Archetype Analysis of golden eagle migration patterns using Bayesian Methods

Abraham Arbelaez, McNair Scholar
The Pennsylvania State University

McNair Faculty Research Adviser:
Ephraim M. Hanks, Ph.D.
Associate Professor of Statistics
Department of Statistics
Eberly College of Science
The Pennsylvania State University

Abstract

Animal migration has the potential to be a very good indicator of environmental changes that could affect us all. It also helps us understand the different species with whom we share this planet. In order to do the mentioned above, we analyzed a monthly golden eagle (*Aquila Chrysaetos*) location data. The main research question was whether covariates, such as age, could be a big factor on their migration routes. This exploration was possible through an archetype analysis, which is a statistical nonparametric approach that represents each individual as a mixture of multiple estimated archetypes. In addition to a traditional archetype analysis, we developed a new approach to archetype analysis, in which covariates are considered, and a subset of the archetypes is defined by existing golden eagles who exhibit known, interpretable behaviors, in order to be fitted using Bayesian methods and Markov Chain Monte Carlo simulations. This approach was developed using R with Machine Learning techniques and Bayesian Statistics. Once the analysis was complete, we were able to exhibit that covariates such as age influence birds' behavior and their migration routes, concluding that the older they get, the more likely they will belong in a non-migratory archetype. This novel approach showcases a new proposal for such databases and optimize processes in ecological research.

Keywords: Archetype analysis, Bayesian methods, MCMC, spatio-temporal statistics, R.

1. Introduction

Machine learning is a powerful tool that is currently being used in many areas of science, finance, and industry. However, it can be abstract, and hard to digest due to its algorithms and complex components. Supervised learning has a measure of success (or lack thereof) that can be used to measure effectiveness, whereas unsupervised machine learning or "learning without a teacher" (Hastie et al, 2009) draws inferences from data sets without labels, therefore, it finds patterns when one is not sure what one is looking for. There are multiple unsupervised learning techniques, such as K-means clustering algorithms, Gaussian Mixture Models, Principal Component Analysis (PCA), and so forth. In this paper, I will concentrate on Archetypal Analysis (AA).

AA was first introduced by Cutler and Breiman (1994); they proposed an approach that would characterize the "archetypal patterns" in a data set. Their first example was a question of how many sizes were needed to fit all Swiss soldiers faces in face masks. Contrary to what clustering analysis offers (using the "average" members of certain groups as the prototype), the idea of AA is enclosing the data set into a gradient, where the individuals are weighted combinations of the archetypes. In other words, every soldier has a mask that can cover their face; therefore, the mask will be large enough to cover everyone's face, but a mask that is slightly larger than their face can still be worn.

In this article, we will be analyzing a monthly Golden Eagle (*Aquila Chrysaetos*) telemetry data consisting of 180 bird-years of monthly location observations. The data was obtained over the course of 6 years, where the earliest observations are from 2012 and the latest are from 2018 and it can be seen in Figure 1. Each observation represents one coordinate point as follows: [$X_1, X_2, \ldots, X_{12}$], and [$Y_1, Y_2, \ldots, Y_{12}$] in the Cartesian plane. The data consist of 180 bird-years from 63 unique golden eagles, as some birds were tracked for multiple successive years.

