Quantifying song categories in Adelaide's Warbler (*Setophaga adelaidae*)

Chinthaka D. Kaluthota[1,4], Orlando J. Medina[2], David M. Logue[1, 3]

[1] Behaviour and Evolution Research Group, Department of Psychology, University of Lethbridge. 4401 University Drive West, Lethbridge, AB T1K3M4, Canada. (ORCID: 0000-0001-8151-4289)

[2] Cabo Rojo National Wildlife Refuge. United States Fish and Wildlife Service. PO Box 510, Boqueron, Cabo Rojo, PR 00622-0510, USA

[3] Departamento de Biología, Universidad de Puerto Rico. Mayagüez PR, 00682, USA

[4] Corresponding author email: kaluthota@uleth.ca, Phone: +1 (403) 715-8027

Abstract

Many migratory wood-warblers in the genus *Setophaga* divide their song repertoires into two categories. Category B songs are usually sung before dawn, with immediate variety and short latencies between songs, whereas category A songs are sung exclusively after dawn, with eventual variety and longer latencies between songs. Songs in different categories may also differ with respect to their acoustic structure. We used an unsupervised clustering algorithm to identify song categories in Adelaide's Warbler (*Setophaga adelaidae*), a year-round territorial species. We identified two categories of song types, the characteristics of which are similar to song categories in other migratory wood-warblers. Clusters were not well-separated, suggesting that song categories may not be discrete. Song structures in the two categories were similar, but category B songs were shorter and had fewer notes than did category A songs. On average, dyads of males shared more category B songs than category A songs, and were more likely to use category B songs when song type matching other males. The most important song delivery variable for separating clusters was residual average run length (residual values control for covariation with time of day), followed by percent of songs delivered before dawn, residual latency, and percent of songs used as song-type matches. We recommend a scheme based on the first three variables to classify novel song types.

Key words: Adelaide's Warbler, Neotropical birds, singing modes, song repertoires, song types, year-round territoriality

Acknowledgments

We are grateful to P. Sánchez-Jaureguí, F. L. Tarazona, J. Illanas, K. Medina, and A. Lamela for helping with data scoring. We also thank B. Parker, T. Shlakoff, C. Logue, J. Báez, R. Irizarry, and A. García who processed song recordings, and J. McClure who measured frequency excursion. P. Linhart, D. Spector, and an anonymous referee offered valuable comments on an earlier version of this manuscript. This research was funded in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-06553) to D.M.L. The U.S. Fish and Wildlife Service granted permission to work at the Cabo Rojo Wildlife Refuge (permit 2012-01). Protocols were approved by the Institutional Animal Care and Use Committee at the University of Puerto Rico, Mayagüez (permit 2010 09 17).

Conflict of interest: The authors declare that they have no conflict of interests

In many songbird species, individuals can produce more than one song type. The set of song types that an individual can sing is called a repertoire. Song type repertoires are hypothesised to function in both mate attraction and in competition with rival conspecifics (Catchpole & Slater, 2008). Some species divide their song repertoires into distinct categories (Illes, 2015; Molles & Vehrencamp, 1999; Spector, 1992). This pattern is especially common in wood-warblers belonging to the genus *Setophaga,* many of which use two song categories (Family: Parulidae; Bolsinger, 2000; Demko et al., 2013; Price & Crawford, 2013; Spector, 1992; Wiley et al., 1994). The two song categories may be characterized by distinct song delivery patterns, acoustic structures, and / or behavioral contexts (Spector, 1992). Throughout this report, we follow Spector's (1992) recommendation to refer to these categories as "first category" and "second category" when referencing multiple species, and to use pre-existing terminology when discussing individual species (e.g., "category A" and "category B," in the case of our focal species). Many, but not all, *Setophaga* wood-warblers deliver first category songs in a repetitive manner, with relatively long silent gaps between songs, during the daytime, and in the context of male-female interactions (Spector, 1992). By comparison, second category songs are typically delivered with immediate variety, at higher singing rates, mostly around dawn and in the context of male-male interactions (Beebee, 2004b; Kroodsma et al., 1989; Price & Crawford, 2013; Spector, 1992; Staicer, 1989, 1991, 1996a).

In addition to differences in delivery patterns and behavioral context, songs from the two categories can be structurally distinct. In American Yellow Warblers (*Setophaga petechia*), for example, type I (first category) songs have higher trill rates and wider frequency bandwidths than type II (second category) songs (Beebee, 2004b). Similarly, Price and Crawford (2013) showed that Pine Warblers' (*Setophaga pinus*) first category songs have significantly higher trill rates than their second category songs.

The existence of song categories in wood-warblers is a foundational example of functional variation in song type repertoires (Catchpole & Slater, 2008; Macdougall-Shackleton, 1997). Most published research on song categories in wood-warblers, however, has come from North-temperate, migratory species. These studies may not represent the many wood-warbler species that breed at other latitudes because the vocal behavior of tropical and southern-hemisphere songbirds is often different from that of North-temperate species (Curson et al., 1984; Stutchbury & Morton, 2001). In the present study, we evaluate the evidence for two distinct song categories in a year-round territorial, tropical wood-warbler.

Adelaide’s Warbler (*Setophaga adelaidae*) is endemic to Puerto Rico and the nearby island of Vieques (Toms, 2010). Socially monogamous pairs defend all-purpose territories throughout the year. Previous studies of this species conclude that males have two distinct song categories (Staicer, 1991, 1996a, 1996b). Those studies indicate that males deliver category A (first category) songs after dawn, throughout the year, with eventual variety, and relatively long intervals between songs. Category B (second category) songs occur primarily before dawn, during the breeding season, with immediate variety, and relatively short intervals between songs. Category B songs are also reported to occur later in the day, when they are still delivered with immediate variety and short intervals (although intervals are longer and variety is lower than during pre-dawn singing). Males tend to use Category A songs when interacting with females and Category B songs when interacting with other males.

