Inter-individual Variation in Circadian Rhythm Across Species of Halictid Bees: Characterizing Varying Clocks and Motif Classification

By

Sofía Meléndez Cartagena

A thesis submitted to the Department of Biology

FACULTY OF NATURAL SCIENCES
UNIVERSITY OF PUERTO RICO
RÍO PIEDRAS CAMPUS

In partial fulfillment of the requirements for the degree of

Master of Science

May 26 2021

Rio Piedras, Puerto Rico

Thesis Approval

FACULTY OF NATURAL SCIENCES

UNIVERSITY OF PUERTO RICO

RÍO PIEDRAS CAMPUS

In partial fulfillment of the requirements for the degree of

Master of Science

Thesis Committee:

_____ Advisor
José L. Agosto Rivera, PhD

_____ Co-Mentor
Patricia Ordoñez, PhD

_____
James Ackerman, PhD

_____
Steven E Massey, PhD

Dedication

To all the hands that made this work move forward. This is not only a labor of love, but also of community. To the curious citizens that saw me out in the field collecting and offered their help and wisdom. To my colleagues at the graduate program that kept me sane. To my friends and family, for whom my ramblings about my work must sound like gibberish, but encourage me to keep talking anyway. And of course, to my science enthusiast parents, for helping me grow into a better scientist.

Acknowledgments

I would like to thank my mentors Dr. José L. Agosto Rivera and Dr. Patricia Ordoñez for taking me in and giving me the opportunity to grow as a scientist and as a person. To Dr. James Ackerman, who has seen me grow from a baby scientist into a not-quite-grown scientist. To Dr. Steven E Massey, for asking me the difficult questions and leading me to think about science from a different perspective. I would also like to thank Elvia Melendez-Ackerman for adopting me as an honorary lab member and offering her time, wisdom, and resources.

To my peers from the graduate program, especially Jonathan A Lopez Colon, Diego A. Rosado Tristani and Elif Kardas, without your help, field work would not have been the same.

A special big thanks to my friends, Melanie E. Martines Arrocho, Nadira Yusif Rodriguez, Brian X. Leon Ricardo, and Alicia Figueroa Carlo. You did everything from reading my drafts, helping me practice my talks and even going out in the field with me. Your support is invaluable and I'm lucky to have you as friends.

Lastly, I would like to thank my undergraduate mentees, Alejandro Armas, Luis Roman Lizasoain, Beatriz A. Morales Grimany, Milexis A Santos Vega, and Andrea V. Velez Velez. I loved growing with you and wish you happiness and satisfaction in your future endeavors.

Table of Contents

List of Tables and Figures

Abbreviations List

Abbreviation	Meaning
1. *L.malachurum*	*Lasioglossum malachurum*
2. *S. curvicornis*	*Systropha curvicornis*
3. *L. ferrerii*	*Lasioglossum ferrerii*
4. *L. enatum*	*Lasioglossum enatum*
5. *L.parvus*	*Lasioglossum parvus*
6. *L. gemmatum*	*Lasioglossum gemmatum*
7. LAM	Locomotion Activity Monitor
8. RS	Rhythm Strength
9. R	Rhythmic
10. WR	Weakly Rhythmic
11. A	Arrhythmic
12. L.S	Lomb Scargle
13. M.L	Machine Learning

14. SAX	Symbolic Aggregate approXimation
15. KNN	K Nearest Neighbors
16. PAM	Partitioning Around Medoids
17. BoW	Bag of Words
18. TF-IDF	Term frequency–inverse document frequency
19. PCA	Principal Component Analysis
20. SW	Size of Word
21. PAA	Piecewise Aggregate Approximation
22. DT	Decision Trees
23. RF	Random Forest
24. TP	True Positive
25. FP	False Positive
26. FN	False Negatives

Abstract

How sociality interacts with other behaviours is a long standing question in insect biology. Simultaneously, in chronobiology, there is an unanswered question concerning how sociality influences patterns of daily activity. Past studies have established the viability of utilizing hymenopterans to describe variable circadian behaviour. Here, we intend to take a step further and establish Halictid bees as a model for cross-species comparisons of circadian rhythms. To this effect, we describe four species of Halictid bees and compare the variability of their internal clocks. We found that variability in circadian rhythms parallel complexity in sociality. We also created a machine learning pipeline to facilitate describing heterogeneous locomotor activity data. The computational experiments showed that the bee locomotion dataset transformed by Symbolic Aggregate approXimation and classified by decision trees yielded the best results. With our findings, we expect to set the basis for finding the true influence of sociality on biological clocks.

Author Biography

Sofía Meléndez Cartagena is interested in learning about the evolutionary pressures that select for sociality. In her current work, she is evaluating how sociality may affect circadian rhythms, and is developing methods to best access them. Her pursuits span anywhere from doing extensive field work to spending long hours in front of the computer building Machine Learning models. When not doing science, Sofía splits her time between enjoying the theater of the mind and doing activism work.

Introduction

Introduction

Circadian rhythms are believed to be a universal mechanism of life (Helm and Visser 2010). They determine when flowers open, when animals are hungry, and even when the temperature of a mammal's body might be at their highest or their lowest. It has been stated that these biological clocks follow a circadian rhythm, which is approximately synchronized to Earth's 24-hour rotation using signals from the environment (Roenneberg, T., et.al 2003).

Drosophila melanogaster is the main model organism used for the study of circadian rhythms. Thanks to its controlled genetic stocks created for laboratories, we have been able to describe biological clocks at the molecular level (Dubowy C & Sehgal A. 2017). Nonetheless, the genetic homogeneity and ease of manipulation that makes fruit flies attractive models leaves them lacking when it comes to answering questions about individual differences in behavior. Learning how biological clocks work in organisms that deviate from the canon may allow us to answer questions that are currently left unanswered by the *Drosophila* model. One such question is on how sociality may affect daily activity patterns.

The most studied model by far for cross-comparative studies in reference to sociality are Hymenopterans (ants, wasps and bees) and Isonopterans (Termites) (West-Eberhard 2003). Although they are not usually kept in captivity and their genetics are not as well-understood as *Drosophila*'s, the diversity of social behaviours is well-documented. Furthermore, past phylogenetic studies have found that the genes responsible for the circadian circuitry in the brain as found in Hymenopterans are more similar to those of mammals in contrast to *Drosophila*

(Rubin, et. al 2006 and Ingram, et. al 2012). Their well-known behaviors and similarity with mammals makes Hymenopterans a model of interest for transferability and comparison for mammalian behaviours.

Sociality in insects is defined as a spectrum, and, depending on the study, one might find that there are a myriad of definitions. The two biggest factors that influence those definitions are based on how an individual of a species may relate to other members of the same species, and the adaptations in behaviors concerning reproduction and brood care (Michener 1969, Toth and Rehan 2017). Here, we focus on four distinct levels of sociality: solitary, communal, primitively eusocial, and facultatively eusocial. These levels of sociality are individually defined in Table 1. To clarify, by facultatively eusocial insects we refer to primitively or strongly eusocial species whose life history suggest that they were originally solitary, or have evolved a solitary life cycle, or possess an adaptable developmental system that can express social or solitary behavior (Eickwort 1996).

Table 1: Levels of sociality as adapted from Michener (1969). 1 means present and 0 means absent.

	Caste and Division of Labor	Overlapping Generations	Cooperative Work on Cells	Structurally similar reproductive female caste (if present) may survive alone	Progressive feeding
Solitary	0	NA	0	1	0
Communal	0	0	0	1	1
Primitively eusocial	1	1	1	1	0/1
Strongly eusocial	1	1	1	0	0/1

Previous studies such as (Moore et al. 1998, Bloch et al. 2001, Giannoni-Guzman et al. 2020) have already begun establishing the use of honey bees as non-canonical subjects of study in circadian science. For example, Moore et al. 1998 and Bloch et al. 2001 established that with age, the way circadian rhythms are expressed in bees change. Furthermore, in Giannoni-Guzman et al. 2020, it was found that even within the same age group and apparent caste, there is evidence of shift work, which influences the daily activity patterns of individuals within a hive and reflects in interindividual variation. The honey bee as a subject of study proves to be a powerful tool to answer questions that are too difficult to explore under the constraints of the typical *Drosophila* model. However, if one wishes to study the influence of sociality in circadian rhythms, *Apis mellifera* is a lacking model as due to

apparent lack of diversity, genus-level *Apis* are exclusively eusocial. One would have to compare them to members of different genera to extract any sort of meaningful conclusions on how varying levels of sociality may affect the expression of circadian phenotypes, as was done in Giannoni-Guzman et al. 2014. Ultimately, the optimal way to observe the relationship between sociality and circadian rhythms would be to observe a set of species with high levels of social plasticity.

One such group of bees exist, the tribe Halictini, which in the past has been suggested as a subject of study for the interaction of socio-comparative studies (Schwarz et al. 2007, Bloch and Grozinger 2011, Toth and Rehan 2017). In the first chapter of this work, we describe and compare four different species of halictid bees whose social organization encompasses the spectrum of sociality. *Systropha curvicornis (Scopoli)* is a solitary species (Patiny and Michez 2007; Patiny et al. 2008; Danforth et al. 2008) and *Lasioglossum malachurum* (Kirby) (*L. malachurum)* is an obligately eusocial species (Richards 2000; Wyman and Richards 2003) from Greece. *Lasioglossum (Dialictus) ferrerii (Baker)* is communal. *Lasioglossum (Dialictus) enatum (Gibbs)* has not been actively studied, but it is related to primitively eusocial species (Eickwort 1988), both of these last bees are from Puerto Rico. The bees from Greece were captured the same day and same time while visiting the flowers of *Campanula arvensis*, and in addition, were kept in the same environmental conditions in the laboratory. Similarly, the bees from Puerto Rico were captured at the same day and same time while they visited *Momordica charantia*, *Sida acuta*, and *Bidens alba*.

We found that although *L. enatum* and *L. ferreiri* highly related and share the same environment, there exist distinct differences in their behavior. Furthermore, as an overall observation, bees that are not solitary at the intraspecies level express individual differences in circadian behavior. Which begs the question: what causes these differences, if not the environment nor their species? Identifying the root causes of these intraspecies behavioral differences may give us insight into how biological clocks are entrained by non-environmental cues.

The individual differences exhibited by *L. malachurum* galvanized an exploration towards a preprocessing in the traditional circadian analysis pipeline. Because these bees have such diverse expressions of circadian phenotypes, we found ourselves dividing the bees into discrete groups and describing the population by the use of subpopulations. This process was the manual equivalent of grouping individuals using machine learning. Once we were confident that the groups observed within the population of *L. malachurum* were discrete and informative for circadian purposes, we took the next step and designed a machine learning pipeline.

In chapter 2, we explore the use of said Machine Learning (M.L) pipeline. Machine learning is a subset of Artificial Intelligence. It is defined as the process of solving real-life problems by gathering data sets and using algorithms to create models based on those data sets (Burkov, 2019). Although machine learning has many applications, we employed it for two distinct purposes: clustering and classification. Clustering is a type of unsupervised machine learning where algorithms are not given predetermined categories (labels) to group data into. The data, based on measurements of similarity, instead groups together naturally (Géron,

2018). Classification is a type of supervised learning where data already has the desired solutions (labels). A classification algorithm learns patterns from a labeled dataset and uses this information to classify new data into these same categories (Géron, 2018).

The use of M.L in circadian science is usually relegated to the evaluation of genes and proteins (Agostinelli et. al 2016 and Anafi et. al 2014). It is our understanding that the use of M.L as a preprocessing step to locomotor analysis is a novel application of this tool, which presents the challenge of having to design a pipeline and workflow from scratch.

Using common transformations for circadian analyses (Lomb Scargle Periodogram, Average of Daily Locomotor Activity and Autocorrelation) (Refinetti et. al 2007) we explored the natural grouping of the various transformations of the *L. malachurum* dataset. The approach was univariate, where we applied PAM or K-means to just one of the transformations. This culminated in three different clustering results, one per transformation. None of the three transformations resulted in groups that were of circadian significance, and prompted us to explore the use of classification. We divided the data for *L. malachurum* into three broad categories, and then transformed the time series data using Symbolic Aggregate approXimation (SAX) for ease of analysis and dimensionality reduction. The data was transformed with different combinations of SAX parameters and were classified using K Nearest Neighbors, Decision Trees and Random Forest classification algorithms. Lastly, we compared the performance of all three algorithms. Decision Trees achieved the best results, followed closely by K Nearest Neighbors.

We hypothesize that sociality is one of the keys to understanding the complexities of circadian rhythms. In this work, we meet our set goal by; First, describing four species of halictid bees with varying degrees of sociality. Second, by developing a preprocessing step using machine learning that is easily incorporated into the traditional circadian pipeline. This preprocessing step facilitates the grouping of individuals of *L. malachurum* into discrete groups once trained. While it stands to be seen if the process is transferable to other species, its effectiveness can nevertheless prove useful in future studies where *L. malachurum* is involved. All of the caveats notwithstanding, our findings can serve as the basis for a larger body of work that elucidates the true relationship between sociality and circadian rhythms.

Chapter 1

Relationship Between Inter-individual Variation in Circadian Rhythm and Sociality: A case Study Using Halictid Bees

Work done in collaboration with: Carlos A. Ortiz-Alvarado, Claudia S. Cordero-Martínez, Alexandria F. Ambrose, Luis A Roman Lizasoain, Milexis A Santos Vega, Andrea V Velez Velez, Jason Gibbs, Theodora Petanidou, Thomas Tscheulin, John T. Barthell, Victor H. González, Tugrul Giray, José L. Agosto-Rivera

Relationship Between Inter-individual Variation in Circadian Rhythm and Sociality: A case Study Using Halictid Bees

Abstract:

The bee family Halictidae is considered to be an optimal model for the study of social evolution due to its remarkable range of social behaviors. Past studies in circadian rhythms suggest that social species may express more diversity in circadian behaviors than solitary species. However, these older studies did not make appropriate taxonomic comparisons. To further explore the link between circadian rhythms and sociality, we examine four halictid species with different degrees of sociality, three social species of Lasioglossum, one from Greece and two from Puerto Rico, and a solitary species of Systropha from Greece. Based on our previous observations, we hypothesized that species with greater degree of sociality will show greater inter-individual variation in circadian rhythms than solitary species. We observed distinct differences in their circadian behavior that parallel differences across sociality, where the most social species expressed the highest inter-individual variation. We predict that circadian rhythm differences will be informative of sociality across organisms.

Introduction:

Understanding the evolutionary link between solitary and eusocial lineages, and their adaptive behaviors, such as those expressed in reproduction and brood care, is a perennial question in insect evolutionary biology (Toth and Rehan 2017). Insects, and in particular hymenopterans, have been useful in observing how

sociality is related to other types of behaviors, such as competitor effects (Peters et al. 2017). A potential behavior to be evaluated is circadian rhythms, as it has been proposed to be governed by demands arising from sociality not only in insects but also in mammals (Mistlberger 2004; Giannoni-Guzman et al. 2014; Beer and Helfrich-Förster 2020).

Circadian rhythms can be viewed as a biological clock that most living organisms possess. These biological clocks regulate processes such as gene expression, behavior, body temperature, and sleep-wake patterns. Biological clocks follow a rhythm that is approximately synchronized to Earth's 24-hour rotation using signals from the environment called zeitgebers. This process of synchronization needs active reestablishment, and it is called entrainment (Roenneberg et al. 2003).

Circadian rhythms have been studied in a wide range of organisms, from plants, invertebrates, birds, and mammals (Helm and Visser 2010). The traditional model animal to study this phenomenon is the fruit fly, *Drosophila melanogaster*, where biological clocks are described at the molecular level (Dubowy and Sehgal 2017). Although this model has been pivotal to the understanding of circadian rhythms, the lack of genetic diversity in the fruit fly reduces the relevance of the model because it limits questions regarding individual differences in rhythms.

Past studies, such as those of Bloch et al. (2001) and Moore et al. (1998), revealed that the rhythmicity of honey bees changed with age. Additionally, Giannoni-Guzman et al. (2020) showed that foragers in the wild display discrete categories that suggest temporal shift work. An earlier study from Giannoni-Guzman

and colleagues (2014) compared the endogenous period of three different variants of honey bees (*Apis mellifera carnica*, *Apis mellifera caucasica* and *Apis mellifera* gAHB) as well as similarly sized insects from different orders and families. They found that honey bees and paper wasps (*Polistes crinitus* and *Mischocyttarus phthisicus*) had a larger degree of circadian period variation within the population in comparison to *D. melanogaster*. They mentioned several possible explanations for their observations, one of them being sociality. In a more recent study, Beer and Helfrich-Förster (2020) explore this connection further and note that the development of the circadian circuitry varies between an eusocial (*Apis mellifera*) and a solitary species (*Osmia bicornis*). In particular, they observe that eusocial individuals are born with an undeveloped circadian clock while the solitary individuals emerge with it fully developed, and attribute these differences to their opposite levels of sociality. However, because these two past studies were done with species spanning from different taxonomic groups, it would be difficult to support their claims without taking phylogeny into account. Nevertheless, these works do give a basis to ask how circadian rhythms vary and are an integral part of the survival strategy and organization of these animals. Moreover, it leads us to consider that the level of sociality in different organisms may play a role in their daily activity patterns. Specifically, one could suppose that complexity in levels of sociality of an insect may be reflected in individual differences in circadian rhythms of individuals of the same population.

Halictidae (*Hymenoptera*) is a bee family considered to be a great model for the study of social evolution due to its exceptional diversity in respect to social

behavior within and among species and populations (Schwarz et al. 2007). *Lasioglossum Curtis* is one of the two genera in the tribe *Halictini* that displays eusocial behavior, but also includes solitary representatives and a range of intermediate social categories (Danforth et al. 2003; Gibbs et al. 2012). Additionally, past studies have shown plasticity in the social behavior even among populations of the same species (Eickwort et al. 1996; Field 1996; Field et al. 2010; Richards et al. 2003; Soucy and Danforth 2002; Richards 2000). Depending on environmental conditions such as elevation, latitude and seasonality, halictid bees might display different modes of sociality. Species with social nests may revert to solitary behavior at high latitudes and altitudes (Eickwort et al. 1996; Packer et al. 1983; Field et al. 2010) or based on access to mates (Yanega 1988/1989). Jeanson et al. (2008) studied members of a solitary species, *Lasioglossum (Ctenonomia)* sp. NDA-1 and observed the results of having them nest in pairs. They observed that after some time together, the individuals in the nest started to show signs of division of labor. This plasticity and diversity of behavior, in addition to the close taxonomic relation, makes *Halictidae* an optimal model for observing the relationship between sociality and circadian rhythm (Bloch and Grozinger 2011).