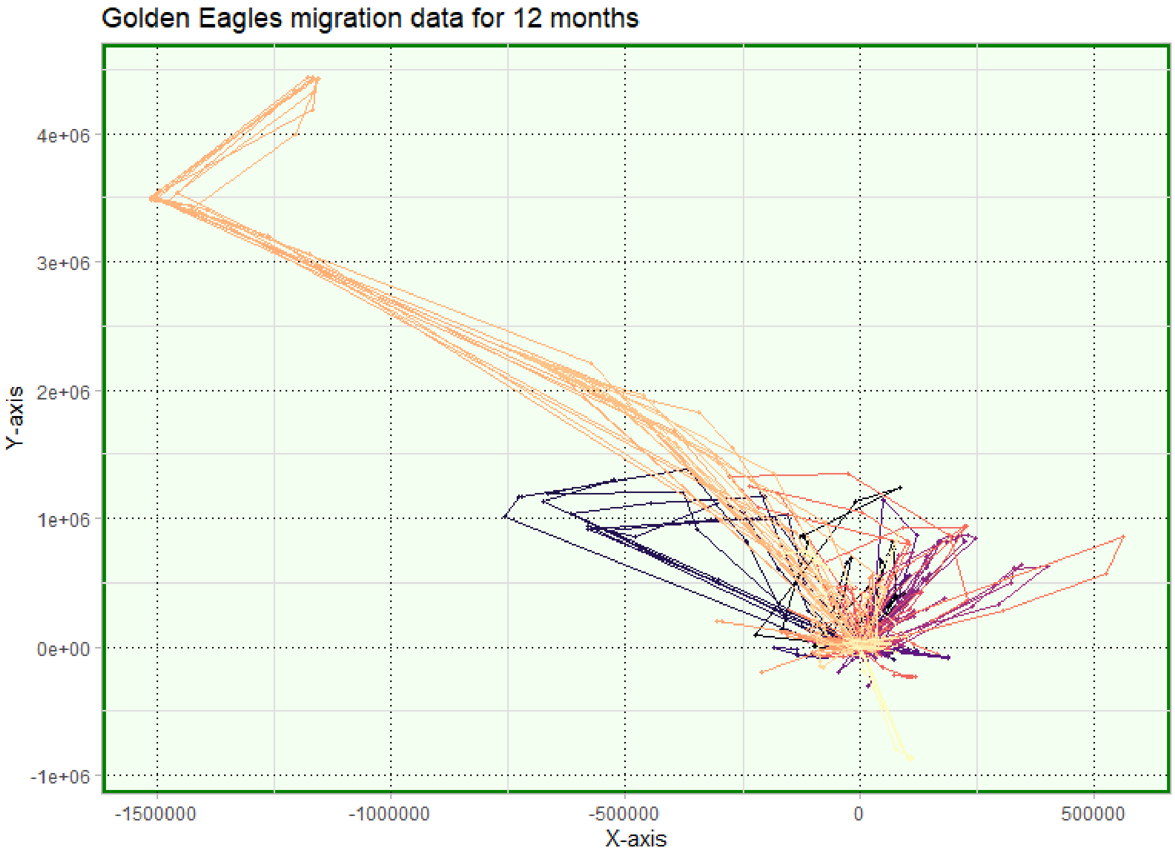


Figure 1: Plot for 63 unique birds with 180 bird-years, where different shades of colors represent the 180 different bird-years. This data set was normalized so that all start at (0,0)

Ecologists have increasingly used hierarchical Bayesian statistical (Conn et al, 2018; Hobbs and Hooten, 2015; Sahu, 2022; Kéry and Royle, 2020) since they can account for uncertainty in ecological analysis (Cressie et al, 2009), and provide an approach to model latent patterns common in ecological systems.

One thing that distinguishes animals is a power they have of moving themselves from place to place (Gray, 2013). This power allows us to say that different birds can move differently from each other. Lack (1968) notes that in many bird migrants, a higher proportion of juveniles than adults migrate, in other words, birds move differently as they age. This paper revolves around two premises on animal behavior: different birds move differently, and birds move different as they age.

In this work we develop a novel Bayesian Hierarchical Model to provide a data driven classification of bird migration strategies, and to explain how birds change migratory behavior as they age. The remainder of this manuscript is organized as follows. In Section 2, an outline of the data processing and data visualization can be found. In Section 3, the results are shown along with different plots obtained. In Section 4, there is a brief summary and discussion concerning our findings to wrap up our writing.

2. Methods

This section is organized as follows. In Section 2.1, an outline of the data processing and data visualization can be found. In Section 2.2, a background information on Archetypal Analysis considered in this work is provided along with the first model on our data. In Section 2.3 the use of covariates (such as age) to improve the model can be found with an AA analysis and Bayesian approach. Lastly, Section 2.4 talks about the implementation of algorithms such as Markov Chain Monte Carlo to satisfy our study.

2.1 Exploratory Data Analysis

We conducted an Exploratory Data Analysis (EDA). As noted previously, our data consisted of 180 different bird-years from 63 different eagles, which meant that there were some birds that were tracked more than one year. The birds that were tracked the most were 4C.Angus_11 and 4C.Eddys_11. Figure 2 shows data from these two eagles tracked over seven years. Despite the fact that these two birds have the most observations (more data leads to lower estimation variance, which results in better predictive performance), we can see that there is a lot of variance and uncertainty. This paper tries to tackle this problem with a Bayesian approach, which will be explained later in Section 2.2, and with the use of age as a covariate to improve our model.

The constant data manipulation from wide tables to narrow tables, and vice-versa was crucial to create visualizations such as Fig. 1 and Fig. 2 with the `ggplot2` library and animations that were created with the `gganimate` library. Code can be found in the Appendix of this document.