Although the categorization of songs into two distinct modes is easily achieved for some wood-warbler species, Staicer (1996b) found that Adelaide’s Warbler songs could not be easily categorized because the acoustic properties of songs in the two categories overlap. Song categorization is further complicated by the fact that different males can assign a given song type to different categories (throughout the manuscript we use the term “assign” to indicate that a bird delivers a song in a way that is characteristic of a given category). For example, one male could assign song X to category A, but a different male could assign song X to category B. This type of individual-specific usage of song types has also been observed in some other wood-warbler species (Beebee, 2004b; Lemon et al., 1985; Price & Crawford, 2013; Spector, 1991; Staicer, 1989).

Staicer (1991, 1996a) used song delivery patterns to categorize songs. In one study, songs sung before dawn (hereafter, “dawn” after Staicer, 1991) and after dawn (“morning”) were taken to represent category B and A songs, respectively (Staicer, 1996a). In another study, Staicer (1991, p. 21) describes a slightly different categorization scheme: “Songs that males sang in the dawn chorus were called B songs and those that males switched to around sunrise were called A songs…. If I recorded a song type only after sunrise but at a rate typical of B songs, and usually in the same bout (consecutive song sequence) with B songs, I classified it as a B song” (a similar description can be found in Staicer, 1996b).

Staicer’s categorization schemes succeeded in elucidating important patterns in Adelaide’s Warblers’ vocal communication system (Staicer, 1996a, 1996b). Nevertheless, we see room for improvement. The first approach described in the previous paragraph has the potential to miscategorise category B songs, which are reported to occur both before and after dawn. The second approach improves on the first, but relies on the analyst to determine typical rates of category B songs and whether a given song is sufficiently linked to other category B songs. We had several motivations to develop a different approach to song categorization. First, we wanted to test whether Adelaide's Warbler repertoires are best divided into two categories, as opposed to one category, three categories, etc. Second, we were interested in which song delivery variables were most useful for categorizing songs. Third, we wanted to develop an algorithm to categorize novel songs (see Chapter 3 of Hastie & Dawes 2010 for a discussion of the benefits of automatic judgement algorithms). Finally, we wanted to compare our results to Staicer's (1991, 1996a, b) results because replication with novel data and a novel analytic approach is a critical, but underutilized, component of the scientific method (Baker, 2016).

The present study examines song categories in male Adelaide's Warbler. We applied an unsupervised clustering algorithm to song type repertoires. Specifically, we ran cluster analyses for all fifteen combinations of four song delivery variables believed to be important for distinguishing song categories. We chose the best number of clusters (k) from each analysis, and then the best analyses based on a given number of variables. We evaluated the four remaining clustering schemes with respect to the number of song categories, their characteristics, their discreteness, and their distributions across individuals. Finally, we looked at patterns of song sharing and compared song structures across categories.

Methods

Field data collection

We recorded nine mated male Adelaide's Warblers at the Cabo Rojo National Wildlife Refuge in south-western Puerto Rico (17º 59' N, 67º 100' W) during the breeding season between March and June, 2012. Males were captured using mist nets and marked with three colored leg bands and a numbered metal band prior to recording.

Individual males were recorded continuously for approximately 3.5 hours per day, for four days each. Consecutive recordings of a given male were separated by at least four days, except on two occasions when recordings were made on consecutive days because of logistical constraints. Observations started 45 minutes before sunrise to ensure that we recorded the first song of the day. Recordists announced the singer's identity after each song, as well as song type matches and fights with neighboring conspecifics. The identity of focal males was confirmed by inspecting the colored leg bands prior to the end of each recording session. Recordings were made with Marantz PMD 661 digital recorders and Sennheisser ME67 shotgun microphones (file format = wav, sampling rate = 44.1 kHz, bit depth = 16 bits). We sampled a large number of songs from a small number of individuals. The decision to emphasize recording effort per individual resulted in increased accuracy of singing variables at the song type within individual level, but the small sample of individuals may have limited our ability to fully capture the range of variation among individuals. This is the same set of recordings used in Schraft et al. (2017) and Hedley et al. (2018).

Scoring and acoustic analysis

We inspected sound spectrograms of all recordings in Syrinx PC v2.6f sound analysis software (John Burt, http://www.Syrinxpc.com; Blackman window; window size = 1024 points). We assigned songs to song types based on their appearance on a spectrogram. Several observers assigned songs to types as they scored the field recordings for each male. Then, one person (D.M.L.) scored song types across males (i.e., decided which songs from different males' repertoires belonged to the same type), and corrected scoring errors. We measured the inter-rater reliability of song type scoring within an individual bird. We randomly selected 100 songs from one individual, and two experienced bio-acousticians used a classification key to score them independently. The result was 100% agreement. This analysis demonstrates that different observers agree on how to score song types within males. Most of the analyses in the present study use "song type within male" as the independent sampling unit, so it is important to verify that song type scoring within male is repeatable across observers. We then estimated the repeatability of scoring song types among individual birds by having an experienced bio-acoustician (C.D.K.) re-classify 22-23 randomly selected songs from each of the nine males (total = 200 songs) using the population-level classification key. In this analysis, 175 of 200 (87%) scores matched the originals. This number is lower than the within-individual repeatability because song structure can vary among individuals. Among-individual repeatability is relevant to the song type sharing analyses in the present study.