To better understand how social behaviors can be associated with circadian rhythms in insects, we have set out to document the rhythm in four species of halictid bees that span a gradient of social complexity. *Systropha curvicornis* (Scopoli) (*Halictidae*: *Rophitinae*), a solitary pollinator specialist (Grozdanić and Mučalica 1966) considered ancestrally solitary within the family Halictidae (Patiny and Michez 2007; Patiny et al. 2008; Danforth et al. 2008), and three species of *Lasioglossum* (*L.*

ferreri, L. enatum and *L. malachurum*), which were selected because of their varying levels of social behavior (Eickwort 1988; Wyman and Richards 2003; Gibbs 2018). Species of *Lasioglossum* likely had a common ancestor capable of eusocial nesting but have reverted multiple times to other levels of sociality (Danforth et al. 2003; Brady et al. 2006; Gibbs et al. 2012). *Lasioglossum* (Dialictus) *ferrerii* (Baker) and *L.* (Dialictus) *enatum* (Gibbs) is found in the Caribbean, whereas *L. malachurum* is found across Europe; the first species nests communally, that is, each individual contributes to nest construction and reproduction (Michener 1974; Eickwort 1988). Although *L. enatum* has not been thoroughly studied, this species is part of a species complex that includes weakly eusocial species. Namely, *L. gemmatum* (Smith) and *L. parvum* (Cresson) from Jamaica and the Bahamas, which exhibit reproductive division of labour (Eickwort 1988). There is no evidence of morphologically defined castes beyond reproductive status in these two species, and thus we assumed this is likely the case for *L. enatum* in Puerto Rico, where we conducted our experiments. In contrast, *L. malachurum* (Kirby) is an obligately eusocial species with a morphologically well-defined queen and worker castes (Richards 2000; Wyman and Richards 2003). *Lasioglossum malachurum* is known to display varying degrees of behaviors depending on location (Richards 2000). In Lesbos, Greece, where we studied this species, they were observed to exhibit a facultatively eusocial behavior (Wyman and Richards 2003).

The social plasticity of *Lasioglossum* and its potential as a model for social evolution leads us to believe that observing this group of bees can give invaluable insight on how social behavior affects biological clocks. To test our idea, we captured

these wild bees as they were visiting flowers and observed them in the laboratory using constant conditions. Based on these observations, we determined the variety of behaviors present, and made inferences on how they associate with their sociality.

Methods and Materials:

Study sites

Puerto Rico:

Lasioglossum ferrerii and *L. enatum* were captured using 15 mL falcon tubes from flowers at the Balneario de Luquillo parking lot (18.38706 N 65.72517W, 3 Meters) in Puerto Rico. This site is characterized by having many vine-type plants, high vegetation density, and it is further located right next to a road where *Momordica charantia* is quite abundant. (Figure 1.A and 1.B). Most bees were caught between 8:00 and 12:00h at the flowers of *Momordica charantia L.* (*Cucurbitaceae*), *Sida acuta Burm.fil.* (*Malvaceae*), and *Bidens alba* (L.) DC.(*Asteraceae*). We also observed them visiting *Euphorbia heterophylla*, which has not been reported in previous literature. Collections took place during the months of December, January, March, and August. In total, we collected 36 bees, 26 of which were *L. ferrerii* and 10 were *L. enatum*.

Greece:

Systropha curvicornis and *L. malachurum* bees were collected between 6:00 and 9:00 h from flowers of *Convolvulus arvensis* (*Convolvulaceae*) that were growing on a recently cut wheat field in Skala Kallonis (39º 10’N 26º 20’E, 0 Meters) on the

Island of Lesbos, Greece. We used 15 mL falcon tubes to catch bees in the field, which would house the individual for the duration of the experiment. Sampling was conducted on July 3 of 2017. From this sampling 118 bees were *L. malachurum* and 34 were *S. curvicornis*.

Laboratory settings

After collection in the falcon tubes, bees were provided with food that lasted for the whole of the observation period. The food recipe we used varied between the studies in Puerto Rico and Greece (As explained below). The main nutrient for both recipes was sugar, and are therefore nutritionally comparable. However, the agarose based recipe (Puerto Rico) was more convenient in terms of ease and speed of preparation due to the fact that an independent water system was not necessary.

Food preparation varied by locations as follows: In Puerto Rico, for every 0.89 ml of water, 1 g of sucrose and 0.1 g of agarose were used. The water was heated in a stirring plate with a magnetic stirrer placed at the bottom. We added the sucrose first to the solution. The agarose was then incorporated upon its dissolution. We left the solution stirring until it turned into a lighter color while being mindful of not letting the solution overheat, as to avoid part of its volume being lost to evaporation. As a form of assurance, we made 3 ml more than what was expected to be used. After all solids were diluted, we quickly pipetted 1 ml into the bottom of a 15 ml centrifuge tube, being mindful of not letting it splash, as to preserve as much solution as possible. Once all of the tubes had their portion of the solution, they were allowed to reach room temperature, and were finally refrigerated. The final product was a gel

that could be kept refrigerated until the day it needed to be used as long as it did not dehydrate.

In Greece, captured *S. curvicornis* and *L. malachurum* were fed with ApiYem (Namik Kemal University with Kosgeb R&D Innovation Project) which is a food substitute composed of 78.5% sugar and 21.5% invert syrup. Food was placed in the cap-end of each tube and a damp cotton was placed in the other end of the tube to provide water to the bees. The water supply was refilled every 2–3 days. Resources were provided *ad libitum* during the complete running period of the experiment.

Locomotor activity monitoring

Each bee was monitored individually for at least seven days in the falcon tube in which they were captured. The tubes were plugged using cotton balls to let air circulate. These tubes were then placed into TriKinetics' Locomotion Activity monitors (LAM16) that, in turn, were put inside incubators that were set to constant conditions. In Greece, the temperature was 26°C, humidity was at 78%, and light conditions were constant darkness. In Puerto Rico, the temperature was 30°C, humidity was at 65%, and light conditions were constant darkness. The differences in the environmental chamber conditions were set to resemble the average daytime parameters at each location.

Species Identification

The individuals caught in Puerto Rico were identified using Gibbs (2018). Samples collected in Greece were identified by an in-field expert, Victor H. Gonzalez (University of Kansas).

Data processing and analysis

Circadian Analysis

Circadian rhythm and locomotor activity for our subjects were analyzed using the MATLAB toolboxes developed in Jeffrey Hall's laboratory (Levine et al. 2002). The outputs provided data on the individual's locomotor activity throughout the experiment in the form of an actogram, average activity plot, and an autocorrelation that also calculates rhythm strength.

To test if the observed differences in circadian patterns across species, we applied a Brown-Forsythe one way ANOVA with a Dunnett's T3 multiple comparisons test using GraphPad Prism version 8.4.3 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com. The variables taken into consideration for this were time, species, individuals, and interspecies variation.

Results:

During the Summer of 2017 on the 3rd of July between the hours of 6:00 am and 9:00 am, *S. curvicornis* and *L. malachurum* were caught as they visited the flowers of *Convolvulus arvensis* on the island of Lesbos. They were transported from the field to the laboratory and placed inside an incubator for 10 days of which 8 were in constant conditions (26 °C and 57% humidity) with the purpose of characterizing their intrinsic biological clock. During this collection, 118 bees were *L. malachurum*, although only 98 survived until the end, and 34 were *S. curvicornis* of which only 4 females survived the study period.

Figure 2.A illustrates the average of individuals evaluated under constant conditions of the Greek, solitary and specialist pollinator *S. curvicornis*. Its period runs slightly short at approximately 22 hours, with the peak of its activity in the early morning and an average rhythm strength of 4.4. The population average is consistent with the individuals examined, as illustrated in Figure 2.B; a randomly selected individual looks fairly similar to the activity plotted for the population average. The average for period was 22.75 with a standard deviation of 0.41, Rhythm Strength had an average of 4 and standard deviation of 0.707.

The consistency displayed by our population of female *S. curvicornis* is contrasted with the diversity observed in the other 3 species analyzed in this study. This is especially true for the eusocial *L. malachurum*, for whom after careful evaluation of the data we had to create a classification schematic (Figure 3) to appropriately describe the phenotypes being displayed by the population. The population average shows that this species has a perfectly circadian 24-hour period under constant dark conditions. Peak average activity of *L. malachurum* is at 6:00 h, with no clear rest periods when all individuals are averaged. When examined individually, we found five distinct patterns of circadian activity patterns (Figure 3.A and Figure 4). These patterns can be divided into 2 large branches (Figure 3.A), those that are rhythmic and those that are arhythmic, i.e, individuals with uniform distribution of activity. Rhythmic individuals varied in the amplitude of their activity rhythm, and were therefore classified as strong or weakly rhythmic. Moreover, both strong and weak categories are subdivided into unimodal or bimodal based on the number of activity peaks per day. For example, a bimodal individual is active during two different

instances of the day, such as in the morning and in the afternoon (Figures 4.B and 4.C), while a unimodal individual is mostly active during a set time of the day (Figures 4.D and 4.E).

Strongly rhythmic individuals (either unimodal or bimodal) constituted 41% of individuals. These patterns are recognized by a strong Rhythm Strength (RS)(Figures 4.B.iii and 4.D.iii) on average higher than 2.67 and clear rest and active periods both in the double plotted actogram (Figures 4.B.i and 4.D.i) and the average activity plot(Figures 4.B.ii and 4.D.ii). Weak rhythmicity (Figure 4.C and 4.E) was observed in 21.6% of individuals and they were characterized by having RS values (Figure 4.C.iii and 4.E.iii) that average on 1.79, but their actograms (Figure 4.C.i and 4.E.i) do not show a clear pattern of locomotor activity. Finally, 38% of individuals were arrhythmic. Both the double plotted actogram (Figure 4.F.i) and the average activity plot (Figure 4.F.ii)for arrhythmic bees do not have any discernible daily pattern of activity or inactivity. Often the autocorrelation (Figure 4.F.iii) does not return any values.

On February 19, 2020, between 8:00 am and 10:00 am, at the Balneario de Luquillo (Figure 1.A and B), 36 bees were captured as they visited *Bidens alba*, *Momordica charantia* and *Sida acuta*. Bees were captured and monitored individually in one tube each (modified from Giannoni-Guzman et al. 2014). In the laboratory, the bees were placed inside an incubator for seven days in constant conditions so we could characterize their intrinsic biological clock.

Of the 36 bees captured, 26 were identified as *Lasioglossum ferrerii* (Figure 1.C and 1.D) and ten as *Lasioglossum enatum* (Figure 1.E and 1.F). Only 22 L. *ferrerii* and 8 *L. enatum* survived the entire observation period and were used for analysis. The average peak of circadian activity for *L. ferrerii* is between 6:00–7:00 (Figure 5.A.ii) with a 23-hour period (Figure 5.A. iii), making it rather short. Individuals fell into two categories: 50% were strongly rhythmic and 50% were weakly rhythmic (Figures. 5B and 5C). The peak of average activity for *L. enatum* is from the fifth to the seventh hours of the day, with a circadian period of 23.8 hours (Figure 6.Aiii), just slightly short of one day. The average peak of activity for *L. enatum* was 6:00-7:00 (Figure 6.A.ii). *L. enatum* also had three patterns of activity with similar characteristics to that of the umbrella categories for *L. malachurum*, and we saw it fit to categorize them in a similar fashion (Figure 6). 12.5% of the observed population fell into the strongly rhythmic category, 25% in the weak rhythmic category, and 62.5% in the arrhythmic category.

In summary, all four of the described species followed unique patterns of behavior (Figure 7.A) characterized by the amount of interindividual variation. *Systropha curvicornis* was the species with the least amount of observed interindividual variation in its daily activity patterns, followed by *L. ferrerii* with two distinct patterns of behavior, then *L. enatum* with 3, and lastly, *L. malachurum* with 5.

Cross species comparison of observed circadian parameters.

Average activity was only significant between *L. ferrerii* and *L. malachurum* with a p-value of 0.0016 (Figure 7.B). Circadian Period (Figure 7.C) on the other

hand showed differences between *S. curvicornis* and *L. ferrerii* with a p-value of 0.0102 as well as *L. malachurum* and *L. ferrerii* with a p-value of <0.0001. Lastly, with even more differences still, Rhythm Strength (Figure 7.D) presented differences between *S. curvicornis* and *L. enatum* (p-value = 0.0014), S. curvicornis and *L. malachurum* (p-value = 0.0177), *L. ferrerii* and *L. enatum* (p-value = 0.0056) and finally, *L. ferrerii* and *L. malachurum* (p-value = 0.0259).

Discussion:

The solitary specialist, *S. curvicornis* as observed in this work, suggests that at least for the females, the population is consistent, displaying a single circadian activity phenotype (Figure 7.A). The activity of these bees is highly rhythmic, and shows little variation across samples with an average RS of 4.4 and the peak of activity appears to be near the hour 6 of the day. Overall, the species exhibits a short period phenotype under constant darkness. This high degree of rhythmicity might be due to *S. curvicornis'* evolutionary history as a foraging specialist of *C. arvensis*, which blooms for a brief period during the morning, a pattern described for another closely related species (*S. planidens*: Gonzalez et al. 2014). A rigorous internal clock is important to be able to anticipate the time when resources are available. For example, the immediate development of *Osmia bicornis*' circadian rhythm (Beer and Helfrich-Förster 2020) may be related to nurishment accessibility, as it has been shown in the past that large quantities of pollen are the key to proper larvae development rather than diversity of pollen (Radmacher and Strohm 2010). All three of these species (*S. curvicornis*, *S. planidens* and *O. bicornis*) lead solitary lifestyles

and consequently must assure the survival of their progeny in an individual manner. A strong circadian rhythm can ensure that a female may find sufficient resources efficiently to feed its young.

All three species of *Lasioglossum* examined were shown to have more than one distinct pattern of circadian activity. The most diverse of the three was the facultatively eusocial *L. malachurum* (Figure 7.A) with 5 distinct circadian behaviors. Two of the sub-categories of these behaviors fall under the strongly rhythmic category, which we are calling binomial and unimodal. These rhythmic categories are characterized for having an easy-to-distinguish pattern in the actogram, clear rest/activity periods in the average activity plot, and an RS higher than one. Another two of the subcategories fall under the weakly rhythmic umbrella. This umbrella, just like the strongly rhythmic category, can be divided into bimodal and unimodal. These categories can be identified by an actogram with no clear pattern, an average activity plot with more or less clear rest/activity pattern, and an RS larger than one. Lastly, there is the arrhythmic category where no discernible pattern can be pinpointed in the actogram nor in the average activity plot and its RS is less than one. A conceptual map on how these categories are identified can be found in Figure 3.A.

To make descriptions comparable across species, we used the same metrics to categorize the other two bees examined in this study. In categorizing *L. ferrerii* and *L. enatum*, our categories worked as a good basis. *L. ferrerii* only had two distinguishable patterns: rhythmic and noisy rhythmic (Figure 7.A). We decided to change the name from weakly rhythmic to noisy rhythmic because it is a better

descriptor (Figure 3.B). Similarly, *L. enatum*, which lives in the same environment as *L. ferrerii*, has 3 distinguishable categories (Figure 7.A). These categories are rhythmic, noisy, rhythmic and arrhythmic. In contrast to *L. ferrerii*, *L. enatum* expressed 5 individuals in the arrhythmic category. Taking into consideration that both of these species of bees were caught in the same environment, and that they belong to the same genus, the results suggest that something other than environmental variables are behind these differences.

The difference in expression of circadian patterns between *L. ferrerii* and *L. enatum* could be explained by competition. Both of these species share the same niche in Luquillo, to the point of being caught in the same flowers during the same range of time. Having a slight difference in rhythmicity can lower the possibility of temporal competition when foraging. *L. ferrerii* on average would be active from 5:00 am to 10:00 am, while average time of activity for *L. enatum* would be from 3:00 am to 8:00 am. Due to that two-hour disphase, it would appear to be less likely that bees from these two species try to visit a flower simultaneously, yet their schedules still have some overlap. These observations are echoed by another study conducted in Greece where they demonstrated that 3 species of carpenter bees (*Xylocopa spp*.) that share the same resources have different circadian rhythms when measured under natural field conditions and also in artificial constant and oscillating conditions (Ortiz-Alvarado et al. in rev.). While the solitary, *Xylocopa* species have interspecies variation in their circadian rhythms, two out of the three examined in Ortiz-Alvarado et al. follow a similar pattern as *S. curvicornis*, where there isn't much, if any individual differences in the populations examined. Therefore, in that particular case,

competitor effects can explain the differences in rhythm across species, but in the case of *L. ferrerii* and *L. enatum*, it cannot explain the individual differences observed at the species level.

At a higher level looking at the statistical analysis of all four halictid bees (Figure 7), some interesting patterns can be noted. In terms of average activity (Figure 7.B), there was only a difference between *L. ferrerii* and *L. malachurum*, none of the other possible combinations of differences occurred. However, the length of the whiskers in the box plots for both *L. enatum* and *L. malachurum* does suggest a level of diversity at the intraspecies level and could be reflective of the number of circadian behaviors observed in these species.

When analyzing the circadian period, the observed differences were between *L. ferrerii* and *L. malachurum* as well as *S. curvicornis* and *L. malachurum* (Figure 7.C). The latter of these pairs have shared environmental conditions when the former pair does not. It is also interesting to note that *L. ferrerii* and *S. curvicornis* cannot be found in the same locations, and yet, they do not appear to have significantly different circadian periods. In fact, for the populations examined, they appear to be comparable.

The picture becomes clearer still when observing the differences in rhythm strength (Figure 7.D). Where those species with a lesser number of circadian phenotypes are more similar to each other, and likewise, those with the most diversity are more similar to each other. In other words, *S. curvicornis* and *L. ferrerii* were both significantly different to *L. enatum* and *L. malachurum*, but not to each

other. Likewise, there was no significant difference between *L. enatum* and *L. malachurum*. Because there is this consistency of differences that is not associated with differences in environments, we believe that the key to explaining the difference in diversity of behaviors may not lay in competition, but in something more endogenous of the species. Nevertheless, more data is needed.

As we mentioned before, *Lasioglossum* as a genus is well-known for having a large diversity in social behaviors. This diversity in sociality may also be reflected in other types of behaviors, and could be the key to explaining the individual differences in circadian rhythm we observed at the interspecies level. A limitation of our work was the sample size and the number of the evaluated populations (one population for each species and low number of individuals, particularly for *L. enatum*), thus, it definitely merits repetition of the work to see if our results can be replicated. Furthermore, in the one species whose population was close to a hundred, more time than usual was needed to evaluate the data, due to the diversity found within it. In future studies we will focus on streamlining the process of describing a species with diverse circadian behaviors such that it will facilitate studies with a higher volume of observations. Additionally, we will continue describing the circadian rhythm of more species of *Lasioglossum* who present social behaviors not evaluated in this study, and will evaluate if there is a relationship between sociality and rhythm, causal or otherwise.

Figure Legends:

Figure 1: Habitat (A-B) and Species Observed (C-F). A) Puerto Rico study site in which *L. ferrerii* and *L. enatum* were captured. B) Some of the vegetation the bees were observed visiting, with flowers belonging to the families: Commelinaceae, Cucurbitaceae and Euphorbiaceae. C) Female of *L. ferrerii* distinguished from the male by its short antenna and pointed abdomen. D) Male of *L. ferrerii*, distinguished by its long antennae and flat abdomen. This species is known for its long head shape and metallic metasoma (Gibbs 2018). E) Female of *L. enatum* distinguished from the male by its short antenna and pointed abdomen. F) Male of *L. enatum*, distinguished by its long antennae and flat abdomen. This species is distinguished by: “tegula punctate, extended posteriorly to form a small angle, mesepisternum punctate and metasoma brown” (Gibbs 2018).