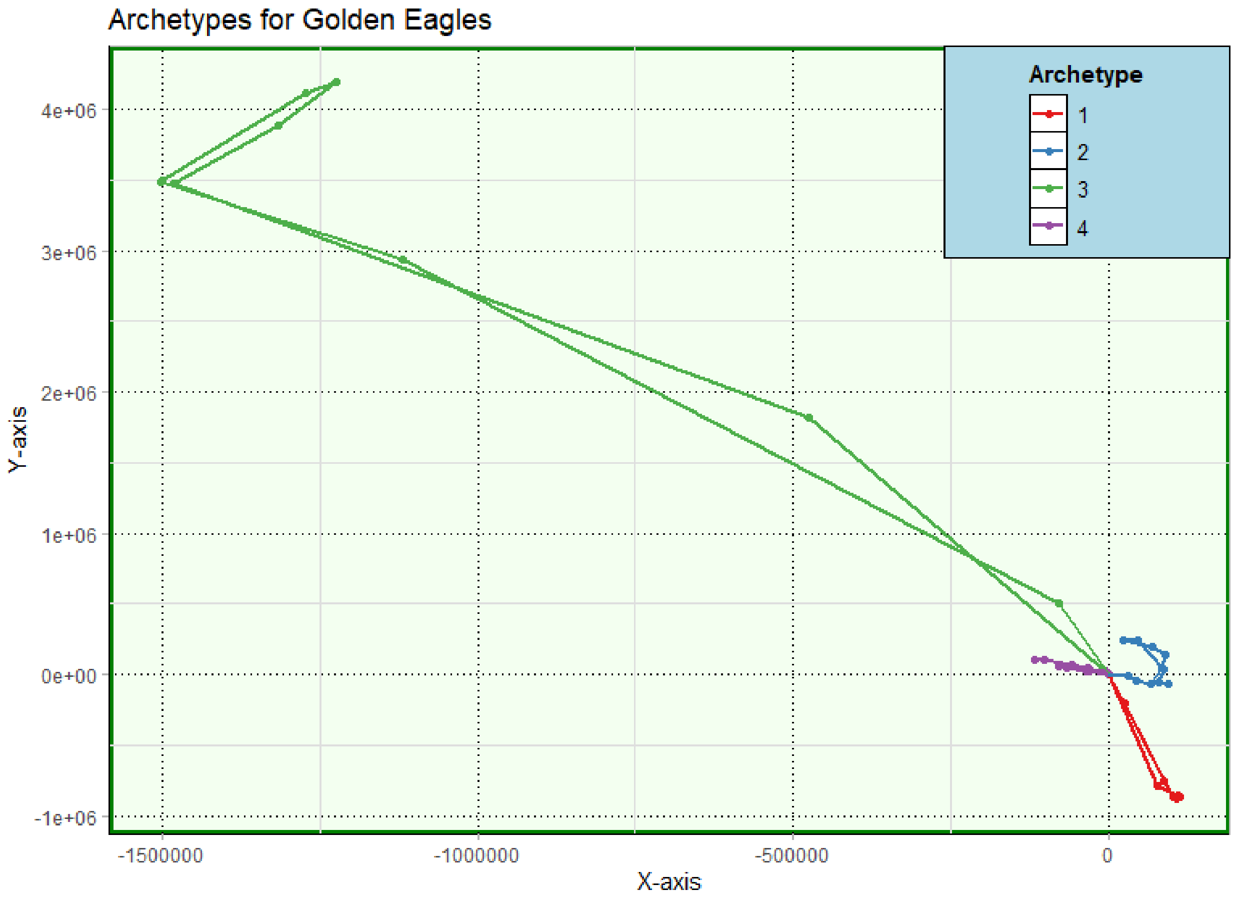


Figure 2: Monthly observations of two birds aged 1 to 7. *4C.Eddys_11* shows a relatively consistent migration pattern. Both birds show a trend of decreasing their migration distance as they age.

The goal of our analysis is to provide a data driven classification of bird migration strategies, and to explain how birds change migratory behavior as they age. To do so, we create two new variables from the telemetry data. The *distance* variable, as its name says, is the total distance traveled by the birds given by (1) where d_i is the distance in form of a scalar quantity from point d_i and d_{i+1} in the data set

$$\sum_{i=1}^{12} d_i$$

which can also be written as

$$\sum_{i=1}^{12} \sqrt{x_i^2 + y_i^2}$$

Where x_i and y_i are the vectors that create scalar d_i.

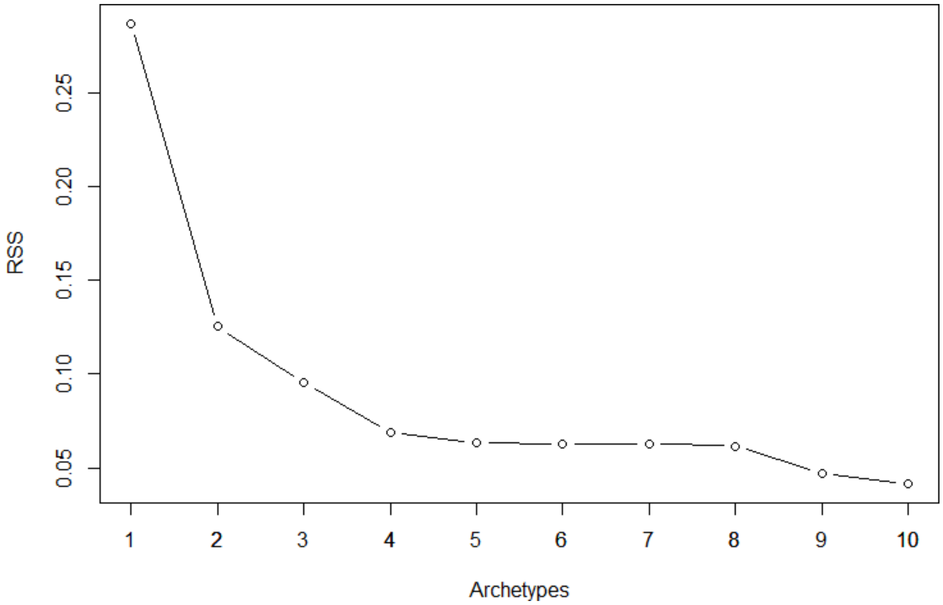
Another variable that we created was *age*. This variable was made by counting the number of repetitions a bird had and transforming them into years. In other words, if an eagle had only one row, its respective age would be 1, whereas the eagles' age in Fig. 2 would be 7 years old.

2.2 Archetypal Analysis

The fundamental idea of AA is to approximate each point in a data set as a convex combination of a set of archetypes (Bauckhage and Thurau, 2009). These are made with the `archetypes` package created by Eugster and Leisch (2009). Our first goal is to explain the variation in migratory behavior in golden eagles. We base our analysis around Archetypal Analysis (AA), an unsupervised learning approach that views each multivariate data point (bird-year in this case) as a weighted average of a set of estimated archetypes.

A simple but effective heuristic tool for choosing the number of archetypes, is the elbow criterion, and for this, we graph a scree plot. A scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA) according to Lewith et al. (2010) (see Fig. 5). The plot consists of RSS (also known as the Residual Sums of Squares) as the y-axis, and the different archetype values as the x-axis. The value of *k* (*k* will be the notation used for the number of Archetypes through the remainder of the manuscript) is selected as the point where the elbow is located (Cabero et al., 2021). The higher the *k*, the lower the RSS. However, although we want to have the lowest RSS to minimize errors and to have a more accurate model, there is a fundamental problem: overfitting. The main idea is to balance out goodness of fit with the fitted data.

Figure 3: Scree plot for archetypes on Golden Eagle data set.