Song recordings with high signal-to-noise ratios, as determined by visual inspection of spectrograms, were subjected to detailed acoustic analysis. We used Luscinia v.2.14 (Lachlan, 2007) sound analysis software to obtain the minimum and maximum peak frequency, number of notes, and song duration for high signal-to-noise-ratio song recordings from the breeding season dataset (settings: maximum frequency = 10 kHz, frame length [equivalent to "bin" or "FFT" length] = 5 ms, time step = 1 ms, dynamic range = 35 dB, dynamic equalization = 100 ms, de-reverberation = 100%, de-reverberation range = 100 ms, high-pass threshold = 1.0 kHz, noise removal = 10 dB). In

Luscinia, users identify focal signals by outlining their images on a sound spectrogram. Sounds in the outlined area that exceed user-defined thresholds for amplitude and duration are labeled with a colored trace that users can compare to the spectrogram to correct errors. The program stores acoustic information about the signal (as defined by the trace) in a database, which users can query (e.g., for the minimum peak frequency for each note). Minimum frequency was subtracted from maximum frequency to calculate the frequency bandwidth of each note. The frequency bandwidth of each song was defined as the average frequency bandwidth of its notes. Trill rate was calculated as note number / song duration (sec). We measured the frequency excursion (FEX) of the same songs using the program FEX Calculator (Podos et al., 2016, J. McClure, https://github.com/BehaviorEnterprises/Fex). FEX is a putative metric of vocal performance that measures changes in fundamental frequency, including changes that are not voiced, over time.

We performed a series of unsupervised cluster analyses to categorize songs. We treated "song type within individual" as the independent sampling unit, because different males may assign individual song types to different categories (Staicer, 1991). There was a high risk of misclassifying rare songs types, so we omitted song types within individual that were recorded 10 times or fewer (10 was chosen as a cut-off point because it is a round number). Clustering was based on variables linked to four of the five attributes hypothesized to separate song categories (Table 1): mean residual latency (a measure of the interval between songs), mean residual run length (a proxy for delivery mode), percent matching (a proxy for social context), and percent of songs sung during the pre-dawn period.

Latency was defined as the time since the focal male's prior song. Latency co-varies with time of day (Staicer 1991). To better isolate latency from time of day, we regressed latency against time and used the residual latency to generate means (logistic regression: $r^2 = 0.387$, $F_{1, 8998} = 5684$, $p < 0.001$, constant = 0.76, beta = 1.0). Run length was the number of songs in a continuous run of a given song type (we only used a single run length value for each run to avoid pseudoreplication). Like latency, run length was correlated with time of day, so we regressed run length against time and used the residual latency to generate means (logistic regression: $r^2 = 0.159$, $F_{1, 5933} = 1124$, $p < 0.001$, constant = 0.84, beta = 1.0).

We chose song type matches by the focal males as our measure of vocal interaction, because it was less likely to occur by chance than other kinds of vocal interaction (e.g., song overlapping, unmatched counter-singing), and so more likely to represent a deliberate interaction on the part of the focal male. Song type matching was scored when a focal male sang the same song type as a neighbor had sung in the previous two seconds. Two seconds was chosen as the cut-off for song type matching because it corresponds to the average duration of a male song (mean = 2.0 sec., n = 2776). Matching was scored if the recordist dictated "song type match" or if examination of sound spectrograms revealed a match.

To calculate the percent of songs delivered before dawn, we first obtained sunrise time from the website *www.timeanddate.com* and scored "time to sunrise" (in seconds) for each song. Negative values of this variable correspond to times before sunrise, and positive values correspond to times after sunrise. We defined the end of the dawn chorus as the time that song rates stabilized after the period of intense pre-dawn singing. Two of us (C.D. K. and D.M.L.) visually inspected a histogram of song delivery times, and determined that 700 seconds after sunrise was the optimal cut-off time. Following (Staicer, 1991), we refer to the period before this cut-off as "dawn," and the period after the cut-off as "morning". We controlled for sampling effort by dividing the proportion of dawn songs from the focal male that belong to the focal song type by the proportion of all songs from the focal male that belong to the focal song type, and multiplying by 100. All data were standardized by subtracting the mean and dividing by the standard deviation prior to cluster analysis.

We ran the *NbClust* function in the R package *NbClust* to cluster the data (Charrad et al., 2014). We chose the k-means analysis with Euclidean distance because it is a simple and effective approach to unsupervised clustering. We ran separate analyses for each of the 15 unique combinations of the four clustering variables (Table 2). For each analysis, we interpreted the number of clusters that produced the highest average silhouette index. The average silhouette index measures the average similarity of objects to other objects in their own clusters relative to objects in other clusters (Rousseeuw, 1987). The index ranges from -1 to +1, where higher values indicate better clustering. We used the silhouette index to choose the best categorization schemes because it is a simple, widely-used metric that reflects our intuitive concept of clustering. At this point, we had 15 clustering schemes (one for each unique combination of variables). We then identified the clustering scheme with the highest average silhouette index for each number of clustering variables, one through four. We were curious whether clustering strength was influenced by sample size, so we calculated separate average silhouette indices for well-sampled (n > 40 songs) song types and poorly-sampled (n ≤ 40 songs) song types for each of these four clustering schemes.

We compared the assignment of song types to categories across males to test the hypothesis that different individuals assign the same song type to different categories. We then tested whether song type sharing differed between song categories. First, we identified song types that were shared and assigned to the same category for all possible pairs of males. We then conducted a paired t-test in Microsoft Excel 2013 (Microsoft Corp., Redmond, WA). This test treated dyads of males as the independent sampling units. It tested the null hypothesis that the mean difference between the number of shared category A songs and shared category B songs shared was zero.

To test whether there were structural differences between song categories, we developed mixed models with the following structural properties as dependent variables: frequency excursion (FEX, unitless), song duration (ms), number of notes, trill rate (notes / sec), minimum frequency (kHz), maximum frequency (kHz) and average frequency bandwidth (kHz; averaged over the notes in the song). The sampling units were individual song utterances, and the dataset was restricted to the high-quality song recordings that were suitable for fine-scale structural analysis. For the structural models, song category was included as a fixed factor, and individual, recording day nested within individual, and song type within individual (not a statistically nested variable) were included as random factors. We chose to include song type in this model because structural variables (but not song delivery variables) are inextricably linked to song type, and therefore different utterances of a given song type are not independent with respect to acoustic structure. Mixed models were constructed with the *lmer* function of the package *lme4* (Bates et al., 2015). We examined residual plots to evaluate model fit. All mixed models are in the online resource.