Figure 2: Female *S. curvicornis* exhibit short period phenotype under constant darkness (<24 h endogenous circadian rhythm). (i) Double-plotted actogram showing the locomotor activity pattern of: (A) The average of all 4 individuals examined of S. curvicornis. (B) A representative individual randomly selected from the population. In a double plotted actogram, each row represents locomotor activity (counts per 30 min) of two consecutive days and the second is repeated such that it is always the beginning of the next row. The x-axis shows the time of day under constant darkness expressed as circadian time (CT). (ii) Average of the locomotor activity patterns of the five days observed into one. The y-axis represents activity over time and the x-axis represents CT (iii) Autocorrelation plots used to determine

the period (p), rhythm index (RI) and rhythm strength (RS). The oscillations indicate periodicity. The asterisk at the third peak of the autocorrelation plot indicates the particular time point used for the determination of the rhythm parameters.

Figure 3: A summary of the variations in the circadian rhythm as observed in: A) *Lasioglossum malachurum*, B) *Lasioglossum enatum* and *Lasioglossum ferreirii*. The circle represents the root of the flowchart, squares represent nodes that branch off and rhombuses represent leaves. In total for malachurum, 5 distinct behaviors were observed.

Figure 4: *L. malachurum* exhibits a variety of circadian phenotypes under constant dark conditions. (i) Double-plotted actogram showing the locomotor activity pattern for 5 days of: (A) The average of all 98 individuals examined and representatives for the following categories: (B) Bimodal Rhythmic (C)Weakly Rhythmic Bimodal, (D) Unimodal Rhythmic, (E) Weakly Rhythmic Unimodal, (F) and Arrhythmic circadian behaviors. (ii) An average activity plot for the five days of observation (iii) Autocorrelation plots used to determine the period (p), rhythm index (RI) and rhythm strength (RS).

Figure 5: Description of the circadian behaviors exhibited by *L. ferrerii* under constant dark conditions. (i) Double-plotted actogram of the locomotor activity from the five-day observational period for: A) All 22 individuals from the data set averaged out into one representative individual. B) A representative individual out of the 11 from the category Strongly Rhythmic. C) A representative individual out of the 11 from the category Noisy Rhythmic. (ii)An average activity plot for the five days of

observation (iii) Autocorrelation plots used to determine the period (p), rhythm index (RI) and rhythm strength (RS).

Figure 6: Description of the circadian behaviors exhibited by *L. enatum* under constant dark conditions. (i) Double-plotted actogram of the locomotor activity from the five-day observational period for: A) All 8 individuals from the data set averaged out into one representative individual. B) The only individual from the category Strongly Rhythmic. C) A representative individual out of the 2 from the category Noisy Rhythmic. D) A representative individual out of the 5 from the Arrhythmic category. (ii)An average activity plot for the five days of observation (iii) Autocorrelation plots used to determine the period (p), rhythm index (RI) and rhythm strength (RS).

Figure 7: Summary of descriptive and inferential statistics. A) Number of circadian categories observed by species. B) Box plot illustrating the difference in average locomotor activity between species. *S. curvicornis* has a minimum of 8.200, 25% percentile of 8.900, mean of 13.63, 75% percentile of 20.28 and a maximum of 22.90. *L. ferrerii* has a minimum of 0.7000, 25% percentile of 2.675, mean of 4.950, 75% percentile of 7.100 and a maximum of 11.20. *L.enatum* has a minimum of 1.000, 25% percentile of 2.875, mean of 8.113, 75% percentile of 6.650 and a maximum of 36.00. *L. malachurum* has a minimum of 6.700, 25% percentile of 10.35, mean of 8.201, 75% percentile of 10.35 and a maximum of 26.30. There was only a statistical difference between *L. ferrerii* and *L. malachurum* with a p-value of 0.0016, DF of 61.66 and t of 3.867. C) Box plot illustrating the difference in circadian

period between species. *S. curvicornis* has a minimum of 22.20, 25% percentile of 22.35, mean of 22.75, 75% percentile of 23.10 and a maximum of 23.20. *L. ferrerii* has a minimum of 21.80, 25% percentile of 21.95, mean of 22.69, 75% percentile of 23.20 and a maximum of 24.200. *L. enatum* has a minimum of 20.00, 25% percentile of 22.40, mean of 23.31, 75% percentile of 24.60 and a maximum of 25.50. L, malachurum has a minimum of 20.20, 25% percentile of 23.50, mean of 24.00, 75% percentile of 24.50 and a maximum of 27.80. Both *S. curvicornis* and *L. ferrerii* were significantly different from *L. malachurum* with p-values of; 0.0102 and <0.0001, DFs of; 5.652 and 52.94 and, t of; 5.179 and 6.565, respectively. D) Box plot illustrating rhythm strength among species. *S. curvicornis* has a minimum of 3.000, 25% percentile of 3.250, mean of 4.000, 75% percentile of 4.500 and a maximum of 4.500. *L. ferrerii* has a minimum of 0.6000, 25% percentile of 1.700, mean of 2.691, 75% percentile of 4.125 and a maximum of 4.600. *L. enatum* has a minimum of -2.500, 25% percentile -0.4250, mean of 0.2125, 75% percentile of 1.200 and a maximum of 2.200. *L. malachurum* has a minimum of -2.100, 25% percentile of 0.7250, mean of 1.754, 75% percentile of 2.800 and a maximum of 4.500. The solitary, S. curivcornis, and communal *L. ferrerii*, were significantly different from the eusocial species, but not each other. Likewise, *L. enatum* and *L. malachurum* were not significantly different. S. curvicornis vs. *L. enatum* ; p-value of 0.0014, df of 7.932 and t of 6.226. *S. curvicornis* vs. *L. malachurum*; p-value of 0.0177, df 4.04 of and t of 5.895. *L. ferrerii* vs. *L. enatum*; p-value of 0.0056, df of 17.09 and t of 3.982. *L. ferrerii* vs. *L. malachurum*; p-value of 0.0259, df of 33.42 and t of 3.054.

Tables and Figures

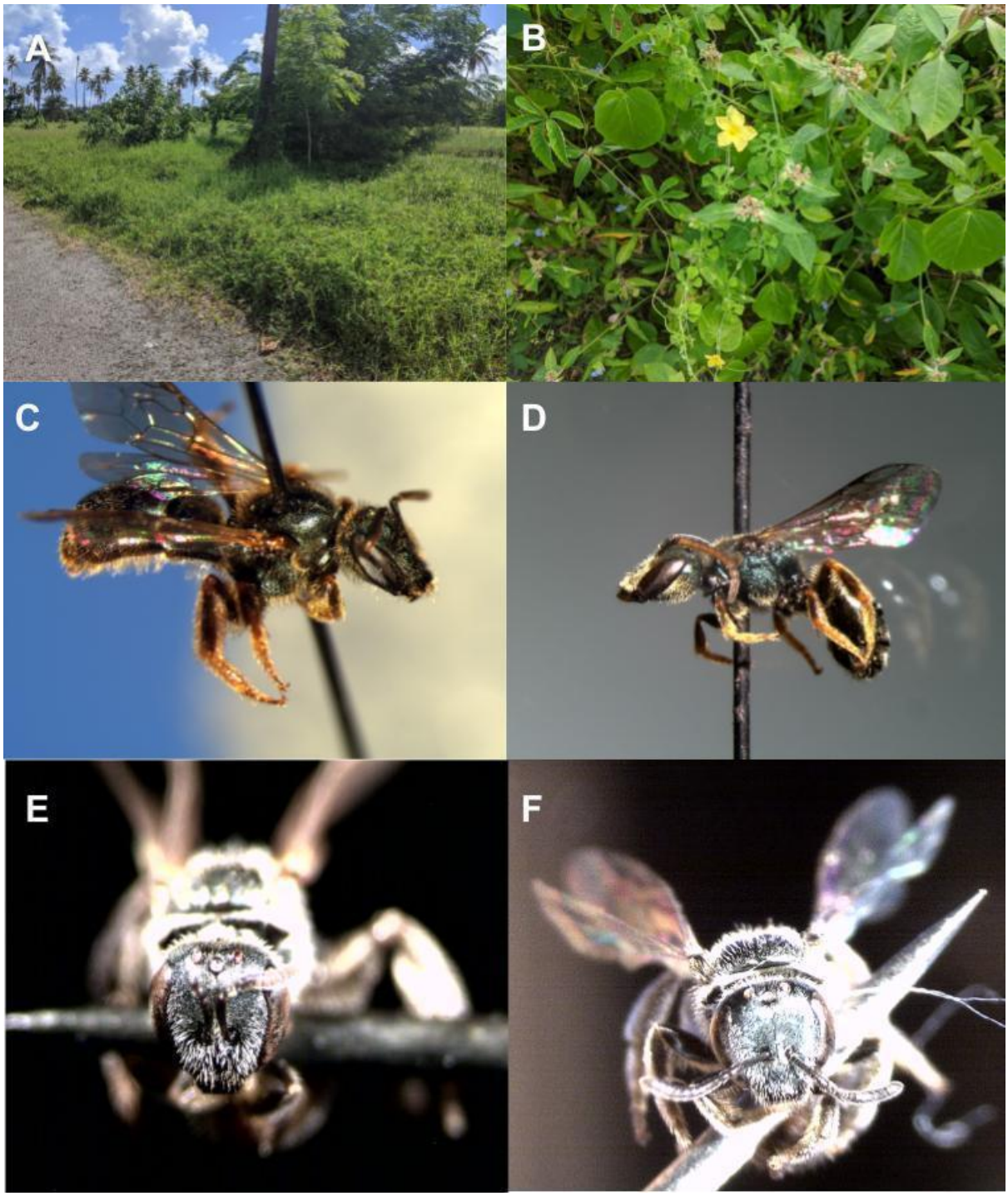


Figure 1: Habitat (A-B) and Species Observed (C-F).

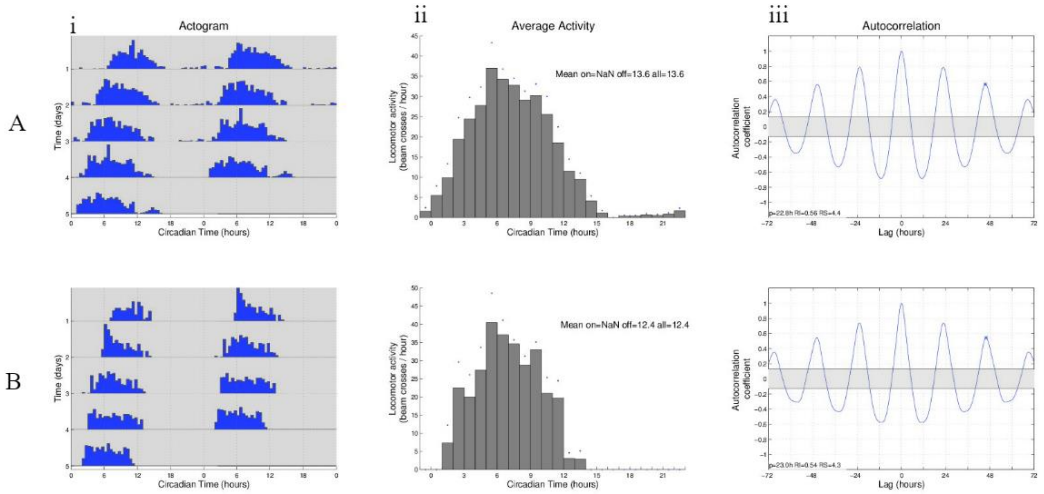


Figure 2: Female *S.curvicornis* exhibit short period phenotype under constant darkness (<24 h endogenous circadian rhythm)

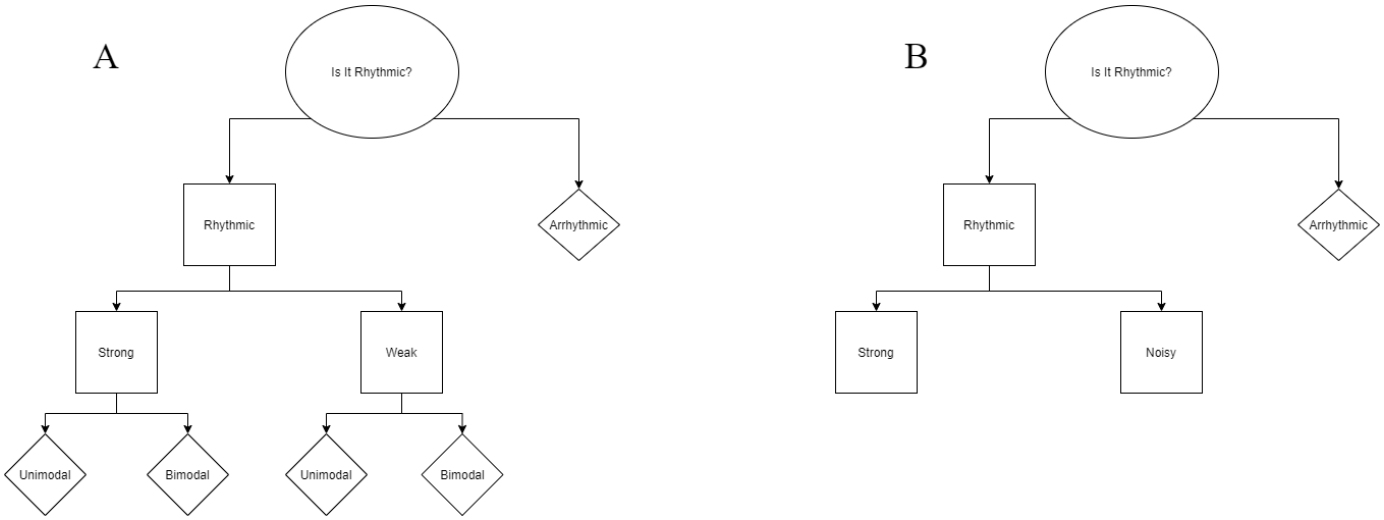


Figure 3: A summary of the variations in the circadian rhythm as observed in: A) *Lasioglossum malachurum*, B) *Lasioglossum enatum* and *Lasioglossum ferreirii*

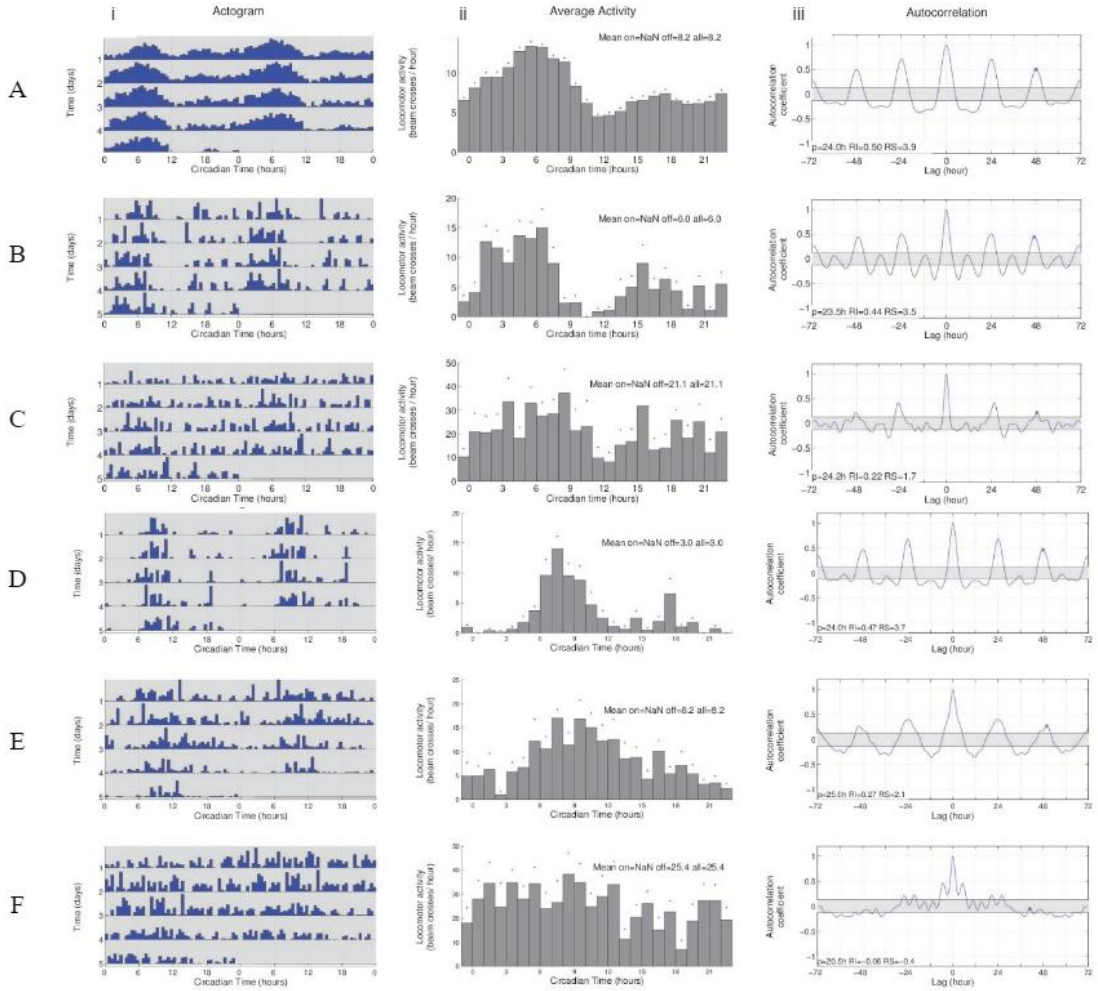


Figure 4: L.malachurum exhibits a variety of circadian phenotypes under constant dark conditions.