Following the elbow criterion, the number of archetypes that is chosen for our data set is *k=4*. Once we pick the number of archetypes, AA will estimate the respective archetypes (α) for the data set. Since we have our desired *k*, our data points will have weights that correlate to the archetypes they are most similar to. We model every bird year's archetype weights as a Dirichlet Distribution (a continuous multivariate probability distribution with a support of $x_1, x_2, \ldots, x_n$ where $x_i \in (0,1)$ and $\sum_{i=1}^{n} x_i = 1$) and depending on the highest weight, we can classify whether the bird is a migrator and whether the bird fits in a determined archetype.

We can visualize these different classifications in Fig 4, which is a plot that shows the four different archetypes created by the unsupervised machine learning algorithm. There are three non-migratory archetypes (archetype number 1, archetype number 2, and archetype number 4), and a migratory archetype (archetype number 3).

It is worth mentioning that there are a lot of cases that belong in either Weight 2 or Weight 4. However, we can find birds that get caught in a cluster between these two weights mentioned above and are not represented by any archetype in specific. This problem is tackled in Section 2.3 considering covariates, manipulating the Archetypal Analysis fitting particular cases and changing the geographical paths of our different *k*s.

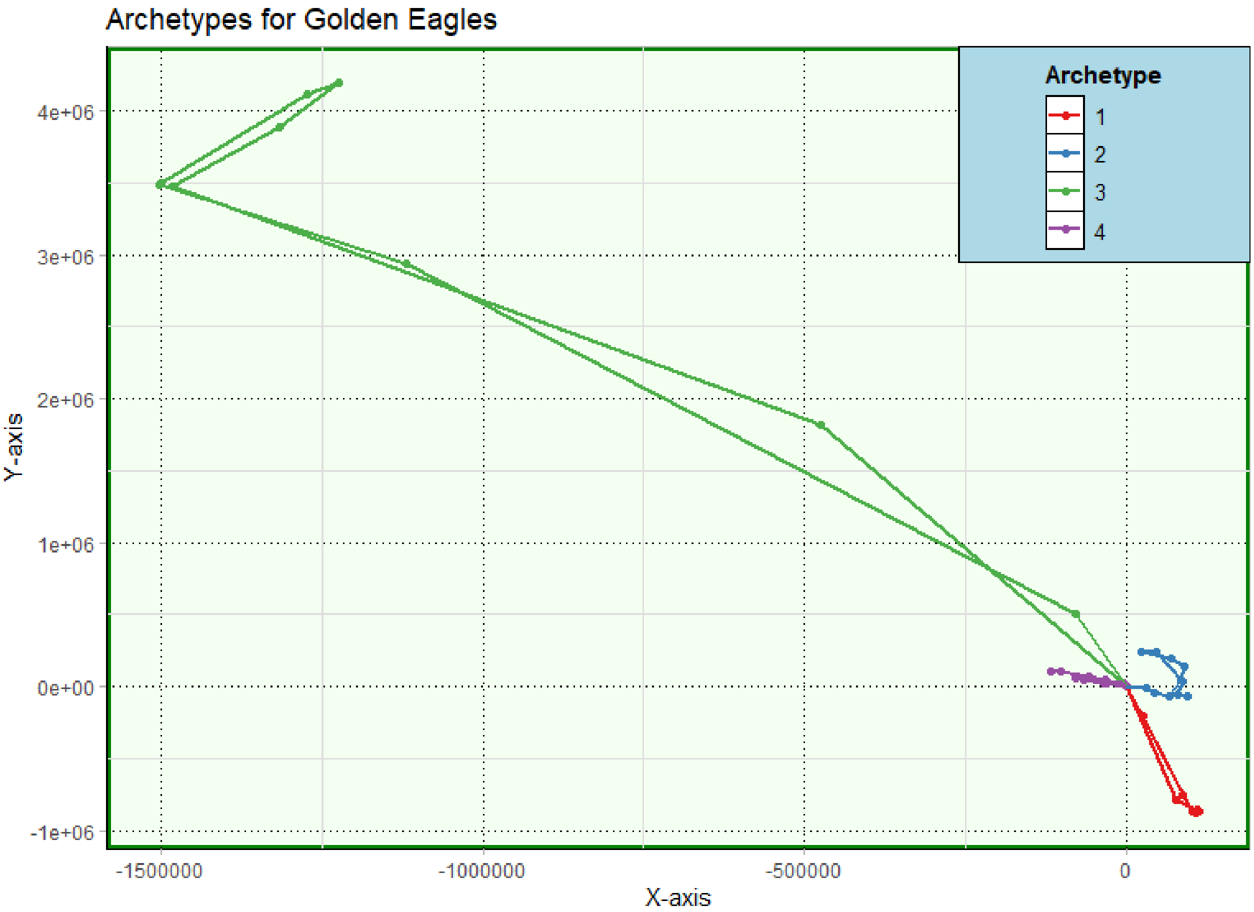


Figure 4: Monthly observations for the four different archetypes for bird-years.

2.2.1 Bayesian Inference and Hierarchical Modeling

Bayesian inference is a method of statistical inference that is gaining more popularity in the ecological field. Ellison (2004) says that Bayesian inference differs from the frequentist[1] inference in four different ways:

- Bayesian inference gives a quantitative measure of the probability of a hypothesis being true in light of the available data, whereas frequentist inference assesses the probability of the data happening given a certain hypothesis.
- Their notions of probability differ: Probability is defined by frequentist inference in terms of long-run (infinite) relative frequencies of events. In Bayesian inference, however, probability is defined as a person's level of belief in the possibility of an event.
- Prior information is used in Bayesian inference along with the sample data, whereas frequentist inference solely employs the sample data.
- Model parameters are treated as random variables in Bayesian inference, whereas they are treated as estimates of fixed, true quantities in frequentist inference.