Results

The breeding season recordings of nine focal males included 9,420 songs comprising 71 song types. Summing all males' repertoires, there were 261 song types within individual (average ± SD = 29.0 ± 4.0 song types per male). Removing uncommon song types (song types recorded 10 times or less within individual) resulted in 9032 song utterances comprising 57 song types and 168 song types within individual (18.7 ± 2.4 song types per male). Of those, 2776 song recordings were of sufficient quality for structural analysis.

The best clustering schemes for all fifteen combinations of song delivery variables are described in Table 2. Most of these schemes included two (seven schemes) or three (six schemes) clusters, but one scheme used five, and another used seven. The average silhouette index was negatively correlated to the number of variables, as is typical of real world clustering problems (n = 15, r = -0.90; Anzanello & Fogliatto 2011).

Of the four clustering schemes with the highest average silhouette indices for a given number of variables, three were based on two clusters, and one (the best two-variable scheme) was based on three clusters. All four schemes included one category characterized by high (> 50%) average percent predawn, low (< 50 s) average latency, short (< 2.5 songs) average run lengths, and high (> 4.9%) average percent song matching (Table 3). We refer to these clusters as "category B" because of their similarity to previous descriptions of category B (Table 1). We refer to the other clusters as "category A" because of their similarity to previous descriptions of category A. The best two-variable scheme includes two clusters that are like category A. We refer to them as categories A1 and A2.

There was considerable variation among clustering schemes in the number of song types assigned to each category for each male (Fig. 1). The best one-variable clustering scheme put most song types within individual (86.3%) in category B. One male's (RDY's) songs were all assigned to category B. The best two-variable scheme assigned 69.0% of songs to category B. Two males were not assigned any A1 songs, and one was not assigned any A2 songs. The best three variable scheme, based on residual run length, residual latency, and percent pre-dawn, assigned 64.3% of songs to category B. All males were assigned A and B songs. Similarly, the best four variable scheme assigned 61.9% of songs to category B and assigned all males both A and B songs.

Cluster plots for the four best clustering schemes all included a relatively dense cluster that corresponds to category B (Fig. 2). The points outside of that cluster were more diffuse. None of the scatterplots showed a clear gap between the category B cluster and the other points. Average silhouette indices were significantly higher for well-sampled song types than they were for poorly sampled song types in three of the four best clustering schemes (Table SX). Separate plots for each male can be found in the electronic supplementary material (Fig. S1).

No song types were recorded from all nine subjects. We identified 41 (71.9%) song types that were shared by at least two birds and 14 (24.6%) song types that were shared by at least five birds. Assignment of song types to categories varied among individuals (Table S1). For example, in both the three-variable and four-variable schemes, 20 (48.8%) song types were used in different categories by different birds. Focusing on song types that are both shared and assigned to the same category, dyads shared more category B songs than category A for all four schemes (Table S1). For example, in both the three-variable and four variable schemes, dyads shared an average of 3.31 (± 2.33) category B songs, but only 1.25 (± 2.33) category A songs.

We studied the clustering results to rank variables by their importance for clustering. The best clustering schemes for a given number of variables formed a nested hierarchy (Table 2). Residual run length was present in all four schemes, residual latency was in three, percent pre-dawn was in two, and percent matching was in one. Among the one-variable clustering schemes, the scheme based on residual run length produced the highest average silhouette index (0.771), followed by schemes based on percent pre-dawn (0.738), residual latency (0.730), and percent matching (0.688). In the four variable scheme, the mean difference in standardized values between clusters was largest for percent pre-dawn (1.68), followed by residual latency (1.44), residual run length (1.04), and percent matching (0.46).

Relative to category A songs, category B songs were shorter, with fewer notes and slightly lower trill rates according to the three-variable clustering scheme (Tables 4, see Tables S2 & S3 for detailed results from the four best clustering schemes). Frequency excursion, maximum frequency, minimum frequency, and frequency bandwidth were not significantly different between song categories. Models based on the best one-, two-, and four-variable clustering schemes produced qualitatively similar results, but the FEX effect was statistically significant in the four-variable scheme, and the trill rate effect was not significant in the one-variable and two-variable schemes.

Discussion

The four clustering schemes with the highest average silhouette index for a given number of variables (hereafter, the "best" schemes) all included one cluster that matched previous descriptions of category B song and one or two clusters that matched previous descriptions of category A songs (Tables 1 & 3; Staicer 1991). Compared to category A songs, category B songs were more likely to be sung before dawn. Category B songs were also delivered in shorter runs and with shorter intervals between songs (controlling for time of day). Category B songs were more likely than category A songs to be used as song type matches with neighbors, suggesting they may be especially important for male-male vocal interactions. These song delivery patterns are similar to those of several migratory species of wood-warblers (Demko et al., 2013; Price & Crawford, 2013; Spector, 1991, 1992). We found that male Adelaide's Warbler repertoires included fewer A songs than B songs (Fig. 1; Table 3), which is consistent with the results of previous studies on Adelaide's Warbler (Staicer, 1991, 1996a) and other wood-warbler species (Lemon et al., 1985; Spector, 1991; Staicer, 1989). Overall, our results strongly support Staicer's (1991, 1996a) characterization of song categories in Adelaide's Warbler.