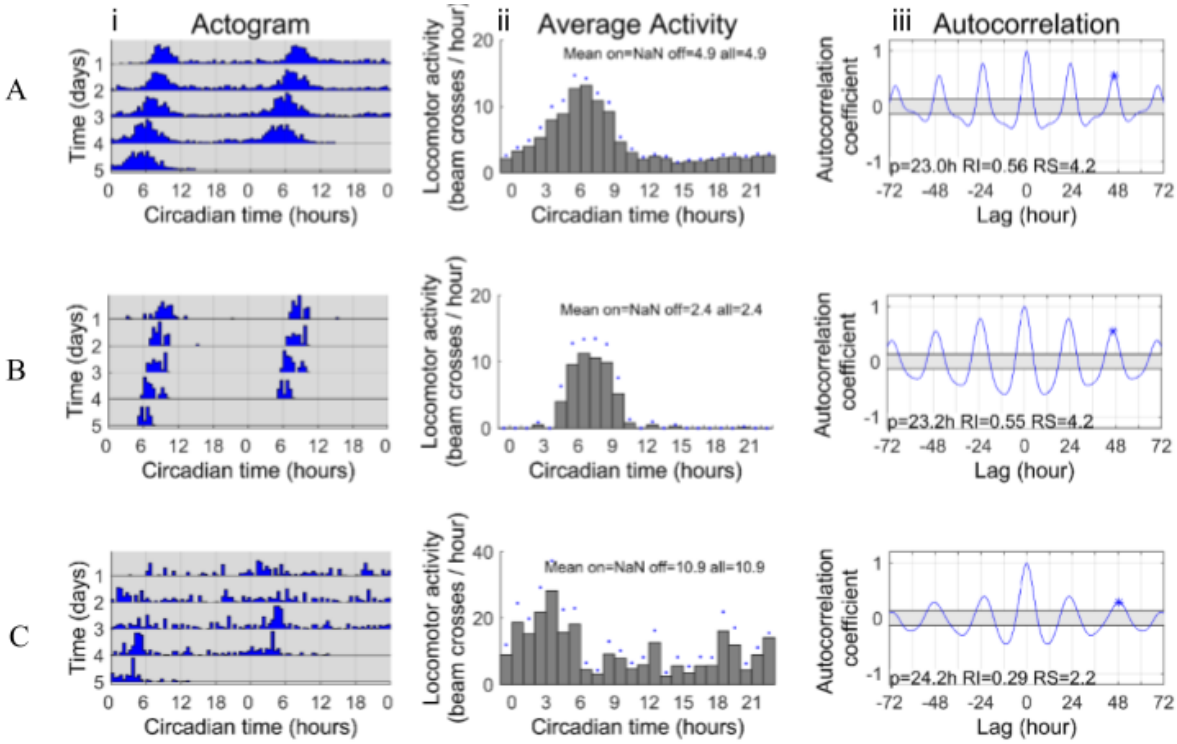


Figure 5: Description of the circadian behaviors exhibited by *L. ferreirii* under constant dark conditions

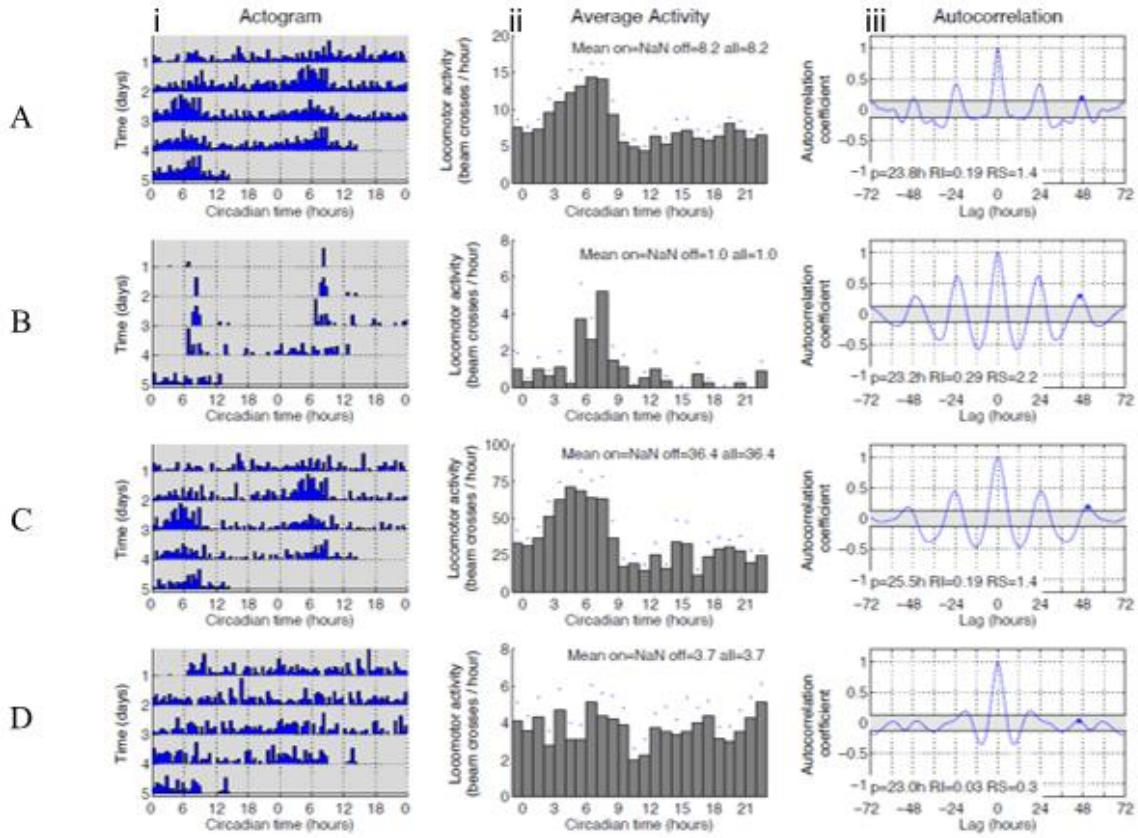


Figure 6: Description of the circadian behaviors exhibited by *L. enatum* under constant dark conditions.

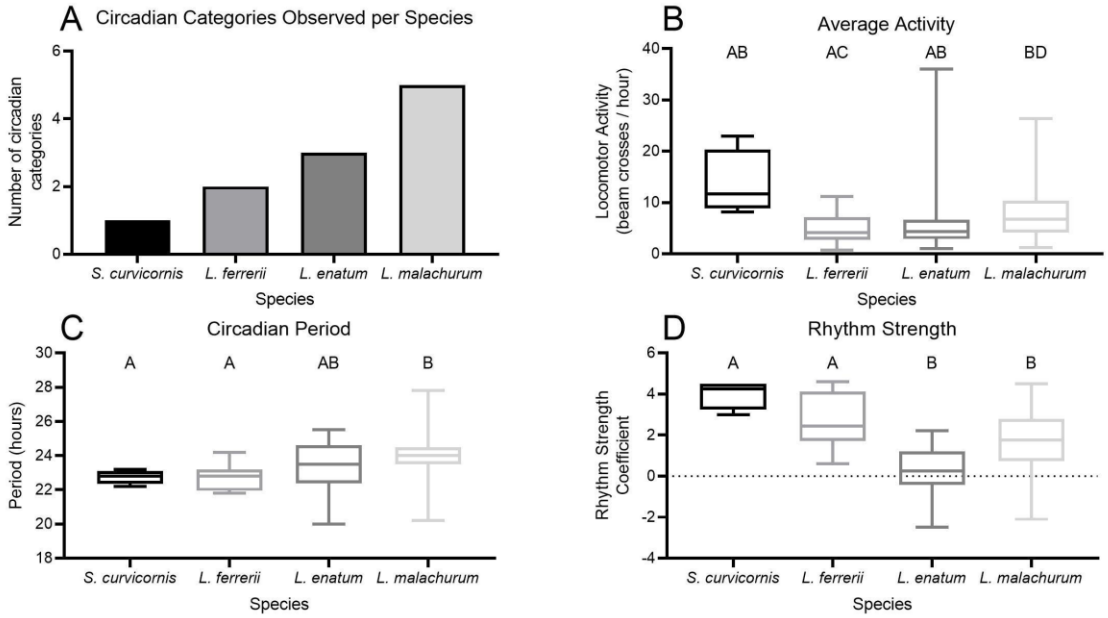


Figure 7: Summary of descriptive and inferential statistics

Chapter 2

A Machine Learning Pipeline for the Classification of Inter-individual Behaviors in Circadian Rhythms of *Lasioglossum* Bees

A Machine Learning Pipeline for the Classification of Inter-individual Behaviors in Circadian Rhythms of *Lasioglossum* Bees

Abstract:

Traditionally, model organisms are used for the study of daily activity patterns. Because the genetics of these organisms are well-known, there has been no need to do systematic *a priori* sorting of the individuals into groups, the categories are built into the analysis pipeline. Recently, there has been a rising interest in using non model organisms in studies, and with them comes unexplored diversity. To facilitate the *a priori* sorting of individuals for analysis, we designed a machine learning pipeline using individuals of *Lasioglossum malachurum* as a case study. We tested both supervised and unsupervised algorithms, and evaluated how well they separated the individuals of the population in discrete groups based on the phenotype of their daily activity patterns. Decision Trees with Symbolic Aggregate approXimation (SAX) transformation achieved the best results. K Nearest Neighbors with Symbolic Aggregate approXimation was a close second. In the future, we aim to test the transferability of this pipeline using insects of the same genus, but different expressions of behavior.

Introduction:

The study of circadian rhythms focuses on characterizing the daily activity patterns of populations. These daily activity patterns are constrained to a period of time that approximates 24 hours and are consistent with the rotation of Earth. Traditionally, model systems, such as fruit flies, mice, rats, and *E. coli* cell lines, are used to evaluate circadian rhythms. Because these models are well-known in terms

of phenotype, no *a priori* systematic sorting of the data sets is usually done, the scientist already knows which individuals have which genotypes and can sort them accordingly.

Recent works in behavioral biology and circadian rhythms have evaluated the use of non-model systems, such as varying species of bees and have found them to be ideal systems to study due to the diversity of behaviors exhibited (Bloch and Grozinger 2011). In particular, halictid bees have been suggested for studies that evaluate the impact of sociality on other behaviors, for example circadian rhythms (Toth and Rehan 2016, Bloch and Grozinger 2011, Danforth et al. 2003). However, because bees express a high diversity of behaviors, they present an added layer of difficulty in terms of systematic evaluation. For example, Gianoni-Guzman and collaborators (2014) saw social insects in comparison to *Drosophila melanogaster* (the fruit fly) displayed a significant amount of variance in terms of period. This means that the length of the internal clock of these social insects (which includes honey bees) varied from individual to individual. In a different work, Gianoni-Guzman (2020) tracked the foraging schedules of honey bees. He found three distinct temporal shift behaviors. The variance in the first study, where one of the populations evaluated was of honey bee foragers, can be explained by the behaviors observed in the second. Studies like Gianoni-Guzman 2014, where there is no prior knowledge of the phenotypes present in a population being evaluated, may benefit from *a priori* sorting based on similarity found within subsets of the population. Studies like that of Gianoni-Guzman (2020) may be of use to train a classification pipeline and serve as a sorting tool for future studies. In cases where no prior labels

are available, a clustering pipeline in conjunction with expert validation can inform the labeling process. Expert validation is defined as having an expert evaluate the clusters and validate their biological significance.

In this work, we evaluate the addition of a grouping step to the circadian evaluation pipeline, as to facilitate the determination of the groups in the non-model systems. To achieve this, we applied both clustering (K-mean and Partitioning Around Medoids) and classification (K Nearest Neighbors, Decision Trees and Random Forest) on a locomotion activity dataset for the halictid bee *Lasioglossum malachurum* (*L. malachurum*). In addition, we transformed the time series of each individual using Symbolic Aggregate approXimation (SAX) to reduce dimensionality in the data set, so that the computational resources needed to evaluate this data set are minimized. Given how circadian activity is traditionally analyzed, we expect that the methods that best capture the shape of daily activity patterns will be the most effective at grouping the bees in groups of circadian significance.

High Level View of the data:

The following 3 sets of graph panels in Figure 1 are examples of what the data looks like after it’s analyzed with a traditional circadian pipeline. Visually, their activity patterns do not look similar. The individual in Figure 1. A is extremely rhythmic, which makes it so the graph of the average daily activity looks fairly similar to the activity of any one day in the actogram. This individual appears to have a bimodal pattern of activity, where the main bout of activity occurs in between the sixth and twelfth hours of the day, and the minor one occurs about an hour later after a period of inactivity. The Lomb Scargle Periodogram suggests that a day for this

individual last about 23.77 hours, and the autocorrelation oscillates, which indicates strong rhythmicity.

In contrast, the individual in Figure 1. B appears to be a repeated yet spread out bimodal pattern. That spreading of the locomotor activity already makes it different from the individual in Figure 1. A. They also differ in the time of day in which they are active. The average activity plot in Figure 1. B has a similar shape to the days of observation, but there is a distinct difference between it and Figure 1.A.. One can see that the gap in activity isn't exactly the same for every day. The Periodogram suggests that this individual's day lasts 22.85 hours, which is slightly shorter than the previous individual's day. The autocorrelation graph does not have as well-defined undulations as the previous individual, which suggests that the individual Figure 1. B, although rhythmic, is not as strongly rhythmic as the subject in Figure 1. A.

Lastly, Figure 1. C is an example of an individual whose rhythm is not circadian (24h), but might be ultradian (12h). Different from the other two Figures, it's difficult to determine a pattern from looking at the actogram, and by consequence, it's not easy to tell if the average is in any way representative of the individual. This makes us rely on the autocorrelation to determine rhythm. As one can see in the autocorrelation plot, this one has a smaller undulations pattern, although it rarely reaches significance. Those small undulations can lead one to believe that this individual might have a much shorter internal day than what we are accustomed to analyzing. Therefore, the periodogram might not be useful to determine the period for these cases.

In all, these three individuals demonstrate that there exists a level of diversity within the data set that might be lost if we evaluate it only using averages. In Figure 2, this theory is illustrated with a single plotted actogram, which represents the activity of all 98 individuals present in this data set. The activity of any one of the days in Figure 2 is not representative of any one day illustrated in the individuals in Figure 1.

To report averages as representative of the species in terms of circadian rhythm, in this case, is misleading. Nevertheless, describing each individual is tedious work. As an alternative, we should be able to group together individuals with similar characteristics and use them to describe the daily activity patterns of the species. To achieve this, we explored the use of both unsupervised and supervised machine learning with an abstract knowledge representation of time series. Because the methods used to evaluate circadian data are dependent on the mean and shape of the activity, we expect that methods that utilize centroids will be the most effective, and we chose SAX for its ability to cluster and classify univariate time series by shape.

Methods:

Animal Model:

Lasioglossum malachurum (L. malachurum)(Kirby, 1802)

L. malachurum is an obligately eusocial halictid bee, also known as sweat bees. They typically nest underground in complex colonies (Wyman, L. M., & Richards, M. H. 2003).

Data collection:

Site description:

The specimens were collected on Lesvos (39°10’N 26°20’E), a Greek island in the northern Aegean Sea off the coast of Turkey. It was summer during the time of collection, and the specimens were captured after being observed between the hours of 0600 and 0900 (Cordero-Martínez, C.S., et al. 2017).

Capture methods

The bees were hand captured as they visited the flowers of *Convolvulus arvensis*, colloquially known as morning glories (Cordero-Martínez, C.S., et al. 2017).

Housing and observation

Each bee was housed individually in a modified centrifuge tube, first in oscillating conditions and later in constant conditions. Oscillating conditions were meant to mimic the changes of light, temperature and humidity in their natural environment, while constant conditions kept all environmental cues constant. While in constant conditions, the bees were not exposed to light. At the bottom of the tube that housed the bee, there was a cotton ball soaked in water for hydration that was refilled every 2-3 days. The body of the tube had small holes to allow for air circulation. On the cap, there was a paste that would function as food for the bee. The paste was composed of a modified version of ApiYen brand bee feed that had no protein, but kept the same amount of sugar.

These tubes were placed into TriKinetics’ Locomotion Activity monitors (LAM), which in turn were put inside GRW-20 CMP3/TBLIN incubators. The incubators were set to mimic the environmental conditions in which *L. malachurum* was captured (Cordero-Martínez, C.S., et al. 2017).

Data exploration:

The code for this process is available at: https://github.com/ComplejoC/CircadianThesis

Data formatting

The data is outputted by the LAM as a data frame contained within a tabular separated value file (.tsv), as specified by the manufacturer in the user manual (https://www.trikinetics.com/Downloads/DAMSystem3%20Software%20Data%20Sheet.pdf). In this data sheet, columns 11-42 are representative of one individual, and each row is the number of times that the subject moved, as detected by the sensor in one minute.

The LAM system is prone to a number of errors, of which we must account for while analyzing the data. If a subject were to fall asleep or otherwise become immobile directly on top of the beam, the system may display the individual as more active than they are in actuality. If someone were to open the doors of the incubator, the sudden entrance of an outside light source can break the beam, and it will count as activity. In our case, we do not need to worry about the lights inside the incubator due to the observations being done in Dark-Dark conditions (DD). Nevertheless, it's important to keep in mind that turning on the lights inside the incubator may cause false activity counts.

Data processing

To be able to properly represent the data, we had to remove dead individuals. Including them in the study would influence the shape of the data, and might not accurately represent the actual behavior of the individual. We used the death

detection algorithm from the Rethomics framework of R packages (Geissmann Q, Garcia Rodriguez L, Beckwith EJ, Gilestro GF. 2019).

When graphing autocorrelations, it is common to detrend to reduce the influence of distortion that is inherent in this type of data. To do this, we used the detreand function from the pracma R package (Hans W. Borchers 2019). Because we wanted to see how the data correlated with itself in both directions, we used the cross correlation function from base R (ccf)(R Core Team 2019) and plotted it using the autoplot function from ggfortify (Yuan Tang 2016), resulting in a bidirectional autocorrelation instead of the unidirectional autocorrelation available in base R.

Visualization:

The double plotted actogram, and the Lomb–Scargle Periodograms were all generated using the Rethomics pipeline (Geissmann Q, Garcia Rodriguez L, Beckwith EJ, Gilestro GF. 2019).

Double Plotted Actogram:

This type of visualization is typically used to represent the Circadian locomotor activity rhythms. Each vertical bar is representative of activity, in our case the number of times the bee broke the sensor laser. The higher the bar, the more active the individual. It's called double plotted, because with the exception of the first day (and maybe the last), every day is plotted twice, two consecutive days are plotted one next to each other, and the second day being re-plotted in the right half right under (Jud, C. et al., 2005).

Lomb-Scargle Periodograms:

Lomb-Scargle is an algorithm that can be used to describe the period of unevenly sampled time series data sets. It allows for approximation of a power spectrum estimator, similar to that of a Fourier transform. The resulting estimators can be used to determine the period of oscillation of a data set (Jacob T. VanderPlas, 2018). In our case, those periods are representative of the length of the day in the internal clocks of our subjects. For ease of visualization, the data was downsampled from 1 minute bins to 30 minute bins. This down-sampled data was used to make average activity plots using ggplot2 (H. Wickman. 2016) and the autocorrelation plots.

Average Activity plots:

These plots are representative of the average activity done by each individual for the duration of the study. We added all days together into one representative day, and divided it by the total time accumulated to get the average.

Autocorrelation plots:

The autocorrelation of a time-series measures how similar a time-series is with a forward or backward shifted version of itself. For signals that oscillate perfectly, the graph of this function oscillates between +1 and -1, t=0 is the highest value registered. To create our autocorrelation plots, we detrended our data using pracma (Hans W. Borchers 2019) and did a self-cross correlation using the stats ccf

function (R Core Team 2019) to see how the data matched itself in both directions.

Clustering Methodology:

Clustering is a type of unsupervised machine learning, in which the data analyzed is unlabeled. That is to say, the machine does not have a reference of how the data groups together, and must learn to do so without input from a user (Géron, 2018). The purpose of this study is to optimize the already existing pipeline for circadian analysis. Therefore, the data transformations used to reduce dimensionality for clustering are those that are already used in circadian science: The Lomb Scargle (LS) Periodogram, Average Daily Activity and Autocorrelations.

Consensus Clustering

Before attempting to cluster, we did a procedure known as consensus clustering. It is a “method to represent the consensus across multiple runs of a clustering algorithm to assess the stability of the discovered clusters” (Monti, S., 2003). The use of the word stability in these instances refers to how much the composition of each cluster changes over repetitions.

We wanted to input a number of clusters for our unsupervised algorithm that were not arbitrary. By doing various repetitions of the clustering procedure and observing which number of clusters has the minimum error, we are allowing the data to speak for itself instead of using our prejudice. For this, we used Wilkerson and Niel’s R package consensusclusterplus that can be found in bioconductor (Wilkerson, D. M and Hayes, Neil D 2010). We used a maximum k of 10 and 10,000

repetitions, using both Euclidean and Manhattan distances. Ultimately, the Manhattan distance was used to minimize cluster collisions.

