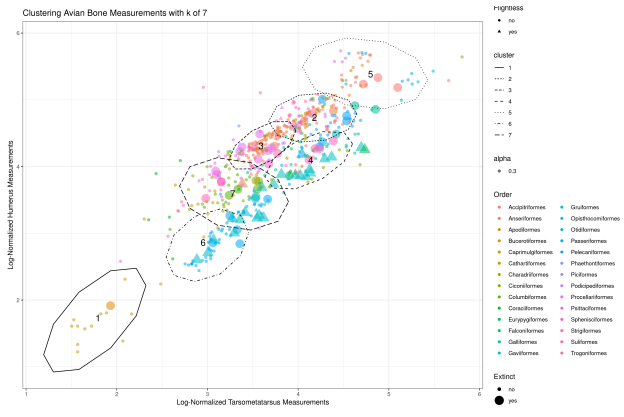

Fig. 4. Distribution of bird measurements with clusters overlain as ellipses and numbers at cluster centers

Finally, I plotted the clusters with a k of seven overlaid onto Figure 1 to identify if there were any trends within the clusters. No noticeable trends were evident in the data (Figure 4).

3 Methods

Body size measurements data was downloaded from data dryad entry 10.5061/dryad.1rn8pk0tb via R script. The groundhog package was used for package version management and library control. Exploratory data analysis was carried out using the tidyverse R package (Wickham *et al.*, 2019). Data was loaded into R using the readxl package, filtered for NA values using dplyr and coerced using base R (Wickham *et al.*, 2022b) (Wickham *et al.*, 2022a). Variables for k-means clustering were chosen based on availability of data; the two variables with the most overlapping available values were chosen for downstream analysis. K-means clustering was performed using the tidymodels R package (Kuhn and Wickham, 2020). Model variables were tidied and gathered using the broom R package (Robinson *et al.*, 2022). Data was plotted using ggplot2 (Wickham, 2016).

4 Discussion

In the original paper the authors found that birds driven to extinction by humans were more likely to be larger, flightless, and live on islands (Fromm and Meiri, 2021). In this analysis I aimed to recreate these findings with an unsupervised learning approach. However, I was unable to do so. One possible explanation is the data points I used. In my analysis I filtered much of the data and ended up only using two data points; Humerus and Tarsometatarsus measurements. Because these two data points are correlated, it is possible that other relationships in the data were not able to be elucidated from this specific combination. Additionally, the authors reached those conclusions by aggregating the size measurements into a single measurement of estimated body size (Fromm and Meiri, 2021). They estimated bird body size by employing a linear regression model using data from all the measurements they gathered to allow for an even comparison between extinct and extant species (Fromm and Meiri, 2021). This esimated body mass metric was then used in their statistical analysis to compare extinct and extant birds and how they came to their conclusions. My analysis did not use estimated body mass, so it is possible that this is the reason I was unable to come to the same conclusions. In the future, one could attempt to employ a different unsupervised approach such as a random forest model using the same estimated body mass measurements to see if that could recreate their results.

5 Conclusion

In this study I aimed to recreate the results of Fromm and Meiri (2021) using an unsupervised k-means clustering approach but was unable to come to the same conclusions. Future studies can employ alternative models using different parts of the data to improve upon my analysis.

References

Fromm, A. and Meiri, S. (2021). Big, flightless, insular, and dead: characterizing the extinct birds of the Quaternary. Type: dataset.

Kuhn, M. and Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*

Robinson, D., Hayes, A., and Couch, S. (2022). *broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J.,

Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686.

Wickham, H., François, R., Henry, L., and Müller, K. (2022a). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Wickham, H., Hester, J., and Bryan, J. (2022b). *readr: Read Rectangular Text Data*. https://readr.tidyverse.org, https://github.com/tidyverse/readr.