In the four best clustering schemes, category B songs were clustered in song delivery space, whereas category A songs were more dispersed (Fig. 2). There was no clear break between categories A and B in song delivery space. The average silhouette index from the four-variable scheme (0.379) suggests only weak evidence of clustering, but the indices from the other schemes suggest moderate-to-strong clustering (range = 0.495-0.771). Importantly, the best sampled song types within male (n > 40 songs), were significantly more clustered than the other song types in three of the four best clustering schemes (the sole exception was the one-variable scheme, Table SX). This finding indicates that sampling error depressed the average clustering values in the full dataset, and validates our decision to collect emphasize intensive sample within individuals (as opposed to collecting fewer samples from more individuals). We conclude that there are song categories in this population, but they may not be entirely discrete. Perhaps song types within male exist on a continuum from "B-like" to "A-like," with many songs clustering near the B-like end of the continuum. Larger samples permitted more precise estimates of average song delivery variables, which increased the separation between the song categories. Interestingly, a recent Pine Warbler study concluded that species' song system lacked, "the clear distinction between song categories typical of most other *Setophaga* wood-warblers" (Price & Crawford, 2013, p.559). Those results are based on recordings from the later part of the breeding season, calling into question whether song delivery patterns may be more distinct earlier in the year. Nevertheless, the hypothesis that song categories may not be entirely distinct merits further research.

The best two-variable clustering scheme included two "A" categories (Table 3, Figs. 1 & 2). Relative to category A2, category A1 was characterized by lower percent pre-dawn, lower latency, and much longer runs. We are reluctant to endorse a three-category hypothesis for several reasons. First, only one of the four best clustering schemes suggested three categories, while the others all agreed that two categories were best. Second, we know of no evidence of three-category systems in other *Setophaga* species. Third, samples from three of our nine subjects were missing either category A1 songs or category A2 songs. We conclude that certain category A songs may be used in particularly long runs, but the bulk of the evidence suggests these songs should be lumped in with category A songs, rather than being placed in a separate category.

As previously described in this species, different individuals assigned specific song types to different categories (Staicer, 1991, 1996b). Any hypothesis to explain the development of function of song categories in Adelaide's warbler must account for this finding. Males shared more category B songs than category A songs, as has previously been shown for Yellow Warblers (Beebee, 2002). This pattern was not found in Pine Warblers (Price &

Crawford, 2013), but again, that study was conducted much later in the breeding season than was ours. The findings that males shared more category B songs, and that B songs were more likely to be used as song type matches, lend credence to the hypothesis that category B songs are especially important in male-male vocal interactions.

The four best clustering schemes all indicated that relative to category A songs, category B songs were shorter and contained fewer notes. Some clustering schemes indicated statistically significant effects of song category on FEX and trill rate, but the effect sizes were small. Contrary to our results, Staicer (1996a) found no difference in song duration between categories. Otherwise, Staicer's (1996a) song structure results agree with our own: no differences in trill rate, minimum frequency, maximum frequency, or frequency bandwidth. Staicer (1996a) did not compare note number or frequency excursion between song categories, but she did find differences in syllable complexity (category B songs were more complex) and frequency spectra (category B songs had lower peak frequencies) that we did not test for.

Structural differences between song categories vary among wood-warbler species. For example, category A songs in Golden-cheeked Warblers (*Setophaga chrysoparia*) are shorter with fewer notes, the opposite of the pattern identified in the current study (Bolsinger, 2000). In Pine Warblers (Price & Crawford, 2013), trill rates are significantly higher in category A songs. Similarly, Yellow Warblers sing category A songs with lower vocal deviation (corresponding to higher rates of frequency modulation) than category B songs (Beebee, 2004b). Taken together with our results, these findings suggest that no single explanation can account for between-category structural variation among wood-warblers with divided repertoires.

We used three lines of evidence to rank song delivery variables by their importance for clustering: the hierarchical structure of the four best models, the average silhouette indices of the one-variable clustering schemes, and the magnitude of between-group differences in the four-variable scheme. The ranks of residual run length, residual latency and percent pre-dawn varied among analyses. Average ranks across analyses suggest that residual run length (mean rank = 1.7) is the most important variable for clustering, followed by percent pre-dawn (2.0), residual latency (2.3), and percent matching (4), which consistently ranked as the least important variable. This ranking suggests a different approach from previous efforts to categorize song in Adelaide's Warbler, which relied primarily on time of delivery and secondarily on latency (see Introduction). We propose three non-mutually-exclusive explanations for the finding that percent matching was weakly predictive of category membership. First, percent matching may not be very different between categories. Second, song type matching may be too rare to accurately estimate percent matching with our sample size (sampling error). Third, we may have mismeasured matching rates (measurement error). Specifically, it is likely we missed some instances of this behavior (e.g., when a neighbor's song was audible to the focal bird, but did not show up on a spectrogram).

We favor the three-variable clustering scheme based on residual latency, residual run length, and percent pre-dawn over the other clustering schemes considered in this study. When variables are not perfectly predictive of category membership, the average silhouette index tends to correlate negatively with the number of variables because each additional variable generates another dimension in which distance can be measured. Comparing average silhouette indices across clustering schemes with different numbers of variables is therefore not a sound method to identify natural clusters. Based on *a priori* expectations about song categories in wood-warblers, we expected that our samples should contain examples of all categories from all males, which was not the case in the one-variable and two-variable schemes. Prior research indicates that two clusters were more likely than three, further discounting the validity of the two-variable scheme (Spector 1992). The three-variable and four-variable schemes were very similar, but we prefer the three-variable scheme because (1) it is easier to measure three variables than four, (2) the cluster corresponding to category B is more compact in the three-variable scheme, (3) and song type matching (which contributes to the four-variable scheme but not the three-variable scheme) is harder to accurately estimate than the other variables. Future studies could assign novel song types to categories by comparing the distance in feature space between novel songs and the centroids of the two song categories using the three-variable clustering scheme (Table 3).

The goal of this study was to test for and characterize song categories. In doing so, we set the stage to address functional differences between song categories in future studies. Several studies of the behavioral context in which birds sing songs from the two categories have concluded that first category songs are directed at females and second category songs are directed at other males in *Setophaga* (Bolsinger, 2000; Kroodsma et al., 1989; Spector, 1991; Staicer, 1996b; Wiley et al., 1994). Other studies, however, have failed to find support for key prediction of one or both of those hypotheses (Beebee, 2004a; MacNally & Lemon, 1985). For example, both females (Beebee 2004a) and males (Beebee 2004a, MacNally & Lemon 1985, Weary et al. 1992, Weary et al. 1994, but see Kelly & Ward 2017) tend to respond similarly to playback of category A and category B songs. Our finding that song type matching and song sharing were higher for category B songs are consistent with the hypothesis that category B songs are important for male-male vocal interactions in this species. It is not clear, however, why category B songs would be limited to the breeding season if they are only used for male-male communication. Nor is clear how between-category variation in song structure might facilitate communication with different classes of receivers. Information about when females prospect for mates, how female presence affects males' choice of song categories, and whether females attend to male-male singing interactions like song type matching (Logue & Forstmeier, 2008), would help to clarify the functional differences between song categories.