K-means clustering

In one instance, we use K-means to cluster our data. The algorithm takes an n number of observations or individuals in our case, and divides it into K clusters. K is the number of clusters as provided by the user. The original centroids around which the data clusters are randomly set and require a seed for replicability in the code. We used the set.seed function from Base R (R Core Team 2019) to generate the centroids, and ran the kmeans function from the stats package (R Core Team 2019). Once the first centroids are set, each individual becomes a member of the cluster with the nearest mean, serving as a prototype of the cluster. Then the algorithm continues, by adjusting the centroids until a partition of the data that minimizes the sum of squares deviation is found (H.-P. Kriegel et al. 2017).

PAM (K-Medoids)

In two instances, to cluster our data, we used the PAM algorithm, also known as k-Medoids, from the R package cluster (Maechler, M et al., 2019). This algorithm is considered a more robust version of K-means, because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances. Just like in K-means, the algorithm is given a K number of centers, the difference being that centers or medoids are actual points within the data set (L. Kaufman P. J. Rousseeuw, 1990).

We clustered the data using 3 different transformations: Average Daily Activity, the AutoCorrelation Coefficient and The Lomb–Scargle periodogram.

Cluster visualization

Because we have 97 individuals with 5 days' worth of locomotion data, graphing the clusters means we have to reduce dimensions to make them human readable. To do this, we used the R package factoextra (Alboukadel Kassambara and Fabian Mundt, 2020), which graphs data by using the two most representative dimensions of the whole and using them as the x/y-axis.

Classification Pipeline:

Phase 1: Creating the Gold Standard Part 1 Building the Baseline

1. Naive Classification:
    a. Teach a group of non-experts how to interpret the graphs usually used for circadian analysis.
    b. Separate them into smaller groups and give them a stack of graphs that they will divide into 3 groups: Rhythmic(R), Arhythmic (AR) and Weakly Rhythmic (WR).
    c. Measure the percentage of classification coincidence between the groups.
2. Experienced (Round two using the same people):
    a. Give a review on how to classify the graphs.
    b. Separate them once again into smaller groups and give them the same stack of graphs that they will divide into the same 3 groups as last time.
    c. Measure the percentage of classification coincidence between the groups.
    d. Compare the percentages of coincidence of the groups between rounds.

Phase 3: Data Transformation

Time series data is famously noisy and unevenly sampled. To remedy the pitfall of using this type of data, one may use transformation methods to make it more manageable, such as Fourier Transformations, Discrete Wavelets, and Lomb Scargle periodograms (Refienetti et al. 2007). We decided to use Symbolic Aggregate ApproXimation or SAX for short. Our choice was based on the capacity of this algorithm to do dimensionality reduction while still staying true to the data. This goes in hand with the other advantage of using SAX, which is that it has lower bounds (J. Lin, 2007). This means that it has the capacity to represent our time series while using a minimum amount of resources.

1. **Z-normalization**

   Z-normalization, also known as z-score normalization and "Normalization to Zero Mean and Unit of Energy" is a normalization method first mentioned by Goldin & Kanellakis (1995). The functionality of this method of time series normalization is to take the elements of an input vector and transform them into an output vector whose mean is approximately 0 with a standard deviation close to 1. The formula to achieve z-normalization is:

$$\frac{x-\mu}{\sigma}$$

   Were x is an element within the time series, $\mu$is the mean value within the time series and σ is the standard deviation of the time series.

2. **Symbolic Aggregate ApproXimation (SAX)**

   This algorithm takes a z-normalized time series and transforms it into

symbolic representation to create words. Typically, any function using this method will have four parameters: The size of the window of time being observed, the length of the words that represent the data within that window, the size of the alphabet from which the words are built, and the type of numerosity reduction (P. Ordoñez et al., 2011). This data representation captures the shape of a time series, while also simplifying and facilitating pattern detection (J. Lin and Y. Li, 2009). For the purposes of implementation in this work, PAA size will be the equivalent of word size, size of the window of time being observed will be referred to as sliding window, and number of letters being used is alphabet size. In Figure 3, is a cartoon of how the transformation would look like for a sliding window of 180, PAA size of 7 and an alphabet size of 4. The time series in Figure 3 would be represented by the symbolic word CBDCCCC.

3. **Bag of words (BoW)**

Once the data is transformed, it is necessary to quantify it. BoW does this by quantizing each extracted word and then counting the frequency of each individual word contained in the time series (Senin & Malinchik, 2013). The final output of this is a table for each individual subject of study for whom we have a “count” of how many words represent them, and how often those words repeat. This by itself is not enough, as it does not highlight how much every word contributes to the overall shape of the time series.

4. **Term frequency–inverse document frequency (TF-IDF)**

This is a type of numerical statistic that measures how much a word

contributes to the overall shape of the data (Neto. J, et al., 2000). Words that are unique within the data set are considered to weigh more and have a stronger effect on the shape of the data. For example, many bees in a hive may display highly rhythmic behaviours, but a group of them may only be active in the morning, another only in the afternoon, while a third group may be active all day long. The time series for a unique bee will be similar in shape to those who are active in the same shift. Therefore, all of the morning bees will have a distinct shape from the afternoon bees and in turn those two will have a different shape from the constant workers, even though all these bees may exhibit rhythmic behaviour.

Data Sets:

Once transformed using SAX, the data set was separated into the training and testing data sets. 90% of the individuals were randomly assigned to the training set, while the rest were used for the testing set.

Phase 4: Supervised Learning

Supervised machine learning is a type of machine learning where some of the data being analyzed already has the desired solutions (labels). The type of algorithms that use labels to learn an established pattern typically have one of two uses. One is to predict a target value based on a given set of numerical characteristics (features), this type of task is called regression. The other is the most common use for supervised learning, which is classification, or separating data into categories. In our case, the algorithm learns the patterns from the labeled data set

and uses this information to classify new data into the same categories (Géron, 2018).

5. **K Nearest Neighbors (KNN)**

The KNN algorithm, as used for classification purposes, is a non-parametric method that uses plurality to determine membership within a group (Bezdek et al., 1986). The user assigns a number K of minimum nearest neighbors that an individual being evaluated must have to be assigned a label (Bhatia, 2010) (Figure 4). The distances between neighbors are measured with Euclidean distances.

Decision Trees

Decision Trees consist of a unique central node that branches out into edges depending on the answer to binary questions. The binary questions represent a test on an attribute for a classification, each branch represents an outcome of the test which eventually ends in terminal nodes or leaves representative of the labels (Leonard, 2017). On the left-hand side on Figure 5 there is an example of how a decision tree could look like using our labels.

6. **Random Forest**

This algorithm is derived from Decision Trees. By definition, it is a combination of tree predictors, or the “forest”. Each tree depends on the values of a random vector sampled independently, and with the same distribution, for all other trees in the forest (Breiman, L. 2001). Based on a measurement of error, the best tree is chosen from the forest as the predictive model (Figure 5 right).

All the code for phases 3 and 4 can be found in:

https://github.com/ComplejoC/CircadianSAX.

Results:

Clustering:

Clustering by Lomb Scargle Periodogram

An intuitive way to start is by separating individuals by the length of their day or periods. We reasoned that individual with similar lengths of day would cluster together. After doing consensus clustering, it was determined that 4 clusters was the optimal way to minimize errors and to have zero overlap between the clusters. Figure 6 is an illustration on how those clusters look like after using the PAM algorithm.

Some clusters notably have more individuals than others, which illustrates that some mean lengths of days are more common than others. The requirements for cluster membership appear to be wider or shorter depending on the cluster. A closer inspection of the clusters, as illustrated in Figures 7 and 8, shows the characteristics of membership for each cluster.

One may be tempted to look at all four graphs while looking for a pattern in clustering. Normally, the way to tell this story would be to only show the periodogram, because having all things together may seem confusing. Nevertheless, it’s important to see all four graphs while validating the clusters for the biological interest, because that is how the circadian expert would evaluate the individuals.

All these clusters in Figures 7 and 8 have in common the shape of the power spectrum, and not the length of their day per se. This leads to bees with different

rhythms, but similarly shaped periodograms being put together. In some cases, like in Figure 7 B, the Lomb Scargle Periodogram could not determine a period for one of the individuals, but could determine it for the other. Furthermore, the period did not have to be similar for the individuals to group together. In some cases, we had individuals in the same cluster, for whom the difference in period would be 5 hours.

Clustering by Autocorrelation

The next intuitive step is to cluster based on autocorrelation, as it is one of the most commonly used measures of rhythmicity. Because of its typical use, it stands to reason that the algorithm would cluster based on rhythm, which is what we are looking to describe. After using Consensus clustering, we determined that PAM with a K of 3 would work best to cluster the bees (Figure 9).

We found no form of collision between the clusters (Figure 9). The dimensions of this plot do not appear to be representative of the data. Similarly, to a PCA, the sum of the percentages of both principal components should be as close to 100% as possible. There also seems to be an over-representation of individuals in cluster 1 and few individuals both in cluster 2 and 3. Because it is known that this data set is diverse, it is worrying to see that most of the individuals grouped together.

Contrary to our hypothesis, rhythmicity is not the feature on which the data is being clustered. It would appear that shape is once again the driving force behind our clusters. In the case of autocorrelation, it would appear to be the thickness of the autocorrelation plot.

The second and third cluster appear to group things together by the thickness of the autocorrelation, where cluster 3 is thicker than 2. Nevertheless, contrary to

what one would expect, it does not appear that all the individuals from these clusters have similar rhythms. To exemplify, in Figure 10. B., the thickness in the autocorrelation of both individuals is similar, but it's clear that the strength of their rhythms is contrasting.

Clustering by Average Activity

For this group we choose to do clustering in a different manner. By using consensus, we determined that the K-means algorithm with a K of 3 was the best approach for this transformation (Figure 11. D.).

The clusters for average daily activity are separated by magnitude of activity, with no consideration as to when the activity is being done. Neither of the locomotion plots (Figure 11. A. and B.) give the impression that any of the clusters have a common period of inactivity. The apparent lack of inactivity is the result of averaging individuals that are active in different times of the day. Figure 11. C illustrates that definitely, is the magnitude of activity that distinguishes all 3 clusters.

Classification:

K Nearest Neighbors (KNN):

To explore which transformation parameters optimized accurate classification, we explored all possible combinations of PAA size and Alphabet Size, as seen in Table 1, where PAA could be equal to: 3,4,6,8,12 or 24, and Alphabet Size could be: 3,4,5,6 or 7. Additionally, we kept Sliding Window Size equal to 48, as we wanted our results to be analyzed in the circadian 24 hours. The best overall scores were achieved with a combination PAA of 4 and Alphabet Size of 3 (Table 1). Overall, no

trend is evident when we experiment with the SAX parameters, i.e the alphabet size and PAA were typified by the combination of data set and classification algorithm.

Often precision for the weakly rhythmic category was NA, which also reflected on the F1 measure. Contrastingly, the rhythmic category was the one which most often had any form of measurement, but it was the arrhythmic category that kept the better scores, although it has a larger amount of NAs than rhythm.

Decision Trees and Random Forest:

With the intention of improving upon the results achieved from KNN, we tested the use of Decision Trees. After fitting the model using the rpart method from caret, we examined which Complexity Parameters were best for creating our predictions. Similarly, we build a Random Forest model using the same SAX parameters and the rf method for caret. In this case, we observed the number of randomly selected predictors to use in our prediction.

Overall, both of the models are consistently better than KNN, and in one instance with PAA 6 and Alphabet 3, Random Forest performed the best with an accuracy of 0.889 (Table 2). Nevertheless, although both tree algorithms had better accuracy, whenever PAA or Alphabet size got larger than 4, the process became more computationally intensive. KNN on the other hand was not as computationally intensive, and on one occasion with PAA 4 and alphabet size of 3, achieved an accuracy of 0.8.

Just like in KNN, tree methods had a large occurrence of NAs in precision, recall and F1 for the weakly rhythmic category. In Figure 12, we illustrate how those NAs happen. In the example, we are using the case of PAA = 4 and Alphabet Size =

5 for KNN, but the same should apply for the other combinations of parameters and classification algorithms.

We calculated accuracy by taking the sum of all correctly classified individuals from Figure 12.A. and divide them by the total number of individuals evaluated. Figure 12.C. illustrates how we calculate Precision, Recall and F-score. Precision was calculated by dividing all the True Positives (TP) by the sum of the TP and the False Positives (FP). In the example, for the WR category, both the total of TP and the sum of TP and FP equals zero, and therefore the calculation leads to an undefined value. Recall is the TP divided by the TP plus the False Negatives (FN). This is illustrated in our example for WR, where there is zero TP and four FN, and when plugged into our formula, it results in recall equal zero. Lastly, F-Score is calculated by doubling the product of Precision and Recall, and dividing it by the sum of Precision and Recall. This last calculation is highly dependent on the results of the two before it, and if both values equal zero, for example, or is even an undefined value, then F-score will not be defined. This is the case in those instances where the results for all possible evaluation metrics for one label is equal to NA.

Discussion:

Clustering:

Clustering by Lomb Scargle Periodogram:

Using the Lomb Scargle periodogram to cluster our data resulted in clusters separated by the shape of the graph instead of the periodicity of each individual. This happened because periodograms are a power spectrum estimator and are usually used to evaluate the presence of oscillations within a dataset. While it does provide a

measurement of periodicity, that is one isolated point within the periodogram, and we were providing to the clustering algorithms all of the data points within the graph. That is to say that in this case, more information about the shape and oscillation of the data did not, in fact, give a clear circadian set of instructions to the clustering algorithm. At least not in terms of the length of the biological clock.

That being said, the clusters appear to have a loose sense of modality. In many cases like in Figure 7.A., the number of curves in the periodogram seem to be reflective of the number of peaks of activity of the bee. This sense of shape is likely what is informing the clusters and should be explored further with proper parameter tuning for the Lomb Scargle transformation.

Clustering by Autocorrelation:

Clustering by autocorrelation resulted in clusters that were either informed by the thickness of the autocorrelation or by simply not having a discernible pattern. The autocorrelation has its usefulness to circadian science in calculating period and rhythmicity. Knowing this, one would expect that rhythm would be the main informing feature to the clustering algorithm in this experiment. However, in reality, what best informed the composition of the clusters was once again the shape of the data. This is why there exist individuals within groups with different rhythmicities, but with a similar shape to their activity pattern.

Clustering by Average Activity:

In this last attempt, we not only managed to cluster in such a way that no collision was detected, but additionally, the characteristics of each cluster was clear and easy to define. Unfortunately, even though the individuals clustered

appropriately by average activity, the groups did not share circadian similarities, which was our goal. For example, the individuals in any given cluster may have the same amount of activity over time, but did not necessarily share the same length of day (period) or even be active during the same hours of the day, among other key aspects.

For all three univariate clustering, successful clustering does not equate to significant positive results in terms of a practical question. The fact that all three clustering applications did not yield a circadian significance does not imply a failure in part of the algorithms, but instead, it shines a light on how complex time series questions can be, and more so if we take the biological significance into consideration. The way a circadian scientist would divide individuals in a population into discrete groups is a multivariate process that would take into consideration all of the transformations of the data that we evaluated individually with clustering. Although other types of transformations could be considered for univariate clustering, a multivariate approach should also be considered. However, multivariate clustering is not a trivial pursuit. Using transformations that can inform the shape and intensity of the activity may definitely inform a good multivariate approach but developing the mathematical and architectural tools necessary for this is a thesis in itself. This is due to the size of time series data. We transform the data to reduce dimensionality, but that does not mean we eliminate the continuous time component when we transform, in fact we accentuate it. To reduce dimensionality further for the sake of clustering may do an injustice to the data set. Therefore, the best way to group this type of data set is one that conserves the fidelity of the data while also minimizing

the amount of computational resources necessary to achieve the task, which is why we did classification.

Classification:

When we validated the results of the clustering analysis, we noticed a number of patterns that appeared to be of circadian nature. Because multivariate clustering is a more complex problem than what we were equipped to handle, these patterns offered an alternative for analyzing these data. We created a user defined set of labels and had a group of experienced and naive users assigned those labels to the data set.

For all three classification algorithms, the weakly rhythmic category was the one that caused the most difficulties. Consistently, it was the category that most often returned NA in the evaluation metrics. This suggests that we should re-evaluate what makes an individual weakly rhythmic, or even subdivide it into smaller categories still. On the other hand, we could also consider that none of the three algorithms used may be appropriate for the type of data we are using. Nevertheless, for the Rhythmic and Arrhythmic categories, all three classification methods performed adequately, which leads us to believe that more appropriate labels are needed.

K Nearest Neighbors (KNN):

Our results suggest that combinations of smaller PAA size and Alphabet size parameters in SAX transformation yield better results when evaluating the model. More often than not, KNN did not return a model with accuracy higher than 0.5,

which suggests that the use of this classification algorithm requires fine-tuning for it to return correctly classified results.

Decision Trees and Random Forest:

In contrast, both tree algorithms consistently got an accuracy over 0.5, although only once did they outperform the highest KNN result. The optimal parameters for the best model using Decision Trees follows the tendency in the literature of smaller PAA and Alphabet size being the most optimal (J. Lin, 2007). Although one could argue that PAA 6 and Alphabet 3 are still relatively small. The better model came at the cost of computational power and took considerably longer to build.

Conclusions:

Clustering, although a good first step to gaining intuition for the behavior of the data, does not result in strong conclusions. For all three attempts, the algorithm used the parameter given in a way we did not expect. For example, for periods using the shape of the data rather than its values. While these experiments gave us a good intuition for the data set, the ultimately did not satisfy our need of grouping bees by circadian phenotypes. The next steps would be to attempt clustering with other transformations or even a multivariate analysis. Because our goal was to facilitate circadian analysis, we decided it would be simpler to group using methods that mimic how a circadian expert evaluates multiple parameters of an individual to describe its characteristics.

In our classification experiments for all possible combinations of parameters evaluated, the Weakly Rhythmic label caused the most difficulty in classification.

Often returning zeros and NAs for precision and recall. At this moment, we have not noticed any common patterns in the WR classifications. Nevertheless, this persistent difficulty in classification does inspire a reconsideration on how we are labeling these individuals. Because the individuals are not consistently being misclassified as rhythmic or arrhythmic, it stands to reason that the errors are not necessarily caused by the users mislabeling the individuals, but that just using three generalized labels could be obfuscating certain behaviors in a category. Therefore, experimenting with separating the weakly rhythmic category into further categories may facilitate the classification process. Furthermore, these more in-depth classifications can help reflect the biological reality of *L. malachurum.*

All is not lost, as we did build two different classification models with 0.80 accuracy or more. This demonstrated that at least two of the categories built with user input were adequate to inform a model. Of the two models The KNN one paired with SAX transformation parameters of PAA 3 and alphabet size 4 (Table 1) is the less accurate at 0.80, nevertheless it was the one that ran the fastest and used the least amount of computation resources. While the Decision Tree model with PAA 6 and Alphabet size 3 (Table 2) had an accuracy at 0.89 but took longer to run and more computational resources. Therefore, the choice of the best model is arguably determined by the resources available to the user and having more than one model from which to choose is beneficial for those who may not have much computing power available to them.

Ultimately, we successfully set the basis for an evaluation pipeline for circadian data that isn’t heterogeneous. This will undoubtedly facilitate the evaluation

of organisms that naturally express multiple phenotypes of circadian rhythms. The next steps are to test this pipeline with other organisms to observe whether or not it is indeed generalizable.

