Modeling Seasonality Patterns for Birds of Northern California

Cody Carroll

6/3/2020

Introduction

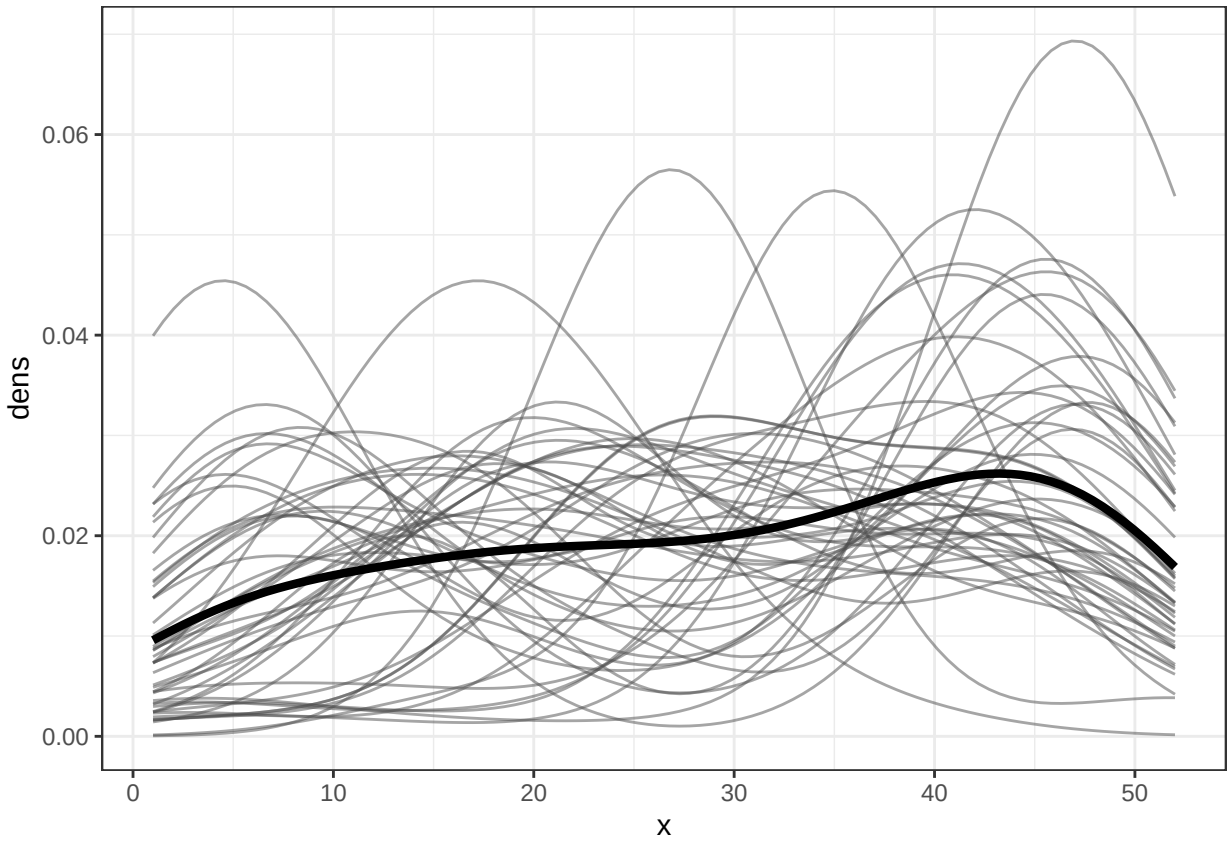
Citizen science projects like iNaturalist are becoming more credible as sources of research-grade data. While evasive species can create biases in absolute counts of observations, studies concerning within-species relative abundance may sidestep these issues by assuming that such reporting biases are uniform in time. Here we consider the relative abundance of 47 species of birds in Northern California over the 2019 calendar year. For a practical reference see for example the handbook (McCaskie and Story 1979). The raw number of observations for a given species per week is obtained from iNaturalist, accessed from their API using the R package `spocc` (Chamberlain, Ram, and Hart 2018; *INaturalist. Accessed May 28,* 2020). Seasonality curves are then constructed from iNaturalist observations using techniques from kernel density estimation. The resulting seasonal trends may then be viewed as a sample of random curves (more specifically, probability densities) and can be studied under the framework of Functional Data Analysis (Wang, Chiou, and Müller 2016). We will use functional principal component analysis (Yao, Müller, and Wang 2005) to model the relative abundance of these species and identify a low-dimensional latent geometry which can be used to group similar bird species by their patterns of seasonality.