In summary, we applied an unsupervised clustering algorithm to Adelaide's Warbler song types, and identified two song categories with different song delivery parameters. The delivery styles in the two categories are similar to patterns described in several other *Setophaga* species, and are largely consistent with previous studies on the focal species. Our findings do not allow us to exclude the possibility that song categories are not entirely discrete. Whether the use of song categories, as defined in this study, varies seasonally remains an open question (Table 1). We found evidence that category B songs are used more than category A songs during male-male vocal interactions, but further studies are required to understand the functional significance of song categories in this species.

References

Anzanello, M. J., & Fogliatto, F. S. (2011). Selecting the best clustering variables for grouping mass-customized products involving workers' learning. *International Journal of Production Economics, 130*(2), 268-276.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi:10.1038/533452a

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Beebee, M. D. (2002). Song sharing by yellow warblers differs between two modes of singing: Implications for song function. *Condor, 104*(1), 146-155. doi:10.1650/0010-5422(2002)104[0146:Ssbywd]2.0.Co;2

Beebee, M. D. (2004a). The functions of multiple singing modes: Experimental tests in yellow warblers, *dendroica petechia*. *Animal Behaviour, 67*(6), 1089-1097. doi:10.1016/j.anbehav.2003.05.016

Beebee, M. D. (2004b). Variation in vocal performance in the songs of a wood-warbler: Evidence for the function of distinct singing modes. *Ethology, 110*(531-542).

Bolsinger, J. S. (2000). Use of two song categories by golden-cheeked warblers. *The Condor, 102*, 539-552.

Catchpole, C. K., & Slater, P. J. B. (2008). *Bird song: Biological themes and variations* (2 ed.). Cambridge: Cambridge University Press.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1-36.

Curson, J., Quinn, D., & Beadle, D. (1984). *New world warblers*. London: A&C Black.

Demko, A. D., Reitsma, L. R., & Staicer, C. A. (2013). Two song categories in the canada warbler (*cardellina canadensis*). *The Auk, 130*(4), 609-616. doi:10.1525/auk.2013.13059

Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world: The psychology of judgment and decision making*: Sage.

Hedley, R. W., Logue, D. M., Benedict, L., & Mennill, D. J. (2018). Assessing the similarity of song-type transitions among birds: Evidence for interspecies variation. *Animal Behaviour, 140*, 161-170. doi:10.1016/j.anbehav.2018.04.008

Hof, D., & Podos, J. (2013). Escalation of aggressive vocal signals: A sequential playback study. *Proceedings of the Royal Society of London B: Biological Sciences 280.1768*, 20131553.

Illes, A. E. (2015). Context of female bias in song repertoire size, singing effort, and singing independence in a cooperatively breeding songbird. *Behavioral Ecology and Sociobiology, 69*(1), 139-150. doi:10.1007/s00265-014-1827-3

Kroodsma, D. E., Bereson, R. C., Byers, B. E., & Minear, E. (1989). Use of song types by the chestnut-sided warbler: Evidence for both intra-sexual and inter-sexual functions. *Canadian Journal of Zoology, 67*(2), 447-456.

Lachlan, R. F. (2007). Luscinia: A bioacoustics analysis computer program. Version 1.0. *Computer program]. Retrieved from luscinia. sourceforge. net on October, 8*, 2012.

Lemon, R. E., Cotter, R., MacNally, R. C., & Monette, S. (1985). Song repertoires and song sharing by american redstarts. *The Condor, 87*(4), 457-470. doi:10.2307/1367942

Logue, D. M., & Forstmeier, W. (2008). Constrained performance in a communication network: Implications for the function of song-type matching and for the evolution of multiple ornaments. *Am Nat, 172*(1), 34-41. doi:10.1086/587849

Macdougall-Shackleton, S. A. (1997). Sexual selection and the evolution of song repertoires. *Current ornithology*, 81-124.

MacNally, R. C., & Lemon, R. E. (1985). Repeat and serial singing modes in american redstarts (s*etophaga ruticilla*): A test of functional hypotheses. *Zeitschrift fu ̈r Tierpsychologie, 69*, 191-202.

Molles, L. E., & Vehrencamp, S. L. (1999). Repertoire size, repertoire overlap, and singing modes in the banded wren (thryothorus pleurosticus). *The Auk, 116*(3), 677-689.

Podos, J., Moseley, D. L., Goodwin, S. E., McClure, J., Taft, B. N., Strauss, A. V. H., . . . Lahti, D. C. (2016). A fine-scale, broadly applicable index of vocal performance: Frequency excursion. *Animal Behaviour, 116*, 203-212. doi:10.1016/j.anbehav.2016.03.036

Price, J. J., & Crawford, C. L. (2013). Use and characteristics of two singing modes in pine warblers. *The Wilson Journal of Ornithology, 125*(3), 552-561. doi:10.1676/13-006.1

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi:https://doi.org/10.1016/0377-0427(87)90125-7

Schraft, H. A., Medina, O. J., McClure, J., Pereira, D. A., & Logue, D. M. (2017). Within-day improvement in a behavioural display: Wild birds ‘warm up’. *Animal Behaviour, 124*, 167-174. doi:10.1016/j.anbehav.2016.12.026

Spector, D. A. (1991). The singing behaviour of yellow warblers. *Behaviour, 117*(1/2), 29-52.