Figure Legends:

Figure 1: Locomotor activity over time data from *L. malachurum* displays heterogeneity. All the 3 figures contain a Double plotted Actogram, a Daily Average activity plot, a Lomb Scargle Periodogram and an Autocorrelation. A double plotted actogram is a visualization typically used to represent the Circadian locomotor activity rhythms. The daily average activity plot is representative of the average activity done by each individual for the duration of the study. Lomb Scargle periodogram is an algorithm that can be used to describe the period of unevenly sampled time series data sets. The autocorrelation of a time-series measures how similar a time-series is with a forward or backwards shifted version of itself. In **A**, **B**, and **C**, is a sample of the diversity of behaviours displayed in the data set.

Figure 2: Average of all 98 individuals in the dataset is not representative of any one individual due to heterogeneity. The single plotted actogram illustrated was created by taking the average of the activity for all 98 individuals done in the four day observational period.

Figure 3 A mock-up of a normalized time series transformed with SAX. A sliding window of 180, PAA size of 7 and an alphabet size of 4 together make the symbolic word CBDCCCC.

Figure 4: A mock-up of how KNN works. The pentagon is a new piece of data plotted into n-dimensional space here simplified as two-dimensional space. By consensus, the pentagon is closest to the group labeled AR and therefore will be classified as such.

Figure 5 Mock-up illustrating Decision Trees (left) and Random Forest (right). For Decision trees only one chart is considered as a classifier, whereas in Random Forest many trees are considered and the best tree is chosen based on a measurement of error

Figure 6: **PAM with K = 4 for the Lomb Scargle Periodogram resulted in highly representative discrete clusters.** We used principal components to plot the clusters, where the X and Y Axes are the principal components of the data set

Figure 7: **Clusters resulting from L. S are separated by the shape of the periodogram Part 1.** In green A) are examples of membership from cluster 1, characterized by two peaks in the periodogram that cross the horizontal line or the first barely does. In orange B) are examples of membership from cluster 2, where the individuals have no peak, or if they have one, it barely touches the horizontal line of the periodogram

Figure 8: Clusters resulting from L. S are separated by the shape of the periodogram Part 2. In purple A) are examples of the membership in cluster 3. Where individuals either have two peaks in the periodogram, where the first is far from touching the horizontal line, or just the one peak. In fuchsia B) are examples of membership of cluster 4. The individuals in this group have two peaks in their periodograms, but the first peak is small and sometimes unstable in terms of shape.

Figure 9: PAM with K =3 for Autocorrelation coefficient clusters of low representation power. Utilizing the same form of dimensionality reduction in Figure 3, we plotted the clusters resulting from PAM. Most of the individuals in the data set are clustered into cluster 1, while the rest divided into the other two clusters.

Figure 10: PAM with K =3 for Autocorrelation coefficient clusters by thickness of Autocorrelation. In green, A) are examples of the membership in cluster 1. Individuals in this cluster have a thinner autocorrelation graph than that of clusters 2 and 3, passing the horizontal line at most once. Nevertheless, the thickness does not seem to be consistent across the membership of cluster 1. In orange, B) are examples of the membership in cluster 2, where the individuals observed pass the horizontal line more than once and show larger density than cluster 1, but less than cluster 3. In purple, c) are examples of the membership in cluster 3, where the individuals observed have the thickest autocorrelation plot.

Figure 11: K-means with K= 3 for Average Daily Activity clusters by frequency. A) and B) show the shape of the activity contained within the clusters. In A), the graph shows the clusters individually, while B) shows them together. C) Illustrates the distribution of average activity within each cluster. D) Principal components illustration of the clusters.

Figure 12: Prevalence in NAs is due to poor consistent classification. Here we illustrate the process of calculating Accuracy, Precision, Recall and F-score, all measurements that we used to test how good our model is. These are the values taken from Table 1 for PAA = 4 and alphabet size = 5. On A is the test labels matched to the predicted labels, in red are the incorrectly classified individuals and in blue are the correctly classified individuals. In B. are the calculations for accuracy, which is the sum of all correctly classified individuals divided by the total of individuals, in addition there is also a confusion matrix. True positives (TP) are the values within each cell or the correctly classified individuals, false positives (FP) and

false negatives (FN) are the incorrectly classified individuals, which are indirectly viewed in the sum of columns or rows. Lastly, in C., a table illustrating the calculations for the rest of the measurements were Pre = Precision and Re = Recall.

Tables and Figures:

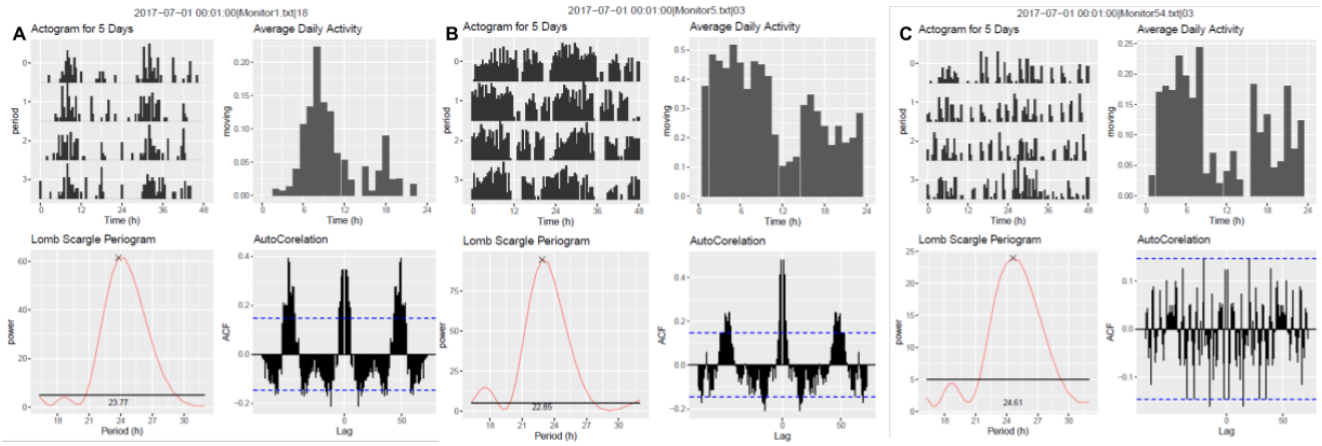


Figure 1: Locomotor activity over time data from *L. malachurum* displays heterogeneity.

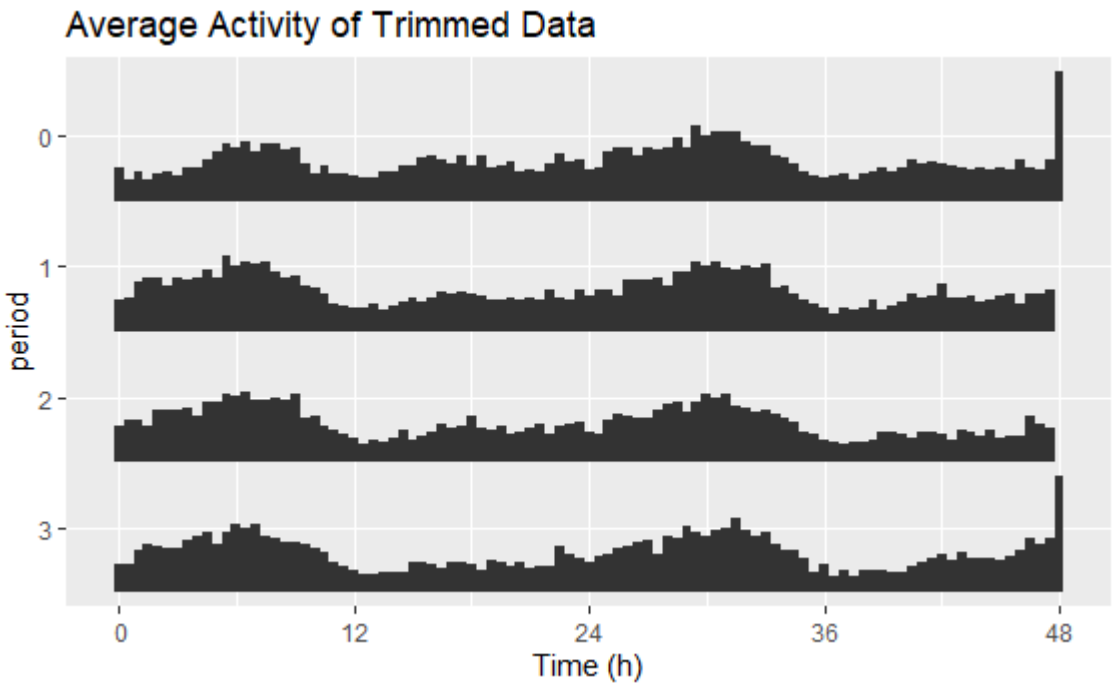


Figure 2: Average of all 98 individuals in the dataset is not representative of any one individual due to heterogeneity.

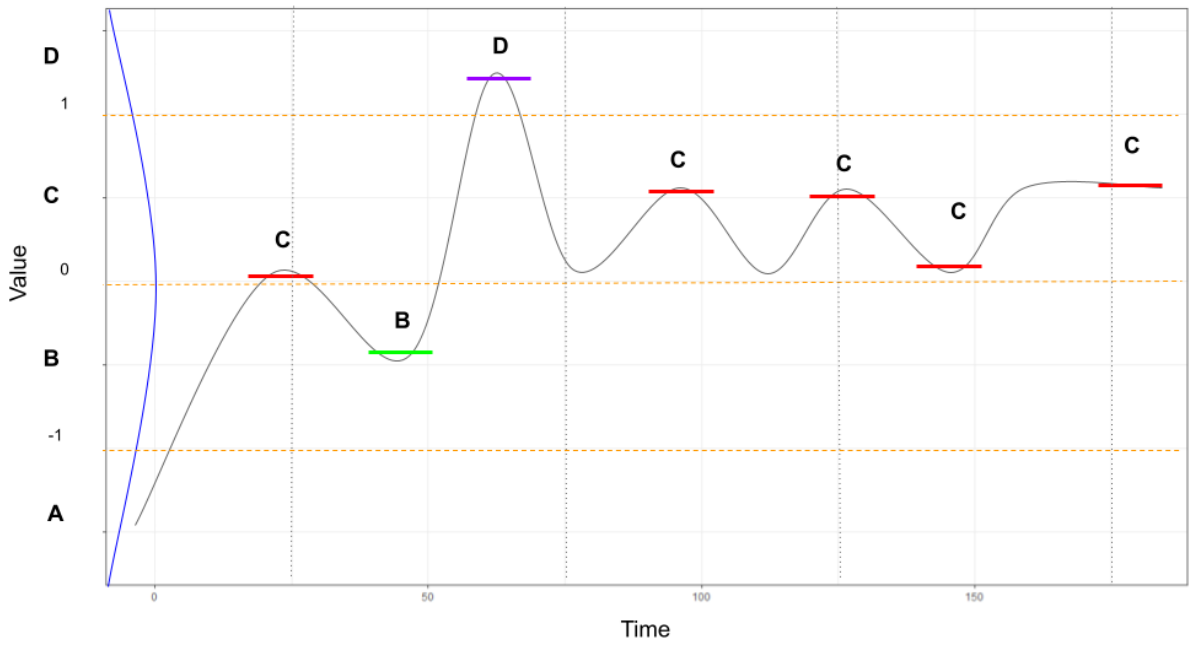


Figure 3 A mock-up of a normalized time series transformed with SAX.

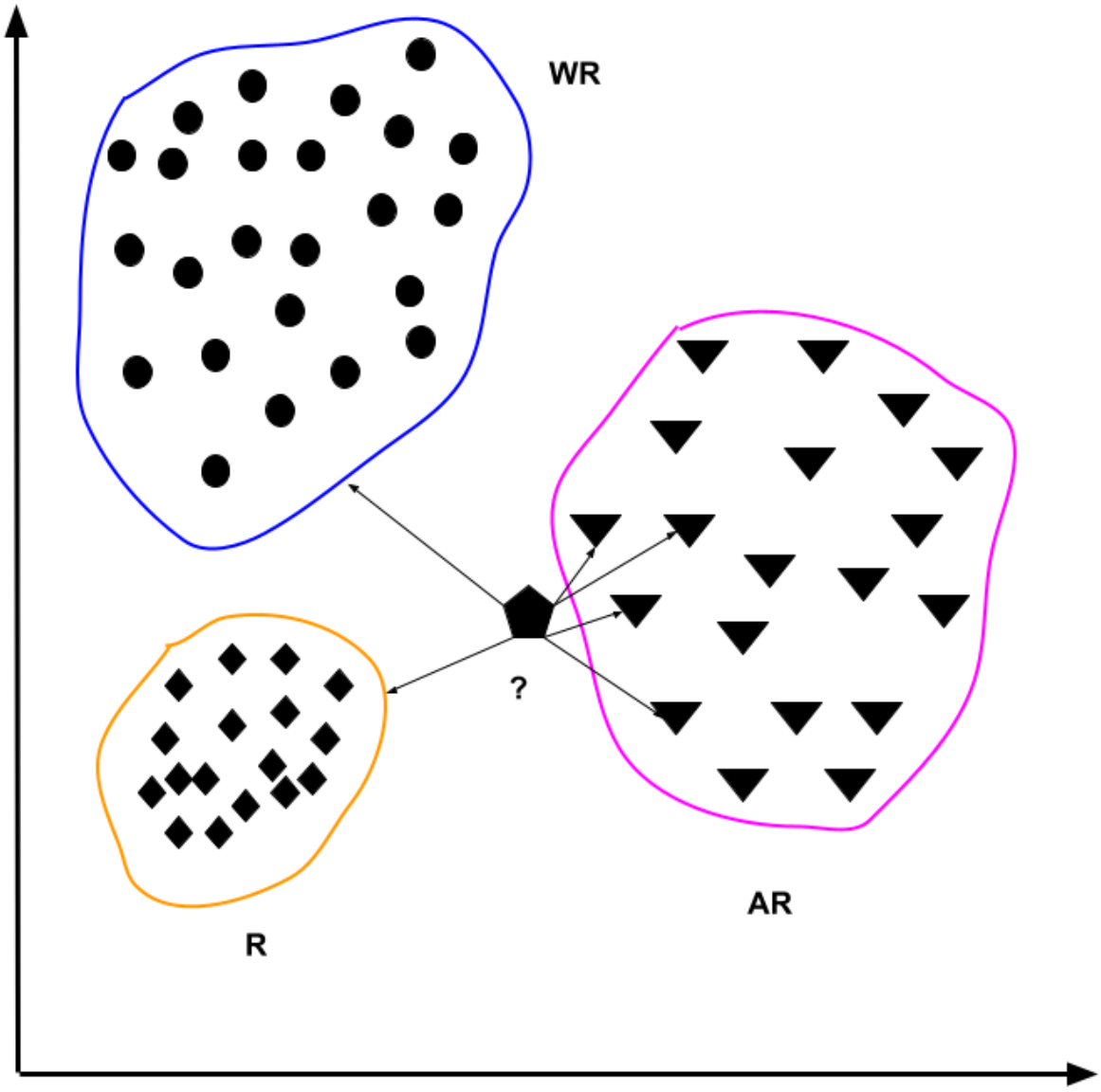


Figure 4: A mock-up of how KNN works.

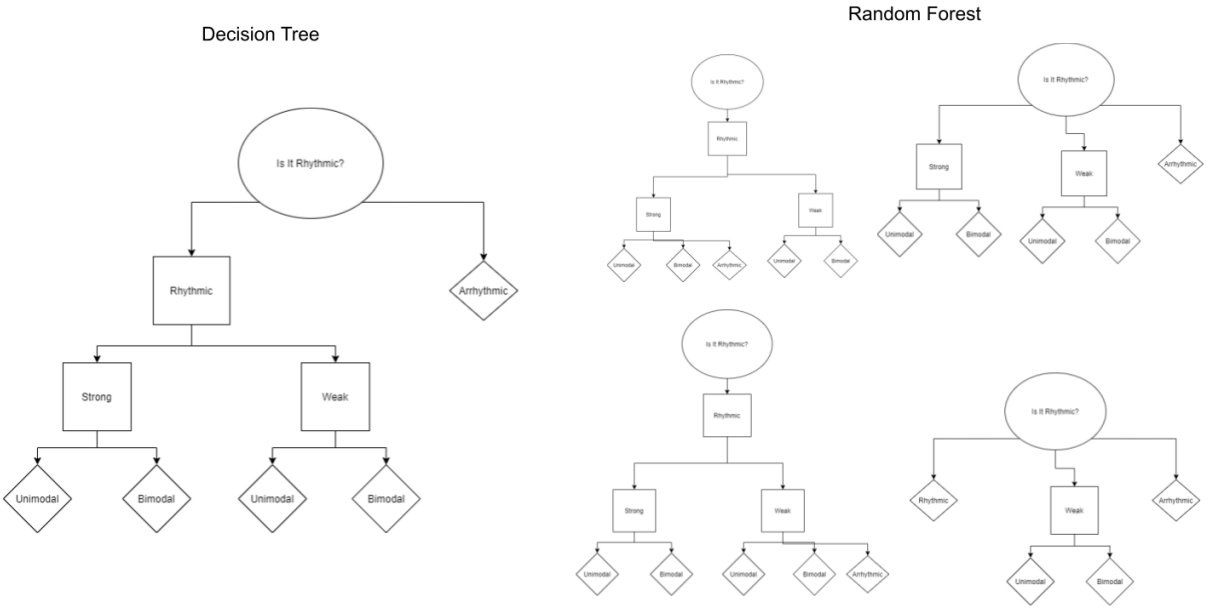


Figure 5 Mock-up illustrating Decision Trees (left) and Random Forest (right).

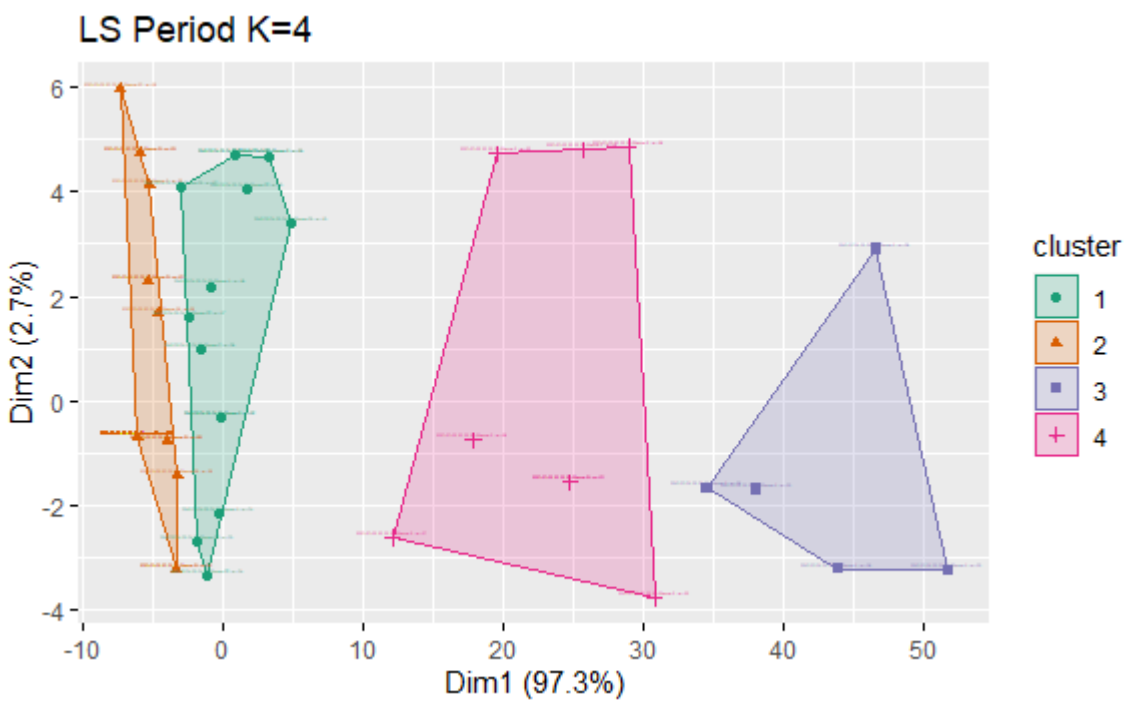


Figure 6: PAM with K = 4 for the Lomb Scargle Periodogram resulted in highly representative discrete clusters.