We approach the analysis of this data set using a Bayesian Hierarchical Model (BHM). Our BHM starts with 2, where $\boldsymbol{Y}_{\mathrm{i}}$ is our data set composed by our 180 eagle with their 24 different observations (monthly observations in the form $X_{i, \ldots,} Y_i$. It has a normal distribution where the k$^{\text{th}}$ column $\mathbf{a}_{\mathrm{k}}$ of $\mathbf{A}$ is the k$^{\text{th}}$ archetype, and $\mathbf{h}_{\mathrm{i}}$ is the archetype weights for data i. σ^2 is our random parameter (standard deviation) that accounts for error.

$$\boldsymbol{Y}_{\mathrm{i}} \sim \mathrm{N}(\mathbf{A}\mathbf{h}_{\mathrm{i}}, \sigma^2),$$

The BHM starts breaking down into branches and we can see this in 3. $\mathbf{A}$ being our archetype has a normal distribution where $\mathbf{Y}$ is our data and $\boldsymbol{\omega}_k$ is the weights of our data. This altogether, form the convex hull of data $\mathbf{Y}$. τ^2 accounts for our variation as σ^2 did in 2.

$$\mathbf{a}_{\mathrm{k}} \sim \mathrm{N}(\boldsymbol{Y}\boldsymbol{\omega}_{\mathrm{k}}, \tau^2),$$

In line with the Bayesian approach, we have to specify some suitable prior distributions for all the random parameters in our model. Therefore, the parameters that we have mentioned and explained in this section, will have prior distributions with their respective support.

We have Exponential distributions for $\tau^2 \sim Exp(10)$ and $\sigma^2 \sim \mathrm{Exp}(10)$, for our variation parameters, and we have Dirichlet distributions for $\boldsymbol{\omega}_{\mathrm{k}} \sim \mathrm{Dir}(1.0)$ and $\mathbf{h}_{\mathrm{i}} \sim \mathrm{Dir}(1.0)$ for the archetype weights.

[1] also called classic, it is an approach to statistics based on a frequency view of probability in which it is assumed that it is possible to consider an infinite sequence of independent repetitions of the same statistical experiment (Everitt and Skronda, 2010)

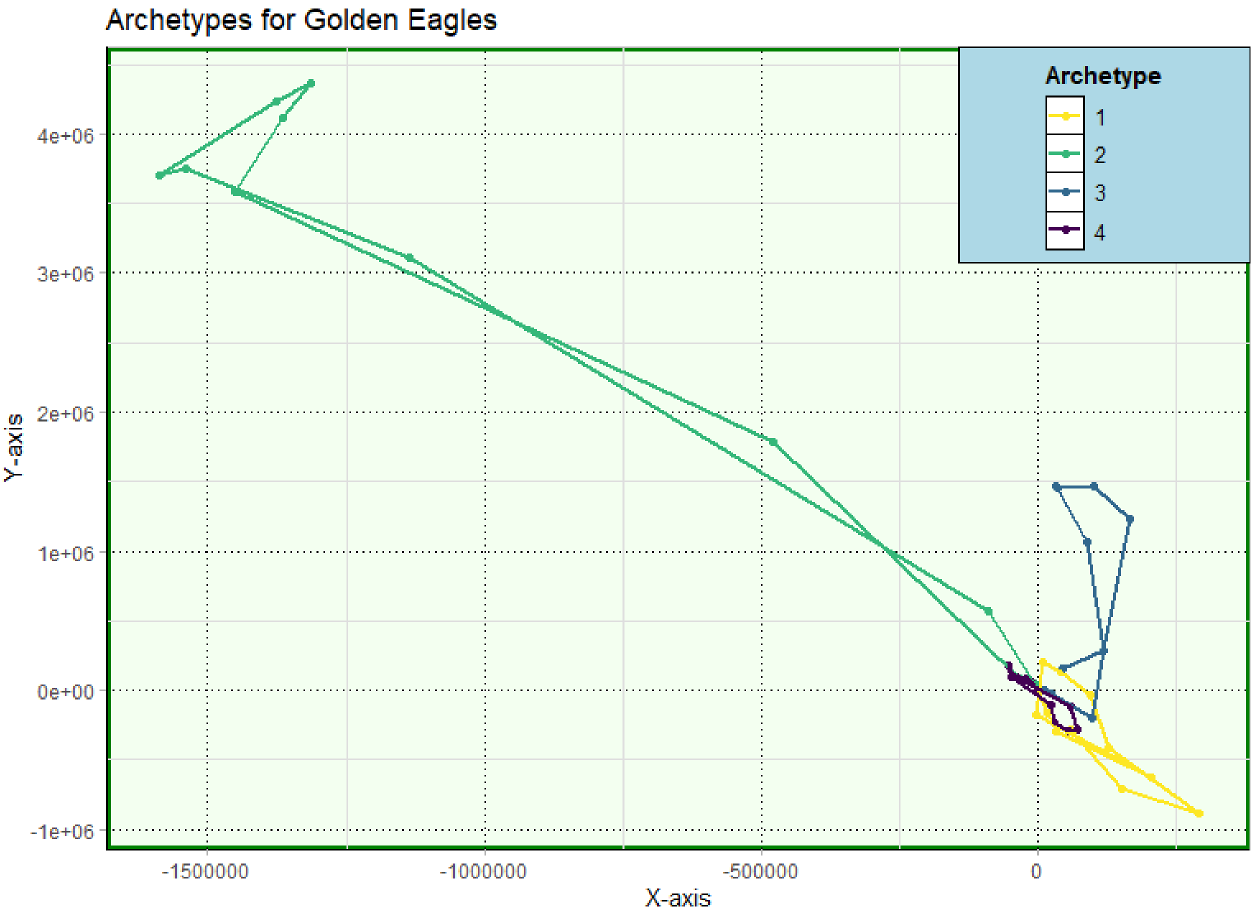


Figure 5: Monthly observations for the four different archetypes for bird-years. Note that these four estimated archetypes have different migration patterns compared to Figure 4.