Spector, D. A. (1992). Wood-warbler song systems. In D. M. Power (Ed.), *Current ornithology* (pp. 199-238). Boston, MA: Springer US.

Staicer, C. A. (1989). Characteristics, use, and significance of two singing behaviors in grace's warbler (*dendroica graciae*). *The Auk, 106*(1), 49-63. doi:10.2307/4087756

Staicer, C. A. (1991). *The role of male song in the scioecology of the tropical resident adelaide’s warbler (dendroica adelaidae).* (PhD), University of Massechussetts.

Staicer, C. A. (1996a). Acoustical features of song categories of the adelaide's warbler (*dendroica adelaidae*). *The Auk, 113*(4), 771-783.

Staicer, C. A. (1996b). Honest advertisement of pairing status: Evidence from a tropical resident wood-warbler. *Anim Behav, 51*, 375-390.

Stutchbury, B. J., & Morton, E. S. (2001). *Behavioral ecology of tropical birds*: Academic Press.

Toms, J. D. (2010). Adelaide's warbler (*setophaga adelaidae*). *Neotropical Birds Online.* version 1.0. Retrieved from https://doi.org/10.2173/nb.adewar1.01

D.M Weary, R.E Lemon, S Perreault. (1992). Song repertoires do not hinder neighbor–stranger discrimination. *Behavioral Ecology and Sociobiology*, 31 , 441-447.

D.M Weary, R.E Lemon, S Perreault. (1994). Different responses to different song types in American redstarts Auk, 111, 730-734.

Wiley, H. R., Godard, R., & Thompson, A. D. (1994). Use of two singing modes by hooded warblers as adaptations for signalling. *Behaviour, 129*(3), 243-278. doi:http://dx.doi.org/10.1163/156853994X00631

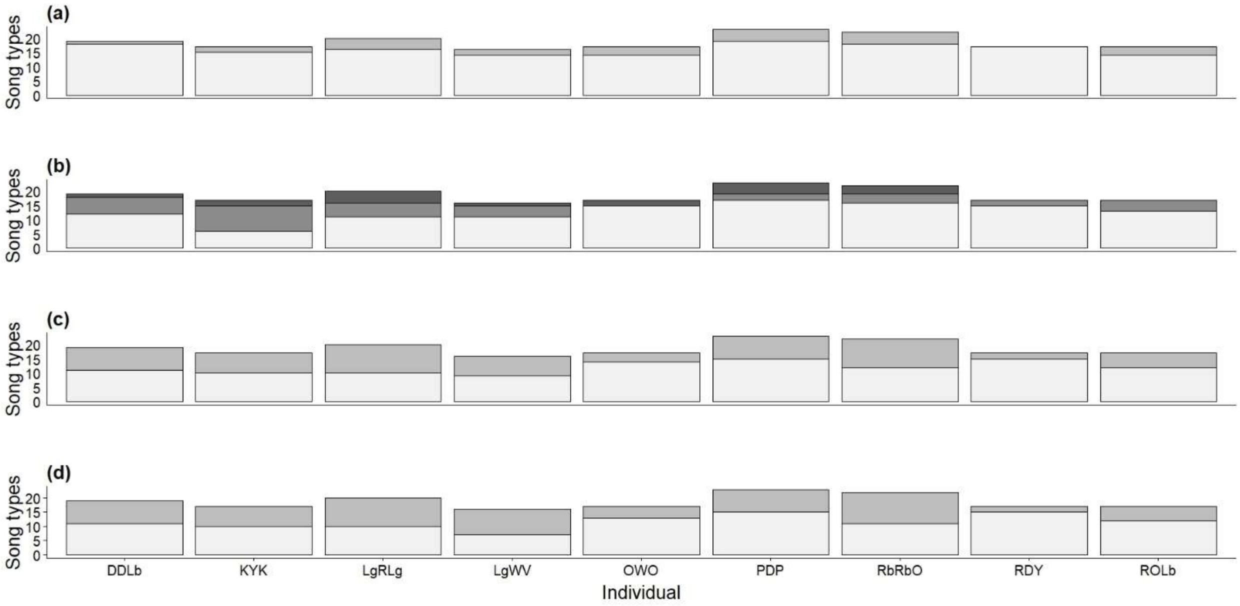

Figure 1. Assignment of song types to song categories by individual male Adelaide's Warbler according to the best (a) four-variable, (b) three-variable, (c) two-variable, and (d) one-variable clustering schemes. White bars represent category B songs, and gray bars represent category A songs. For the two-variable clustering scheme, dark gray represents category A1 and light gray represents category A2.