From Points to Curves

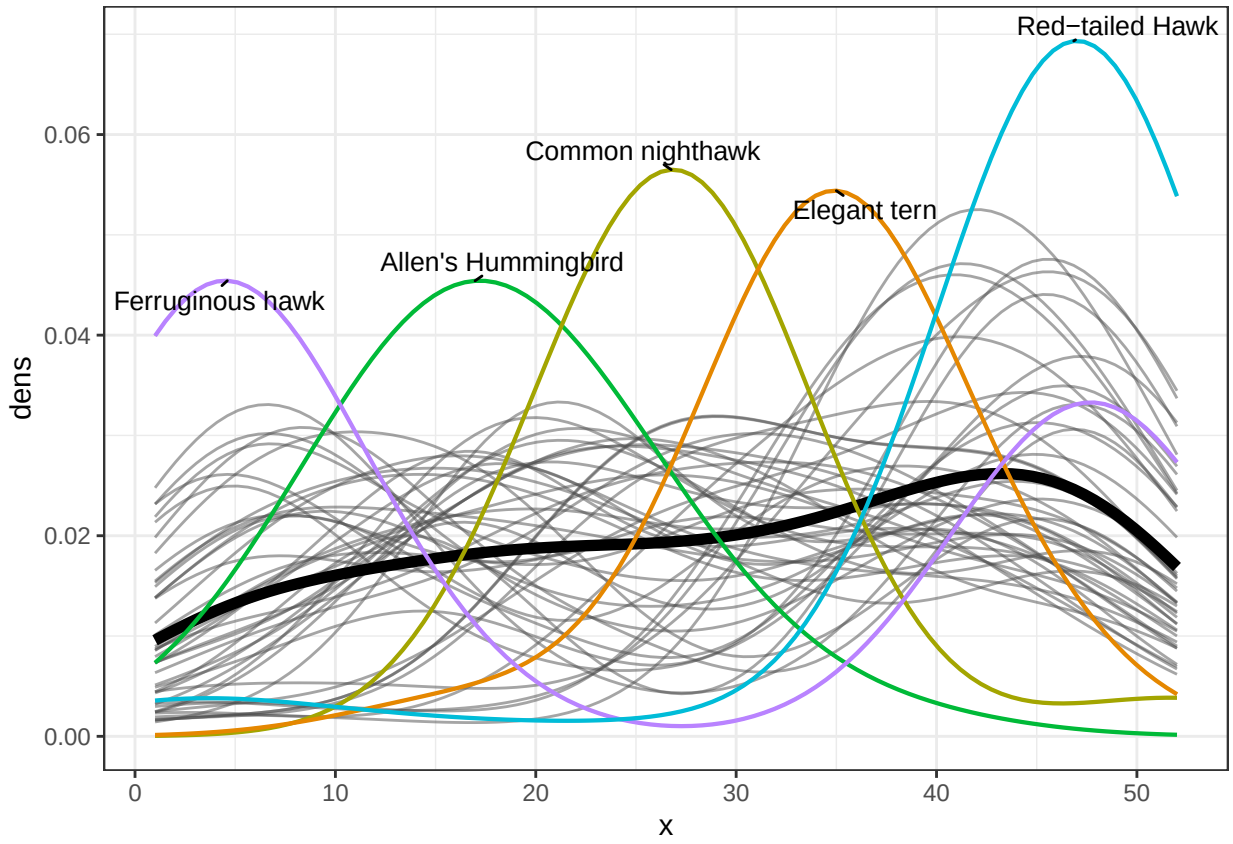
For the i^{th} species during the j^{th} week of the year, we observe the number of observations reported on iNaturalist, $x_{ij},\ i = 1, ..., 47,\ j = 1, ..., 52$. Using a Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ we obtain estimates for the seasonality curves given by:

$$f_i(t) = \frac{1}{52h} \sum_{j=1}^{52} K(\frac{t - x_{ij}}{h}),$$

where the bandwidth h is a tuning parameter which modulates smoothness of curves and is chosen here to be 6 weeks based on visual inspection. The estimated seasonality curves for the 47 birds are shown below.