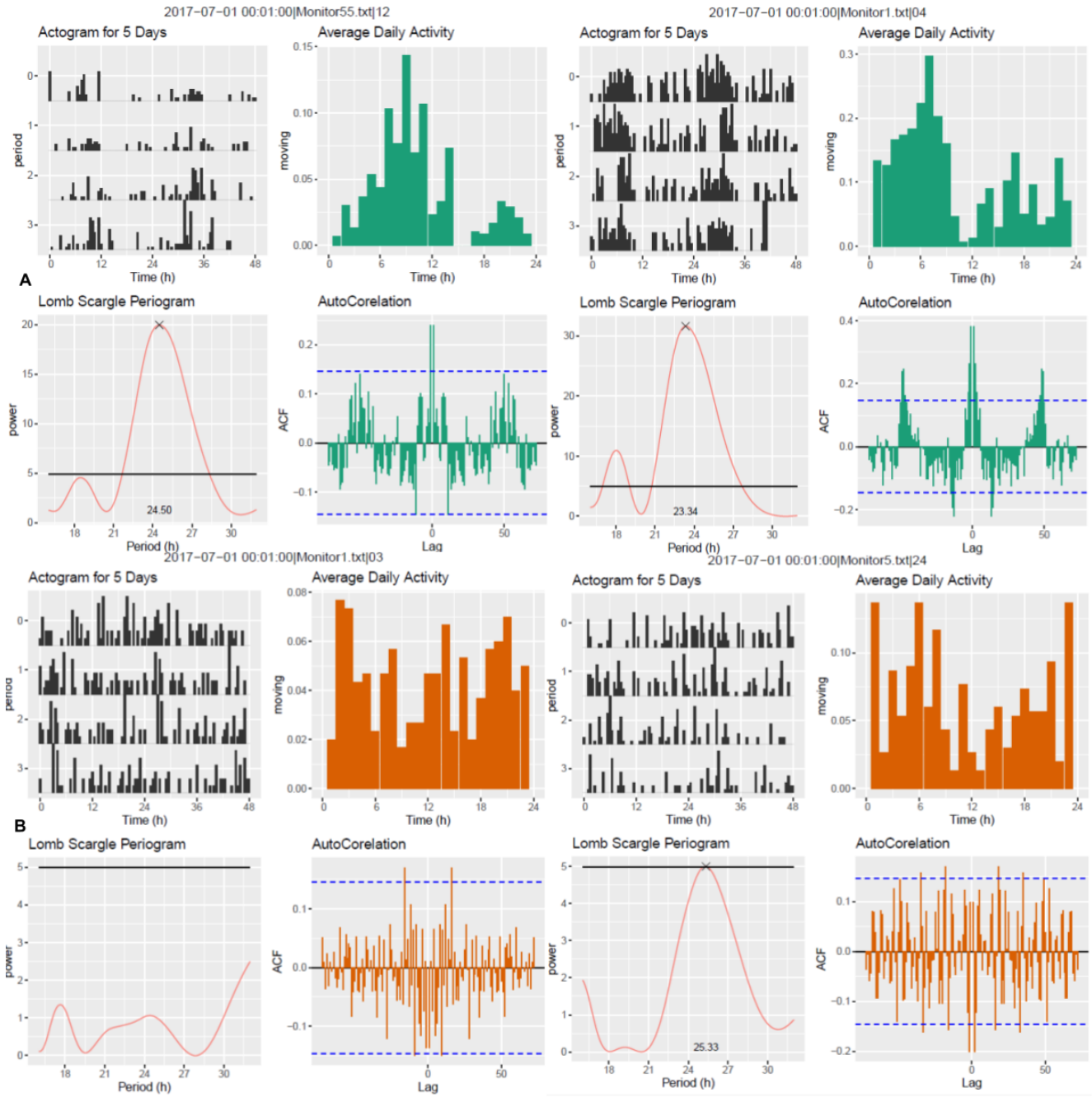


Figure 7: Clusters resulting from L. S are separated by the shape of the periodogram Part 1.

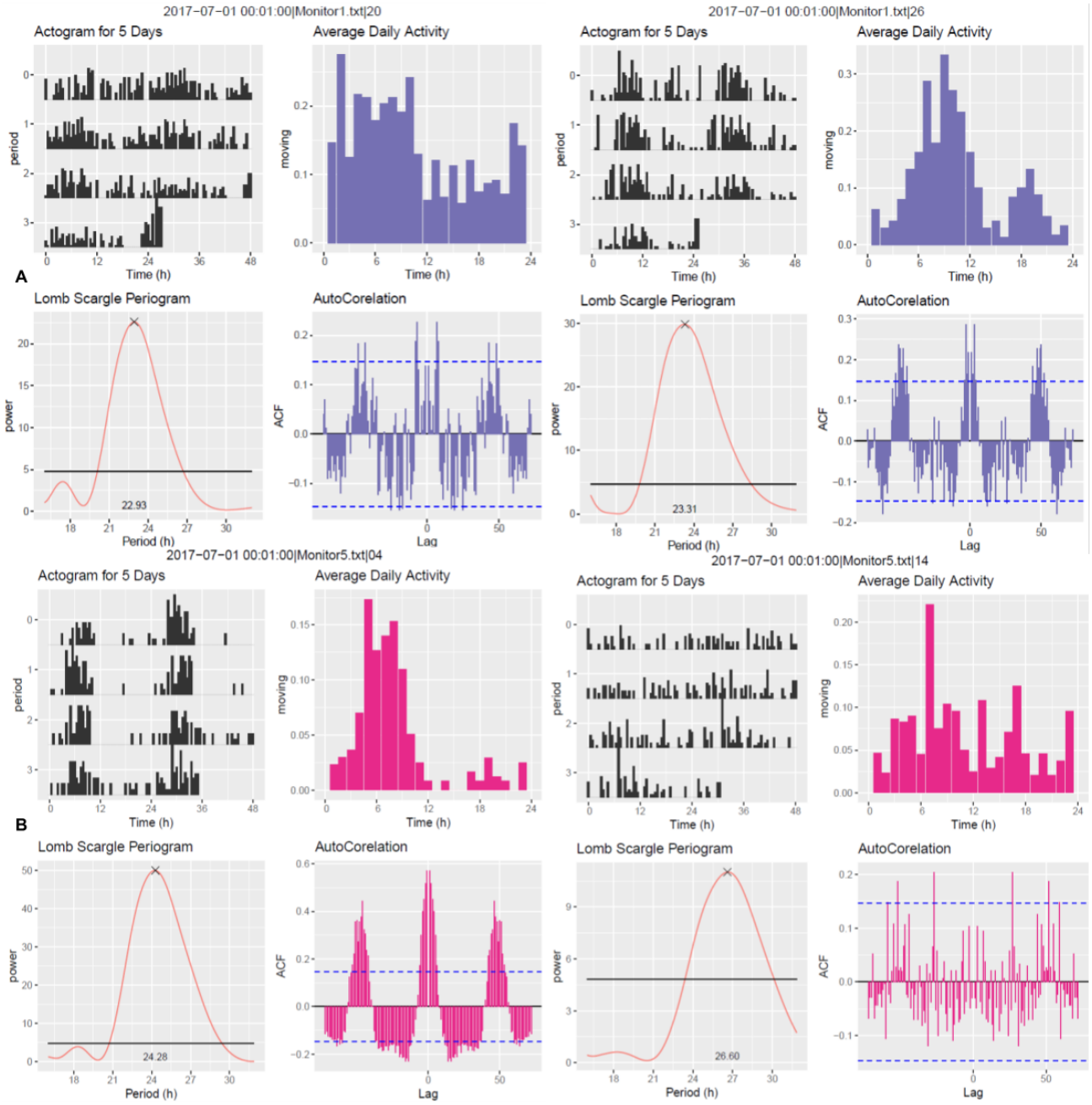


Figure 8: Clusters resulting from L. S are separated by the shape of the periodogram Part 2

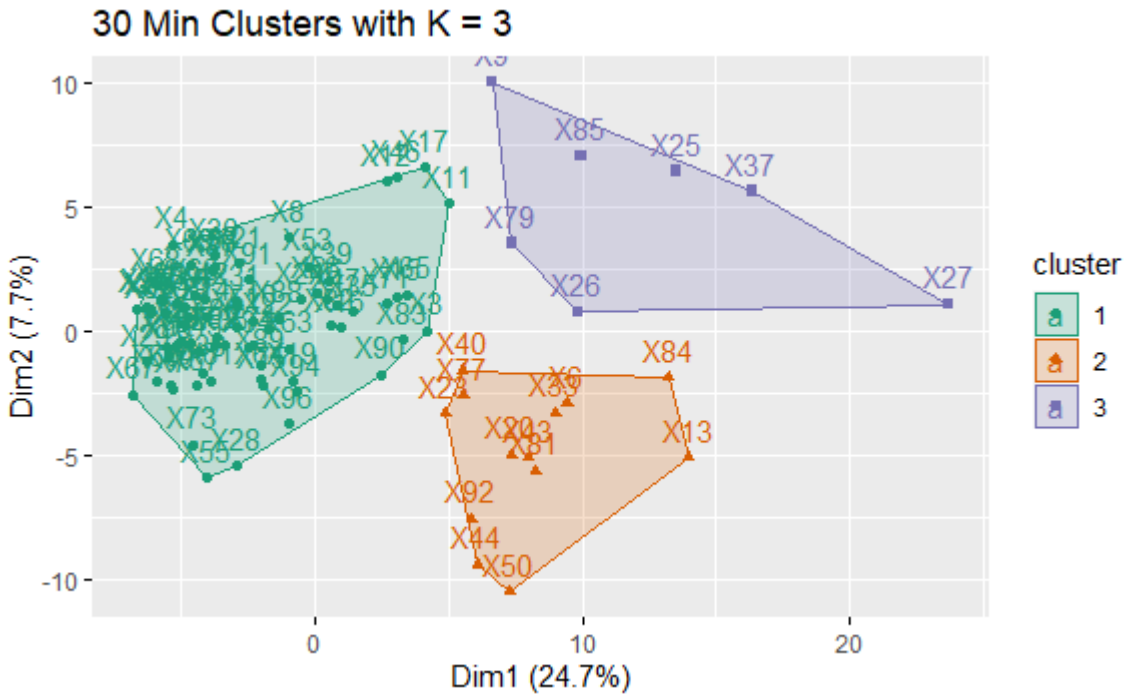
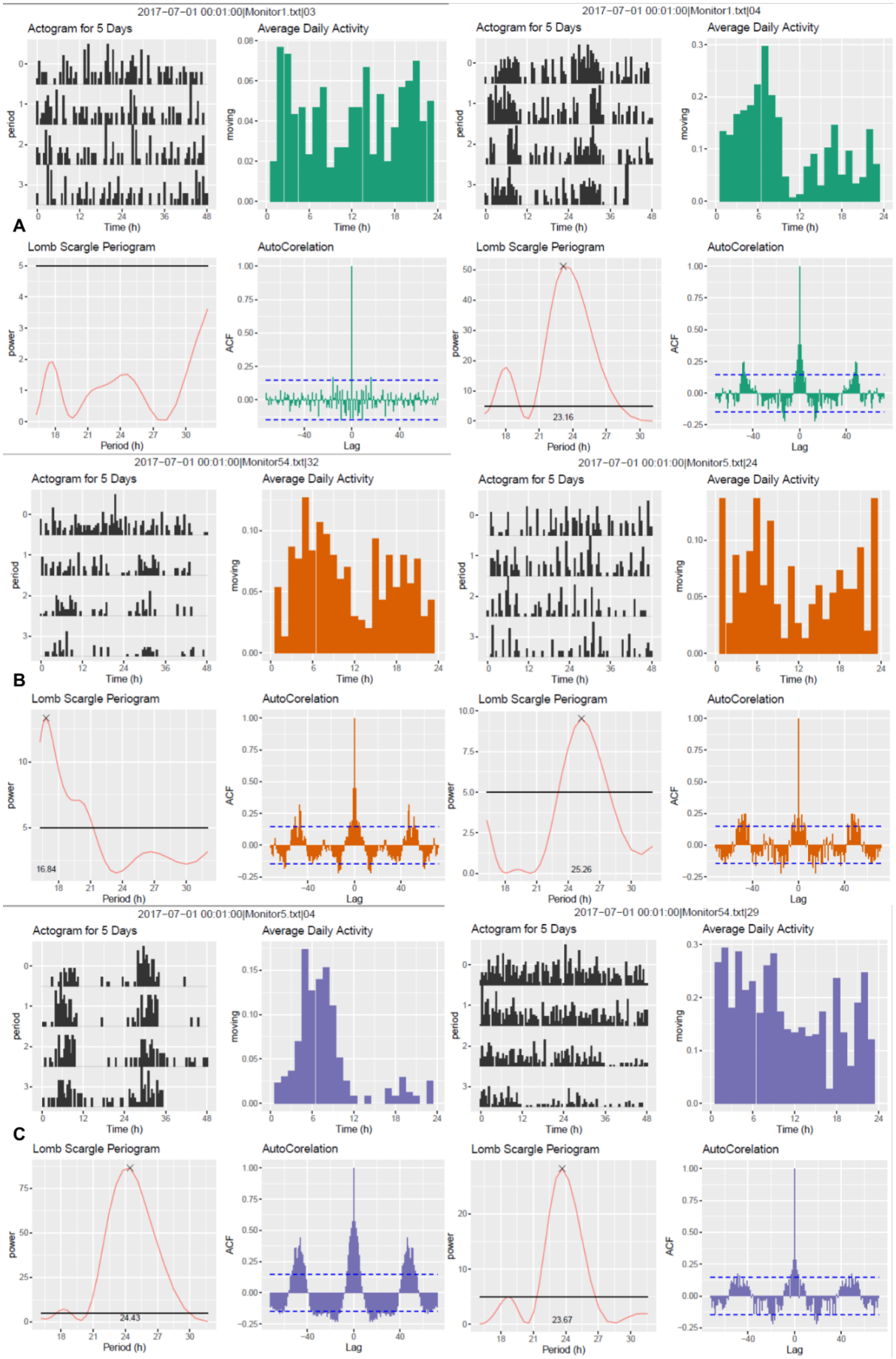


Figure 9: PAM with K =3 for Autocorrelation coefficient clusters of low representation power.


2017-07-01 00:01:00|Monitor1.txt|03
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
2017-07-01 00:01:00|Monitor1.txt|04
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
A
2017-07-01 00:01:00|Monitor54.txt|32
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
2017-07-01 00:01:00|Monitor5.txt|24
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
B
2017-07-01 00:01:00|Monitor5.txt|04
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
2017-07-01 00:01:00|Monitor54.txt|29
Actogram for 5 Days
Average Daily Activity
Lomb Scargle Periogram
AutoCorelation
C
period
moving
power
ACF
Time (h)
Period (h)
Lag
23.16
16.84
25.26
24.43
23.67

1660

Figure 10: PAM with K =3 for Autocorrelation coefficient clusters by thickness of Autocorrelation.

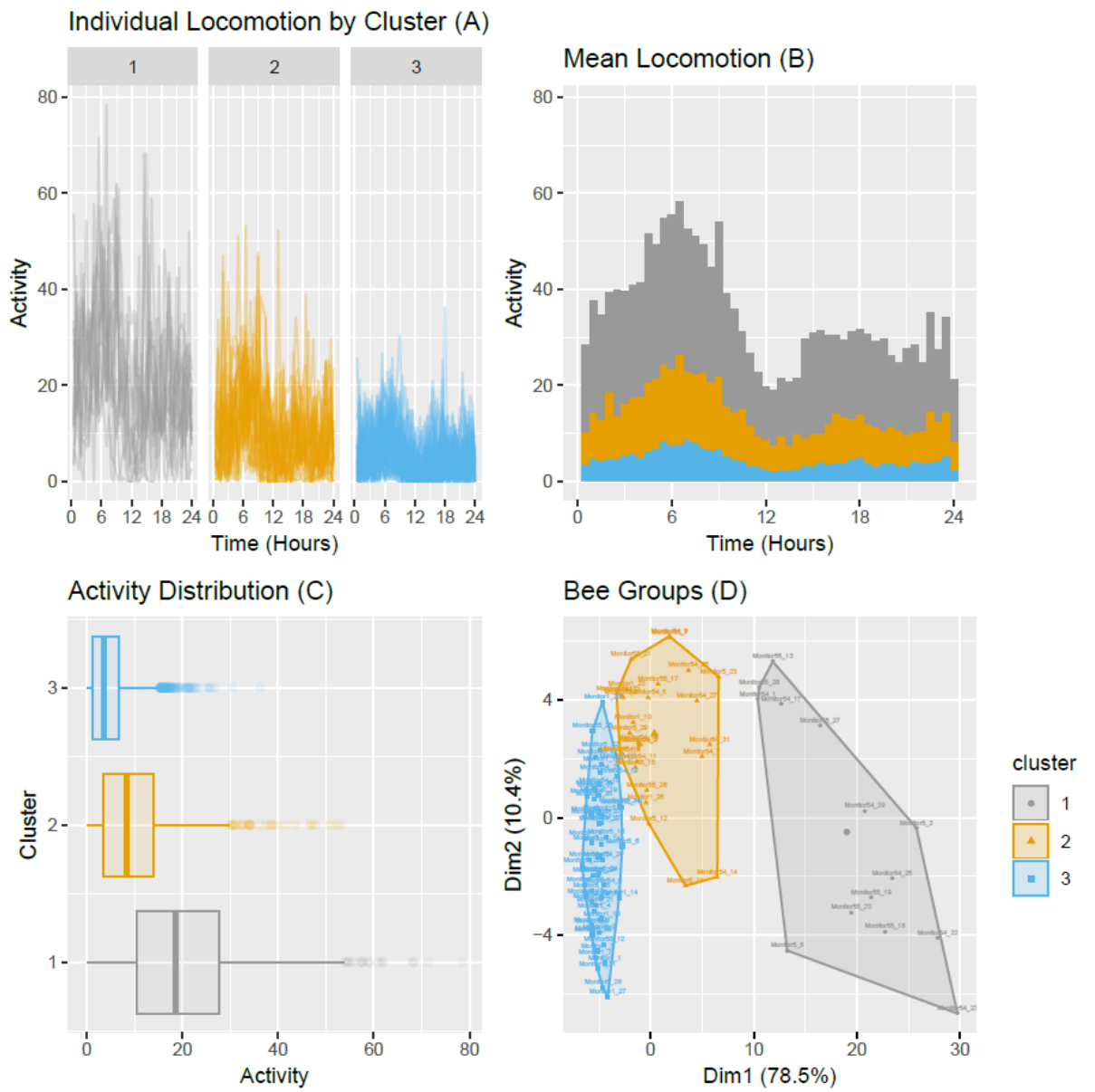


Figure 11: K-means with K= 3 for Average Daily Activity clusters by frequency.