2.3 Bayesian Hierarchical Modeling considering a covariate

It was mentioned previously how birds moved different as they aged, and how one of our goals was to explain how birds changed migratory behavior as they aged. For this reason, we need to change the model we built in Section 2.2.1. Recall our Bayesian Hierarchical Model: $\boldsymbol{Y}_{\mathrm{i}} \sim \mathrm{N}(\mathbf{A}\mathbf{h}_{\mathrm{i}}, \sigma^2)$. Our model has a normal distribution where the kth column $\mathbf{a}_{\mathrm{k}}$ of $\mathbf{A}$ is the kth archetype, and $\mathbf{h}_{\mathrm{i}}$ is the archetype weights for data $\boldsymbol{Y}_{\mathrm{i}}$. Then, we have $\mathbf{h}_{\mathrm{i}}$, modeled with a Dirichlet prior distribution as $\mathbf{h}_{\mathrm{i}} \sim \mathrm{Dir}(1.0)$.

Even though our Bayesian model accounted for uncertainty, we can note that it doesn't have age at all. In this Section, we will modify our previous model considering a covariate (*age* specifically), for the sake of improving model's accuracy, reducing uncertainty and accounting for our variable of interest (*age*).

We originally modeled the archetype weights as $\mathbf{h}_{\mathrm{i}} \sim \mathrm{Dir}(1.0)$, where the $\mathrm{Dir}(1.0)$ was a diffuse prior. To model the effect of *age* on migratory behavior, as captured by the archetype model, we specify a prior for $\mathbf{h}_{\mathrm{i}}$ that varies with the *age* of the individual bird.

$$\mathbf{h}_{\mathrm{i}} \sim \mathrm{Dir}(\alpha_{\mathrm{i}}),$$

where α_{ik} will be the new parameter for the prior distribution of $\mathbf{h}_{\mathrm{i}}$. This α_{ik} is modeled as

$$\alpha_{\mathrm{ik}} = \mathrm{e}^{(\mu_{\mathrm{k}}+\beta_{\mathrm{k}}\mathrm{X}_{\mathrm{i}})},$$

where μ_k is the intercept of the weight model, β_k is the coefficient for the effect of age on the weight of the $^{\mathrm{kth}}$ archetype and X_{i} is the age of the bird in bird year *i*. Recall now in Section 2.1 where it was mentioned that a variable called *age* was created based of the number of years observed each individual had.

Since parameters such as μ_k and β_k were added, we assigned their prior distributions, with $\mu_{\mathrm{k}} \sim \mathrm{N}(0{,}10)$, and $\beta_{\mathrm{k}} \sim \mathrm{N}(0{,}10)$. For clarity, we repeat our full model as follows

$$\boldsymbol{Y}_{\mathrm{i}} \sim \mathrm{N}(\mathbf{Ah}_i, \sigma^2)$$
$$\mathbf{a}_{\mathrm{k}} \sim \mathrm{N}(\boldsymbol{Y}\boldsymbol{\omega}_{\mathrm{k}}, \tau^2)$$
$$\boldsymbol{\omega}_k \sim Dir(\mathbf{1.0})$$
$$\tau^2 \sim Exp(10)$$
$$\mathbf{h}_i \sim Dir(\alpha_{\mathbf{i}})$$
$$\alpha_{ik} = e^{(\mu_k+\beta_k X_i)}$$
$$\mu_k \sim N(0{,}10)$$
$$\beta_k \sim N(0{,}10)$$
$$\sigma^2 \sim Exp(10)$$

2.4 Markov Chain Monte Carlo

We conducted inference on our BHM using Markov Chain Monte Carlo Methods. The use of Markov Chain Monte Carlo (MCMC) allows to learn about unknown elements of our model by performing numerous random draws from the posterior distributions of those unknowns conditioned on the data (Hobbs and Hooten, 2015). We implemented MCMC using the `nimble` package in `R`. We assessed convergence by visual inspection of the chains and by computing the effective sample size of each parameter. The ESS (effective sample size) for all parameters was larger than 30000.

3 Results

The estimated archetypes can be seen in Figure 6. These four archetypes were obtained with the implementation of our Bayesian Hierarchical Model with *age* (See Section 2.3). We can say that archetype 2 is relatively a non-migratory archetypes, as opposed to archetype 1, archetype 3 and archetype 4.

We were able to see the evolution of our grouping process through this paper. Firstly, we started with a purely algorithmic classification based of AA in Figure 4. Secondly, we developed a BHM and the migration routes changed significantly as shown in Figure 5. Lastly, we added *age* as it can be seen in Figure 6.