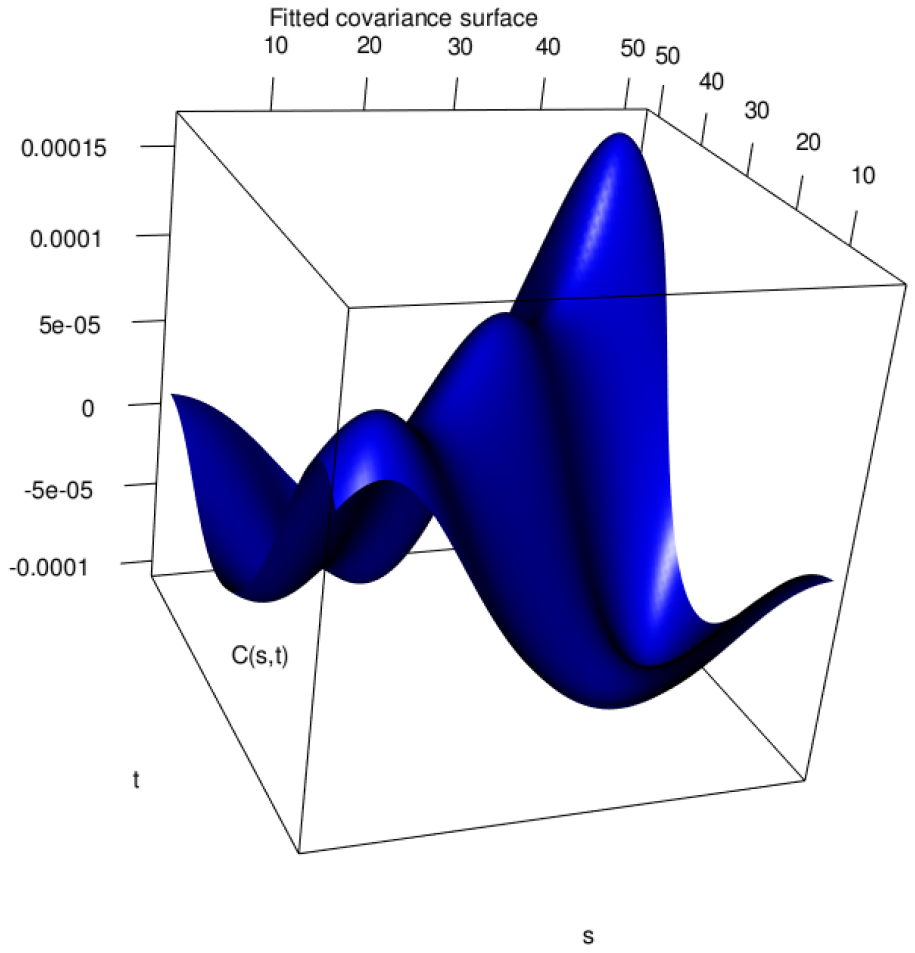
The bold black curve denotes the pointwise mean seasonality, which is roughly constant with a slight skew toward the later weeks of the year. This suggests the sample contains birds with a variety of seasonalities, though perhaps fall and winter birds are slightly over-represented. Note the 5 "outlier" curves which stand out above the crowd. These curves represent species which demonstrate extreme seasonality for each season, with winter being split into "early winter" and "late winter." This is the first sign of an interesting geometric phenomenon present in this data: we expect one representative peak per season, but because the calendar year artificially splits the winter months, the corresponding winter peak is divided into two. The Ferruginous hawk's seasonality exemplifies this: it has a global maximum during the early weeks, but experiences another local maximum toward the end of the year. If we could recognize the underlying geometry on which these curves lie, a kernel density estimate would join these two peaks into one! We will investigate this geometry further using functional principal component analysis, which will also perform dimension reduction on the curves and facilitate comparisons in seasonality patterns across species.