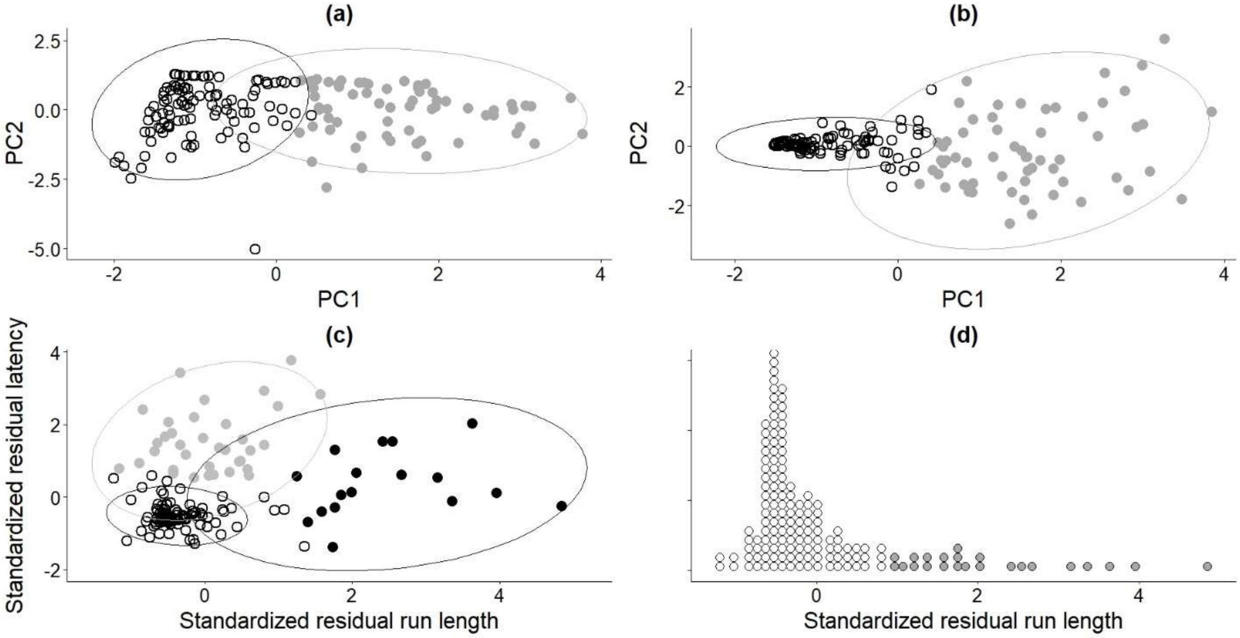

Figure 2. Cluster plots showing the best (a) four-variable, (b) three-variable, (c) two-variable, and (d) one-variable clustering schemes. Open circles represent category B songs, and filled circles represent category A songs. For the two-variable clustering scheme, black circles represent category A1 songs and gray circles represent category A2 songs.

Table 1. Hypothesized song delivery patterns of category A and category B songs in Adelaide's warbler.

Attribute	Category A	Category B
Time of year*	All year	Breeding season only
Time of day	After dawn	Before dawn and sporadically after dawn
Social context	Male-female interactions	Male-male interactions
Delivery mode	Eventual variety	Immediate variety
Latency between songs	Longer	Shorter

After Staicer (1991)

* Time of year was not examined in the present study.

Table 2. Fifteen clustering schemes for Adelaide's Warbler song types. Variables included in each model are marked with an "X."

% predawn	residual latency	residual run length	% matching	# of variables	best k	ASI of best k
X	**X**	**X**	**X**	**4**	**2**	**0.379**
X	**X**	**X**		**3**	**2**	**0.495**
	X	X	X	3	2	0.433
X		X	X	3	5	0.404
X	X		X	3	2	0.402
	X	**X**		**2**	**3**	**0.617**
X		X		2	3	0.583
X	X			2	3	0.552
X			X	2	3	0.491
		X	X	2	3	0.490
	X		X	2	3	0.450
		X		**1**	**2**	**0.771**
X				1	2	0.738
	X			1	2	0.730
			X	1	7	0.688

ASI = Average silhouette index.
The models with the highest ASI for each number of variables are in bold font.

Table 3. Means ± standard deviations of song delivery variables from the best (highest average silhouette index) categorization schemes based on one, two, three, and four song delivery variables.

# of variables	cluster	% predawn	latency (s)	residual latency	run length (songs)	residual run length	% match	category
4	1	71.87 ± 26.64	21.92 ± 16.48	1.64 ± 17.59	2.24 ± 2.71	0.09 ± 0.41	5.62 ± 4.98	B
	2	7.06 ± 13.43	112.43 ± 48.25	70.87 ± 50.59	5.85 ± 4.60	1.26 ± 1.49	3.40 ± 4.23	A
3	1	69.89 ± 28.38	23.79 ± 18.54	2.74 ± 18.16	2.15 ± 2.23	0.07 ± 0.35	5.17 ± 4.54	B
	2	6.31 ± 12.24	115.09 ± 48.86	73.52 ± 51.16	6.25 ± 4.91	1.37 ± 1.51	4.05 ± 5.24	A
2	1	63.21 ± 33.81	27.41 ± 22.69	3.16 ± 15.68	2.35 ± 2.38	0.1 ± 0.43	5.02 ± 4.48	B
	2	3.22 ± 6.47	90.72 ± 41.89	45.1 ± 42.49	11.65 ± 5.32	3.31 ± 1.13	4.16 ± 6.70	A1
	3	15.40 ± 24.69	135.80 ± 46.55	102.09 ± 40.99	3.93 ± 2.48	0.6 ± 0.7	4.25 ± 4.92	A2
1	1	4.07 ± 6.93	102.95 ± 61.51	55.34 ± 62.87	11.45 ± 4.97	2.93 ± 1.17	3.67 ± 5.86	A
	2	54.02 ± 37.30	49.02 ± 49.87	23.68 ± 43.79	2.37 ± 1.71	0.15 ± 0.44	4.95 ± 4.63	B

Table 4: Effects of song category on song structure, based on the three-variable clustering scheme. Estimated effects are relative to category A songs. See Table S2 for equivalent data from the other clustering schemes.

	duration (ms)	**notes**	**trill rate (notes / s)**	**FEX**	**F_{min} (kHz)**	**F_{max} (kHz)**	**bandwidth (kHz)**
fixed effects							
Intercept	2098.16	25.00	11.91	66.11	3178.41	6137.28	2951.89
category B	-123.67***	-2.12***	-0.31*	1.44	-46.02	-22.22	32.10
random effects							
song type	5479	2.72	0.28	16.15	16.805	11.78	17.82
bird:day	3998	1.01	0.15	6.14	5.767	11.32	7.52
Bird	691	0.52	0.06	2.25	0.283	3.81	1.29
Average ± SD							
A	2119.66 ± 291.14	25.70 ± 5.24	12.09 ± 1.56	65.87 ± 9.70	3190.03 ± 402.29	6115.27 ± 475.84	2925.25 ± 451.16
B	2119.66 ± 291.14	25.70 ± 5.24	12.09 ± 1.56	65.87 ± 9.70	3190.03 ± 402.29	6115.27 ± 475.84	2925.25 ± 451.16

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$