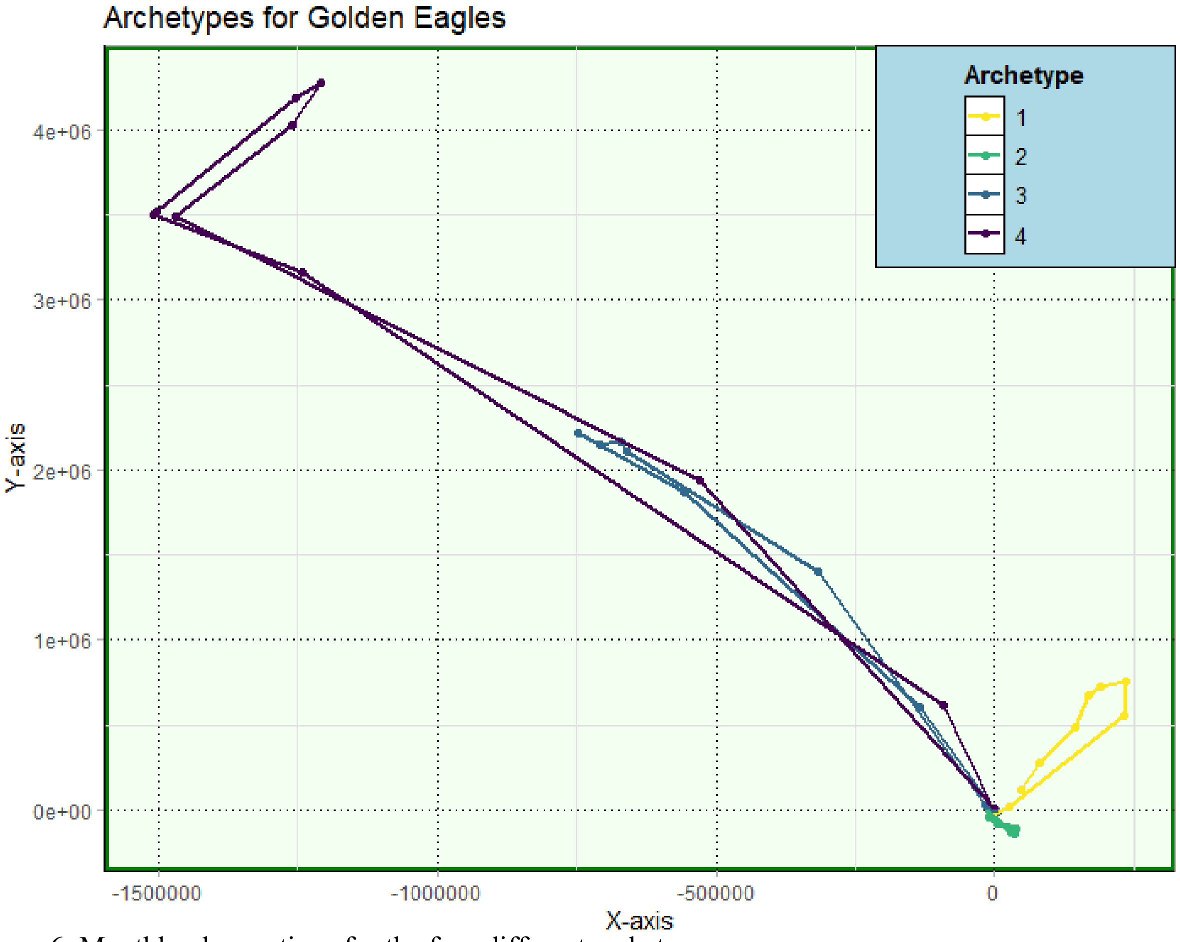


Figure 6: Monthly observations for the four different archetypes.

Once we had the different archetypes migration routes, the goal was to find if the covariate *age* was a significant factor for each bird and their respective archetype pattern.

The result of the analysis mentioned above can be shown in Figure 7. This stacked percentage bar chart illustrates how average weights corresponding to archetype number 2 increase with age, while all other archetypes, especially archetype number 3, decrease. Archetype 1, and archetype 4 decrease as well, but at a lower rate.

These percentages were obtained from the MCMC output, and from:

$$\frac{e^{\widehat{\mu_k}+\widehat{\beta_k}X_i}}{\sum e^{\widehat{\mu_k}+\widehat{\beta_k}X_i}}$$

where $\widehat{\mu_k}$ and $\widehat{\beta_k}$ are the posterior means for the effect of age on the weight of the kth archetype.