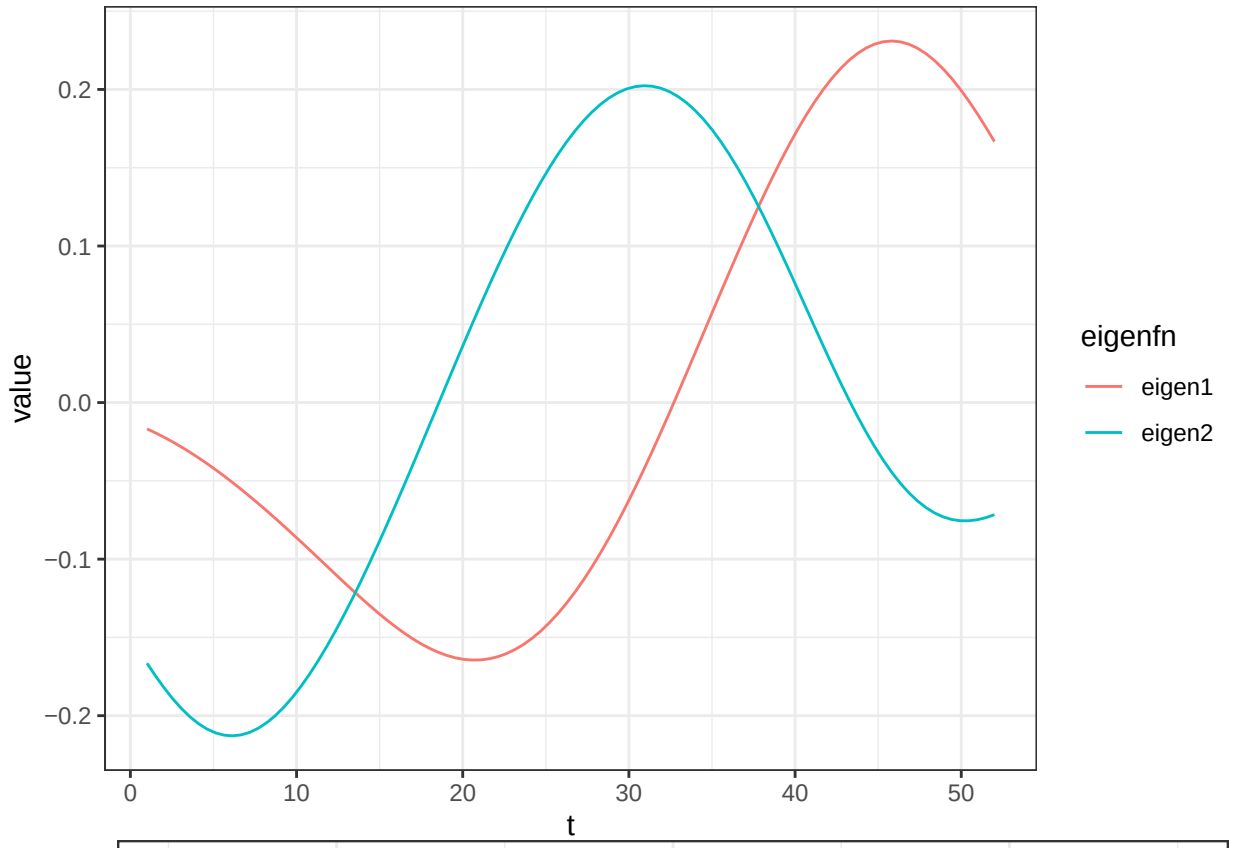
Functional Principal Component Analysis

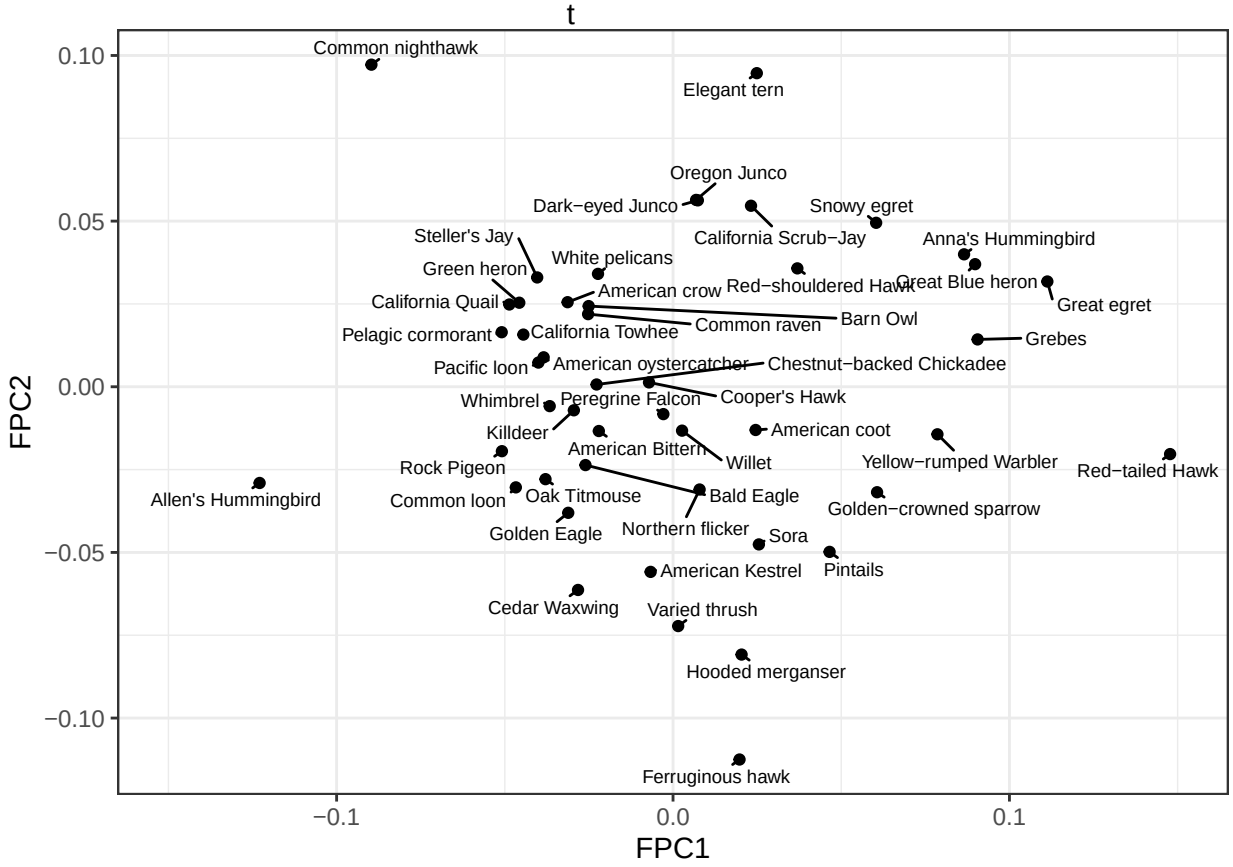
While a curve-based representation of seasonality is intuitive and easy to read for an individual species, understanding and comparing the trends of several species at a time is more difficult. It is impossible to easily read off the species for every curve in the crowded figure above. This difficulty can be attributed to the high-dimensional nature of curves: a huge number of individual data points make up a single curve observation. Here each curve is approximated with by interpoling 100 data points generated from the kernel density estimate. Variation in swaths of these dimensions may come together to form major trends which have meaningful interpretations, or may simply constitute random noise and should be considered as nuisance error. To understand the difference and identify the main patterns of variation in the curves, functional principal component analysis (FPCA) can afford us a low dimensional representation of the curves based on the Karhunen-Lo'eve decomposition for stochastic processes (Castro, Lawton, and Sylvestre 1986). We describe the technical machinery in the appendix and display the results of FPCA here. Statistical computing was performed using the `R` package `fdapace` (Carroll et al. 2020).

The first step in FPCA is to estimate the covariance surface for the sample of curves. The undulating shape of the surface indicates that prevalence is typically negatively correlated across opposite seasons, e.g. birds with high summer seasonality tend to have low winter presence. To see this, fix a time point t and take a cross-section of the surface in either direction. The maximum and the minimum are roughly 25 weeks or 6 months apart from each other for all cross-sections. The cyclic nature of the seasons is suggested by the positive covariance on the edges of the off-diagonal. Here the covariance rises from the negative trough back into positive values in the corners, which suggests that January counts are positively associated with December counts. An intuitive way of visualizing this observation is to consider tesselating the surface across the plane: the edges of the surface align roughly across tesselated copies!



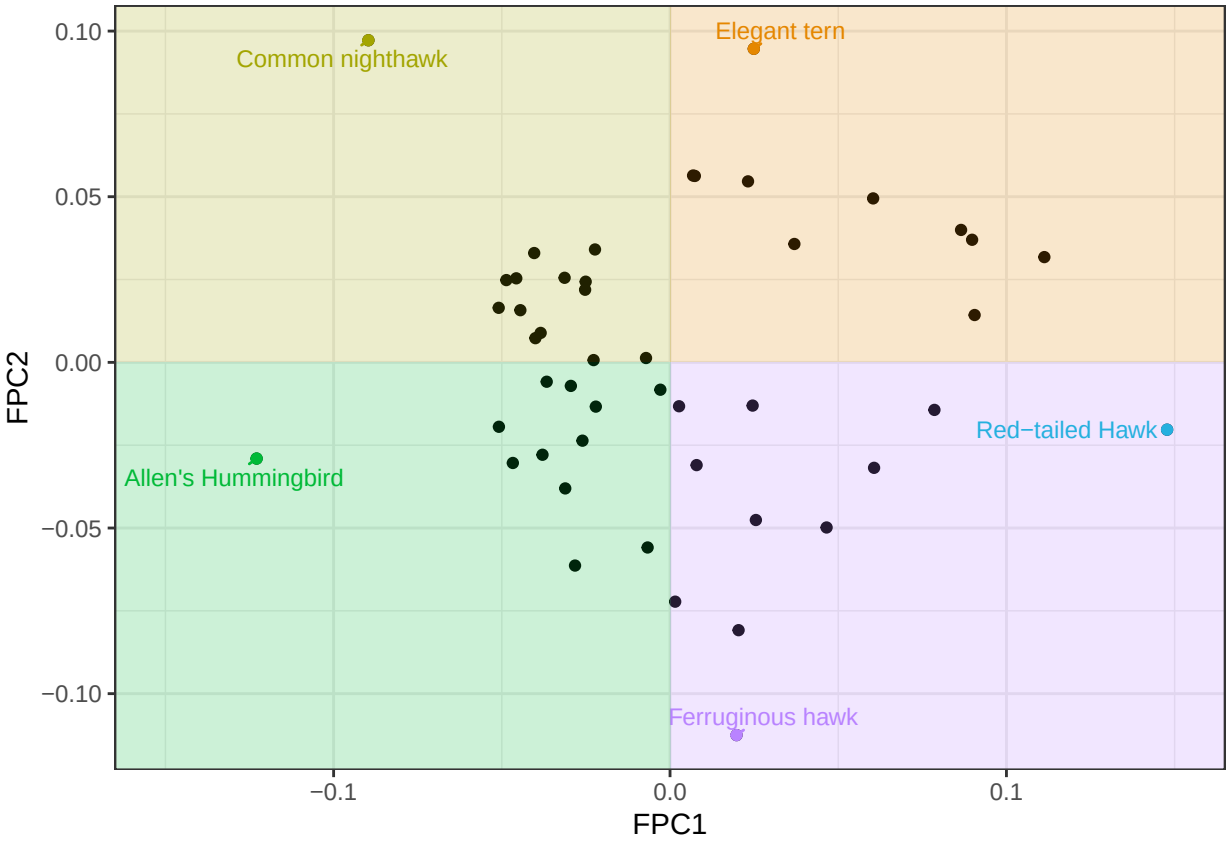
Using the functional analogue of spectral decomposition we can obtain the eigenfunctions of this surface and their corresponding scores for each individual species. The results are displayed below using a 2-dimensional truncation which captures 90% of the original variation. The first eigenfunction represents a contrast between late spring and early winter: species with high FPC1 scores will be more common in early winter and less prevalent in spring. The second eigenfunction follows a similar shape but is shifted in time: species with high FPC2 score will be more prevalent during summer and less common during late winter. In FPC space, we can group species together much more clearly than we could using only their curve representations. For example, the Oregon and Dark-Eyed Juncos exhibit very similar patterns of seasonality: from their high FPC2 scores we can understand that these birds exhibit late summer/early fall seasonality. To see the connection between specific seasons and FPC scores more explicitly, we showcase a few representative birds below.