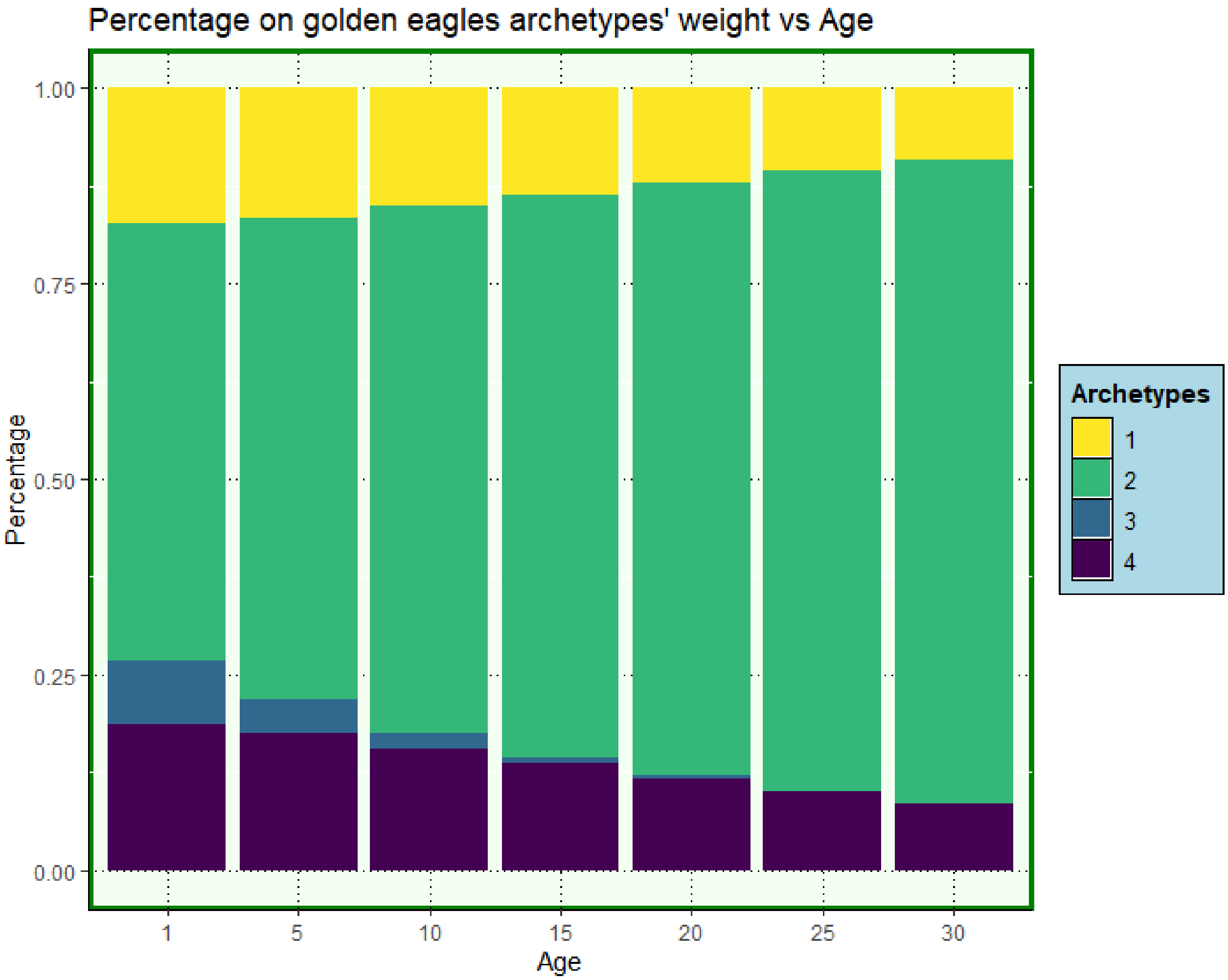


Figure 7: Stacked percentage bar chart that shows how archetype number 1 increases per year.

Archetype 2 is a non-migratory archetype. We observe how there is an increase on the weight of this certain archetype through the years. This has a constant positive trend in the lifespan of a golden eagle. This archetype also has the feature of having the highest weight for every single year. Contrary to archetypes number 1 and 4, which do not have a notorious percentage compared to the non-migratory type by the end of the 30th year. It is important to highlight that archetype number 3 is disappears and is practically non-existent by the 20th year. As shown in Figure 6, Archetype 2 is the non-migratory archetype, whereas the rest are migratory.

The main research question was whether covariates, such as age, could be a big factor on the eagles' migration routes. This question was answered in the form of a stacked bar chart (See Figure 7. The effect of the covariate *age* can be thought as a great indicator of whether the animal will belong to a determined archetype and how the animal will behave migration wise. Therefore, *age* is a big factor on their migration routes. One can conclude that the older that the golden eagles get, the higher the likelihood that they will belong to a non-migratory archetype.

4 Discussion

We analyzed a monthly golden eagle (*Aquila Chrysaetos*) location data in this paper. This exploration was made using an archetype analysis, which is a statistical nonparametric approach that represents each individual as a mixture of multiple estimated archetypes. In addition to a traditional archetype analysis, we developed a new approach to archetype analysis, in which covariates are considered, and a subset of the archetypes is defined by existing golden eagles who exhibit known, interpretable behaviors, in order to be fitted using Bayesian methods and Markov Chain Monte Carlo simulations.

It is important to highlight the use of Bayesian hierarchical modeling in animal movement data. BHM can provide reliable models that can result in straightforward and understandable insights. The fitting of *age* was trouble-free with the use of this approach.

A possible extension to our current approach could be the use of more covariates, such as sex and temperature. This could provide insight into when a golden eagle's movement is being motivated by hormonal and/or environmental factors, as opposed to a purely migrational effect with an algorithmic analysis. This could allow for models that predict birds' movement to have a better understanding of the animal behavior and have an expectation for whether the bird will migrate.

References

Bauckhage, C. and Thurau, C. (2009). Making archetypal analysis practical. In Joint Pattern Recognition Symposium, pages 272–281. Springer.

Cabero, I., Epifanio, I., Piĕrola, A., and Ballester, A. (2021). Archetype analysis: A new subspace outlier detection approach. Knowledge-Based Systems, 217:106830.

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., and Hooten, M. B. (2018). A guide to Bayesian model checking for ecologists. Ecological Monographs, 88(4):526–542.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications, 19(3):553–570.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. Technometrics, 36(4):338–347.

Ellison, A. M. (2004). Bayesian inference in ecology. Ecology letters, 7(6):509–520.

Eugster, M. and Leisch, F. (2009). From spider-man to hero-archetypal analysis in r.

Everitt, B. S. and Skrondal, A. (2010). The cambridge dictionary of statistics.

Gray, J. (2013). How animals move. Cambridge University Press.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer.

Hobbs, N. T. and Hooten, M. B. (2015). Bayesian models. In Bayesian Models. Princeton University Press.

Kéry, M. and Royle, J. A. (2020). Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and Advanced Models. Academic Press.

Lack, D. (1968). Bird migration and natural selection. Oikos, pages 1–9.

Lewith, G. T., Jonas, W. B., and Walach, H. (2010). Clinical research in complementary therapies: Principles, problems, and solutions. Elsevier Health Sciences.

Sahu, S. K. (2022).Bayesian modeling of spatio-temporal data with R. Chapman and Hall/CRC.

Appendix

The coding of this project can be found in the author's personal GitHub account.

abraham-arbelaez.github.io