We recognize the species on the edge of the point clouds as the species which had the outlying curves

mentioned before. They represent archetypal trends for each season. We approximate each season's region in FPC space with a colored quadrant: green corresponds to spring, yellow to summer, orange to fall, and purple to winter. Here we really see the underlying geometry in action: the flow of time corresponds to a clockwise rotation through FPC space. Starting at the bottom with the Ferruginous hawk and sweeping out clockwisely, we pass through all of the seasons before returning back to winter. Another striking observation is that the observed point cloud forms a ring, mimicking the circular interpretation of seasonality. This suggests that the birds in our sample are highly seasonal with only a few showing constant levels of prevalence throughout the year. This geometry motivates an alternative representation of the curves: a more authentic representation of the data would plot the original set of seasonality curves on the surface of a cylinder, instead of a flat plane, in order to mirror the cyclic behavior and avoid the discontinuity created by the calendar year.



Future Directions

While I originally started this project with the goal of comparing iNaturalist and eBird data, I encountered difficulty in accessing eBird's API through `spocc`. A worthwhile follow up would be to replicate this analysis using eBird data to see whether the circular geometry in FPC space remains intact. Other functionally-minded directions for the analysis of seasonality curves may map the densities into unconstrained L^2 space with the log-quantile density transformation (Petersen and Müller 2016) before performing FPCA, which may yield a more parsimonious representation, or consider a time-warping based curve alignment approach where estimated warping functions may be used to cluster seasonalities (Marron et al. 2015). Other uses for this framework may concern phenological trends: the highlighted quadrants may be more finely tuned to better reflect the calendar date-based boundaries of each season before being used for comparisons across geographic regions.

It's also worth mentioning that the sample of 47 birds considered here is also by no means complete: further research may extend this analysis to a more exhaustive survey of Northern Californian birds.

Appendix

Functional PCA

We consider a generic seasonality curve $f(t),\ t \in \mathcal{T} = [1, 52]$ with mean curve $\mu(t) = E(f(t))$ and covariance surface $G(s,t) = \text{Cov}(f(s), f(t))$, which has eigenfunctions $\varphi_1(t), \varphi_2(t), \ldots$. The Karhunen–Lo'eve representation theorem states that

$$f(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \varphi_k(t), \tag{1}$$

where the scores $\xi_k = \int_{\mathcal{T}} (f(t) - \mu(t))\varphi_k(t)dt$ satisfy $E(\xi_k) = 0$, $\text{Var}(\xi_k) = \lambda_k$ and $E(\xi_k \xi_l) = 0$ for $k \neq l$. Here ξ_k is the functional principal component score (FPC) of $X(\cdot)$ associated with the k^{th} eigenfunction φ_k. We can view FPC scores as projections of the stochastic process onto the directions described by the eigenfunctions.

By truncating the vector representation to a finite number of K components one reduces the infinite dimensionality of the curve and can approximate the original stochastic process through its most important modes of variations. That is, FPCA provides a fit of the original curve

$$\hat{f}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_k \hat{\varphi}_k(t) \tag{2}$$

References

Carroll, Cody, Alvaro Gajardo, Yaqing Chen, Xiongtao Dai, Jianing Fan, Pantelis Z. Hadjipantelis, Kyunghee Han, Hao Ji, Hans-Georg Mueller, and Jane-Ling Wang. 2020. *Fdapace: Functional Data Analysis and Empirical Dynamics*. https://CRAN.R-project.org/package=fdapace.

Castro, PE, WH Lawton, and EA Sylvestre. 1986. "Principal Modes of Variation for Processes with Continuous Sample Curves." *Technometrics* 28 (4): 329–37.

Chamberlain, S, K Ram, and T Hart. 2018. "Spocc: Interface to Species Occurrence Data Sources. R Package, Version 0.8. 0."

INaturalist. Accessed May 28, 2020.

Marron, James Stephen, James O Ramsay, Laura M Sangalli, and Anuj Srivastava. 2015. "Functional Data Analysis of Amplitude and Phase Variation." *Statistical Science*, 468–84.

McCaskie, Guy, and Nick Story. 1979. *Birds of Northern California: An Annotated Field List*. The Society.

Petersen, Alexander, and Hans-Georg Müller. 2016. "Functional Data Analysis for Density Functions by Transformation to a Hilbert Space." *The Annals of Statistics* 44 (1): 183–218.

Wang, J-L, J-M Chiou, and H-G Müller. 2016. "Functional Data Analysis." *Annual Review of Statistics and Its Application* 3: 257–95.

Yao, F, H-G Müller, and J-L Wang. 2005. "Functional Data Analysis for Sparse Longitudinal Data." *Journal of the American Statistical Association* 100 (470): 577–90.