Table 1: Results of KNN by varying PAA Size (PAA) and Alphabet Size (Alphabet). Sliding Window (SW) was continually equal to 48.

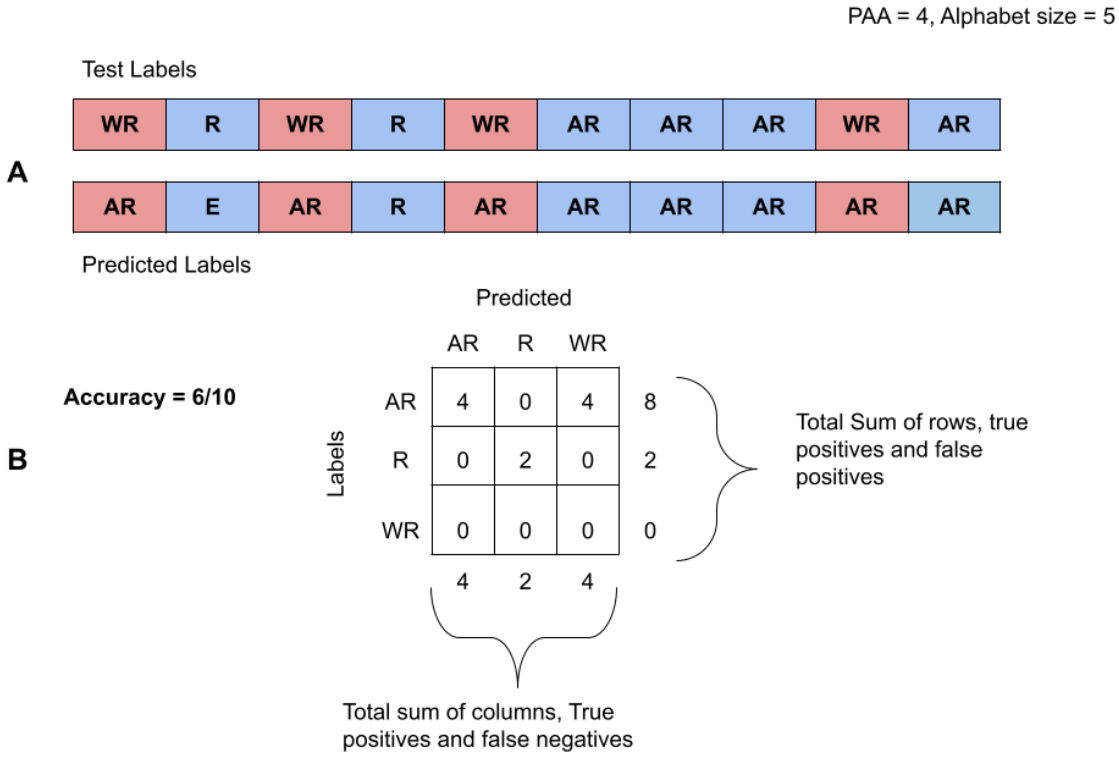
Parameters			Precision			Recall			F1			Accuracy
PAA	Alphabet	SW	R	WR	AR	R	WR	AR	R	WR	AR	
3	3	48	0.5	0	0.25	1	0	0.25	0.67	NA	0.25	0.30
3	4	48	1	0	0.375	0.5	0	0.75	0.5	0.67	NA	0.40
3	5	48	0.5	0	0.4	1	0	0.5	0.67	NA	0.4	0.40
3	6	48	0.5	NA	0.5	1	0	0.75	0.67	NA	0.6	0.50
3	7	48	1	0.67	0.5	0.5	0.5	0.75	0.67	0.57	0.6	0.60
4	3	48	0.67	0.75	1	1	0.75	0.75	0.8	0.75	0.86	0.80
4	4	48	1	0.5	0.6	1	0.25	0.75	0.67	0.4	0.67	0.60
4	5	48	0.67	NA	0.57	1	0	1	0.8	NA	0.73	0.60
4	6	48	1	1	0.57	0.5	0.5	1	0.67	0.67	0.2	0.70
4	7	48	0.3	0	0.5	0.5	0	0.75	0.4	NA	0.6	0.40
6	3	48	0.5	NA	0.5	0.5	0	1	0.5	NA	0.67	0.50
6	4	48	0.33	0.33	0.5	0.5	0.25	0.5	0.4	0.29	0.5	0.40
6	5	48	1	NA	0.5	1	0	1	1	NA	0.67	0.60
6	6	48	NA	NA	0.4	0	0	1	NA	NA	0.57	0.40
6	7	48	0.67	NA	0.57	1	0	1	0.8	0.73	NA	0.60
8	3	48	NA	0.4	NA	0	1	0	NA	0.57	NA	0.40
8	4	48	0.26	NA	0.67	1	0	0.5	0.4	NA	0.57	0.40
8	5	48	0.17	NA	0.75	0.5	0	0.75	0.25	NA	0.75	0.40
8	6	48	NA	NA	0.4	0	0	1	NA	NA	0.57	0.40
8	7	48	0.5	0.5	0.57	0.5	0.25	1	0.5	0.4	0.73	0.60
12	3	48	NA	NA	0.4	0	0	1	NA	NA	0.57	0.40
12	4	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
12	5	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
12	6	48	0.25	0.5	0.5	0.5	0.5	0.25	0.33	0.5	0.33	0.40
12	7	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
24	3	48	0.25	0.5	0.5	0.5	0.5	0.25	0.33	0.5	0.33	0.40

24	4	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
24	5	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
24	6	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20
24	7	48	0.2	NA	NA	1	0	0	0.33	NA	NA	0.20

Table 2: Results Trees by varying PAA Size (PAA) and Alphabet size (Alphabet). Sliding Window (SW) was continually equal to 48.

Parameters			Precision			Recall			F1			Accuracy	Algorithm
PAA	Alphabet	SW	R	WR	AR	R	WR	AR	R	WR	AR		
3	3	48	1	NA	0.5	0.75	0	1	0.8571	NA	0.667	0.667	DT
3	3	48	1	0.5	0.667	1	0.5	0.667	1	0.5	0.667	0.778	RF
4	3	48	1	NA	0.5	0.75	0	1	0.75	NA	0.667	0.667	DT
4	3	48	1	0	0.5	1	0	0.667	1	NA	0.571	0.667	RF
6	3	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	DT
6	3	48	1	1	0.75	1	0.5	1	1	0.667	0.8571	0.889	RF
3	4	48	1	NA	0.4286	0.5	0	1	0.667	0.6	NA	0.556	DT
3	4	48	1	NA	0.5	0.75	0	1	0.8571	0.667	NA	0.667	RF
4	4	48	0.8	NA	0.5	1	0	0.667	0.889	NA	0.5714	0.667	DT
4	4	48	0.8	NA	0.5	1	0	0.667	0.889	NA	0.5714	0.667	RF
6	4	48	0.5	NA	0	1	0	0	0.667	NA	NA	0.444	DT
6	4	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	RF
3	5	48	1	NA	0.4286	0.5	0	1	0.667	NA	0.6	0.556	DT
3	5	48	1	1	0.6	0.75	0.5	1	0.8571	0.667	0.5	0.778	RF
4	5	48	1	1	0.6	0.75	0.5	1	0.8571	0.667	0.75	0.778	DT
4	5	48	1	0.5	1	0.75	1	0.667	0.8571	0.667	0.8	0.778	RF
6	5	48	0.6	NA	0.75	0.75	0	1	0.667	NA	0.8571	0.667	DT
3	6	48	0.667	NA	0.667	1	0	0.667	0.8	NA	0.667	0.667	DT
3	6	48	1	0	0.5	0.75	.	0.667	0.8571	NA	0.5714	0.556	RF

4	6	48	0.8	NA	0.5	1	0	0.667	0.889	NA	0.5714	0.667	DT
4	6	48	0.8	NA	0.5	1	0	0.667	0.889	NA	0.5714	0.778	RF
3	7	48	0.8	NA	0.75	1	0	1	0.889	NA	0.8571	0.778	DT
3	7	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	RF
4	7	48	0.8	NA	0.75	1	0	1	0.889	NA	0.8571	0.778	DT
4	7	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	RF
3	8	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	DT
3	8	48	1	NA	0.6	1	0	1	1	NA	0.75	0.778	RF
4	8	48	0.75	NA	0.4	0.75	0	0.667	0.75	NA	0.5	0.556	DT
4	8	48	1	Na	0.6	1	0	1	1	NA	0.75	0.778	RF



C

	AR	R	WR	Formula
Precision	4/8 = 0.5	2/2 = 1	0/0 = Na	TP/(TP+FP)
Recall	4/4 = 1	2/2 = 1	0/4 = 0	TP/(TP+FN)
F-Score	2(0.5+1)/(0.5+1)	2(1*1)/(1+1)	2(Na*0)/(Na+0)	2(Pre*Re)/(Pre+Re)

Figure 12: Prevalence in NAs is due to poor consistent classification.

Discussion/Conclusion

Discussion/Conclusion

When describing the phenotypic expression of circadian rhythms in our four Halictid bees, we noted that rhythm variability is parallel to that of levels of sociality. From least to most varied in behaviors: *S. curvicornis*; Solitary and with one activity pattern, *L. ferreirii*; Communal with two activity patterns, *L. enatum*: Primitively eusocial with three activity patterns and *L. malachurum*; Facultatively eusocial with five activity patterns. Out of the four species evaluated, *S. curvicornis* is the only specialist, and its relationship to *Campanula arvensis* could explain its rigorous biological clock. However, within the context of this work, that explanation falls short, because all three of the *Lasioglossum* species are generalists, yet they express a significant difference across their circadian parameters, not to mention that *L. malachurum* was caught in the same time place and flower as *S. curvicornis*, yet their patterns for daily locomotion are polar opposites. This is similarly reflected in *L. enatum* and *L. ferreirii* sharing the same exact niche and still displaying a different number of activity patterns. It stands to be seen if these results hold up when considering sexual dimorphism and seasonality. Notwithstanding, our observations set the foundation for asking more complex questions about the influences of sociality in the expression of circadian rhythms in Hymenopterans. It is our belief that these observed differences in rhythm are related to shift work and that understanding circadian rhythms may give a strong insight into the mechanisms that lead to the evolution of complex social organizations. Therefore, we find it worthwhile to look further into this relationship and identify if there is a correlation between the task being done by a bee and their circadian phenotype.

While describing the *Lasioglossum* species, we created a labeling scheme that highly resembles a classification decision tree. Similarly, the original process for exploring the reproducibility of the categories observed in the data set was reminiscent of KNN, where a consensus of users separated the data into discrete pre-established groups. We divided the data this way, because the time series behavior being displayed by the *Lasioglossum* bees was highly heterogeneous. To streamline the *a priori* grouping process, we used clustering analysis as well as classification on the *L. malachurum* dataset. A systematic and replicable methodology for preprocessing locomotor activity data will not only make data analysis faster and easier, but it will also strengthen the reliability of the results.

Clustering proved to be a successful method for grouping individual bees. However, the groups were not of circadian significance, as they were not grouped by rhythm. Nevertheless, the process was pivotal in creating a deeper understanding of the behavior of the data set. Classification, on the other hand, was a complete success. We transformed the data using SAX with the intent of reducing the dimensions of the data set while still keeping its shape. Once the data was transformed, we applied three different classification algorithms, with decision trees being the algorithm that best classified the data, KNN achieving the second-best results, and random forest coming in last place. These findings are reflective of how intuition gained by understanding a data set is the most valuable step in any ML workflow. Using methods that approximate how a user may separate data into discrete functional groups was the key to success.

That being said, some systematic difficulties were encountered when evaluating the efficacy of the classification models. To simplify the classification problem, we did not use all five patterns of activity identified in our grouping scheme. Instead, we used the three larger categories (Rhythmic, Weakly Rhythmic and Arrhythmic) to surmise if that was the minimum amount of labels necessary for an effective classification strategy. For both the rhythmic and arrhythmic categories, the minimum amount of information was enough to properly classify the individuals. The weakly rhythmic category, on the other hand, was consistently misclassified in every experiment. There was no perfect combination of SAX parameters and classification algorithms that would result in the consistent correct classification of the weakly rhythmic individuals. Therefore, the minimum amount of information was not sufficient to inform a proper classification model for the entirety of the *L. malachurum* dataset. Thus, in the future it would be advisable to divide the weakly rhythmic classification into smaller categories for better results, as they are probably obfuscated by the simplification of the data.

The approach that we have developed for preprocessing a circadian data set before evaluation is the first of its kind to our knowledge, and as such, cannot be compared to past studies in the field. Our use of SAX for this data set is unique, since in circadian science, it is more common to use techniques like averages/rolling averages, self/auto correlations and Fourier based transformations (Refinetti et al. 2007). Our results from experimenting with optimal transformation parameters using SAX were not only successful in yielding proper classifications, but also consistent with the findings in other fields, where this transformation worked best with smaller

values for its parameters (Lin and Li 2009). In the future, it would be interesting to replicate these experiments and compare the use of SAX transformations with those of wavelets, as wavelets are a commonly used tool in circadian analysis. Furthermore, there is still the question of transferability. Other species of *Lasioglossum* should be classified using the SAX transformed data with either decision trees or KNN to confirm how generalizable is the use of our pipeline.

This work sets the basis for the use of a novel subject of study and an innovative systematic approach to preprocessing circadian data. We characterized the circadian behaviors of four never before described species of halictid bees. The results from analyzing the bees suggests that circadian behavior may have a complementary relationship with sociality. In addition, we developed and tested a unique preprocessing pipeline utilizing machine learning for the purpose of facilitating the description of organisms for whom their circadian phenotypes are unknown. Future endeavors should focus on testing the transferability of the tool. Furthermore, if we wish to strongly conclude that sociality may serve as a zeitgeber for Halictid bees, we would benefit from replications of our study with larger sample sizes and consideration of sex and seasonality, as well as a larger pool of described species.

Literature Cited

Agostinelli, F., Ceglia, N., Shahbaba, B., Sassone-Corsi, P., & Baldi, P. (2016). What time is it? Deep learning approaches for circadian rhythms. Bioinformatics, 32(12), i8–i17. https://doi.org/10.1093/bioinformatics/btw243

Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Anafi, R. C., Lee, Y., Sato, T. K., Venkataraman, A., Ramanathan, C., Kavakli, I. H., … Hogenesch, J. B. (2014). Machine Learning Helps Identify CHRONO as a Circadian Clock Component. PLoS Biology, 12(4), e1001840. https://doi.org/10.1371/journal.pbio.1001840

Beer, K., & Helfrich-Förster, C. (2020). Post-embryonic Development of the Circadian Clock Seems to Correlate With Social Life Style in Bees. Frontiers in Cell and Developmental Biology, 8, 1325. https://doi.org/10.3389/fcell.2020.581323

Bezdek, J. C., Chuah, S. K., & Leep, D. (1986). Generalized k-Nearest Neighbor rules IEEE IJCNN conference View project Big Data Cluster Analysis View project GENERALIZED k-NEAREST NEIGHBOR RULES*. Fuzzy Sets and Systems, 18, 237–256. https://doi.org/10.1016/0165-0114(86)90004-7

Bhatia, N., & Vandana. (2010). Survey of Nearest Neighbor Techniques. IJCSIS) International Journal of Computer Science and Information Security, 8(2). Retrieved from http://arxiv.org/abs/1007.0085

Bloch, G., & Grozinger, C. M. (2011). Social molecular pathways and the evolution of bee societies. Philosophical Transactions of the Royal Society B: Biological Sciences, Vol. 366. https://doi.org/10.1098/rstb.2010.0346

Bloch, G., Toma, D. P., & Robinson, G. E. (2001). Behavioral rhythmicity, age, division of labor and period expression in the honey bee brain. Journal of Biological Rhythms, 16(5), 444–456. https://doi.org/10.1177/074873001129002123

Breiman, L. (2001) Random Forests. Machine Learning 45, 5–32 . https://doi.org/10.1023/A:1010933404324

Burkov, A. (2019). The hundred-page machine learning book. Canada: Andriy Burkov.Fumo, D. (2017, August 17).

Cordero-Martínez, C.S., et al. (2017). The role of circadian rhythms, humidity and temperature oscillations on the foraging patterns of S. curvicornis and L. malachurum. Poster

Danforth, B. N., Conway, L., & Ji, A. S. (2003). Phylogeny of Eusocial Lasioglossum Reveals Multiple Losses of Eusociality within a Primitively Eusocial Clade of Bees (Hymenoptera: Halictidae). Syst. Biol, 52(1), shrimp. https://doi.org/10.1080/10635150390132687

Danforth, B. N., Eardley, C., Packer, L., Walker, K., Pauly, A., & Randrianambinintsoa, F. J. (2008). Phylogeny of Halictidae with an emphasis on endemic African Halictinae. Apidologie, Vol. 39, pp. 86–101. https://doi.org/10.1051/apido:2008002

Delphia, C. M., & Gibbs, J. (2020). New Island Records for Lasioglossum (Hymenoptera: Halictidae) from the Virgin Islands, West Indies. Journal of the Kansas Entomological Society, 92(2). https://doi.org/10.2317/0022-8567-92.2.479

Dubowy, C., & Sehgal, A. (2017). Circadian rhythms and sleep in Drosophila melanogaster. Genetics, 205(4), 1373–1397. https://doi.org/10.1534/genetics.115.185157

Dubowy, C., & Sehgal, A. (2017). Circadian rhythms and sleep in Drosophila melanogaster. Genetics, 205(4), 1373–1397. https://doi.org/10.1534/genetics.115.185157

Eickwort G.C. (1988). Distribution patterns and biology of West Indian sweat bees (Hymenoptera: Halictidae). In Liebherr J.K. (Ed.), Zoogeography of Caribbean Insects (pp. 232–253). Ithaca: Cornell University Press.

Eickwort G.C., Eickwort J.M, Gordon J., Eickwort M.A. (1996) Solitary behavior in a high-altitude population of the social sweat bee Halictus rubicundus (Hymenoptera: Halictidae), Behav. Ecol. Sociobiol. 38, 227–233.

Field, J. (1996) Patterns of provisioning and iteroparity in a solitary halictine bee, Lasioglossum (Evylaeus) fratellum (Perez), with notes on L. (E.) calceatum (Scop.) and L. (E.) villosulum (K.). Insect. Soc. 43, 167–182

Field, J., Paxton, R.J., Soro, A., Bridge, C. (2010). Cryptic plasticity underlies a major evolutionary transition. Curr. Biol. 20, 2028–2031.

Geissmann Q, Garcia Rodriguez L, Beckwith EJ, Gilestro GF (2019) Rethomics: An R framework to analyse high-throughput behavioural data. PLoS ONE 14(1): e0209331. https://doi.org/10.1371/journal.pone.0209331

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In O'Reilly Media.

Giannoni-Guzmán, M. A., Aleman-Rios, J., Melendez Moreno, A. M., Diaz Hernandez, G., Perez, M., Loubriel, D., … Agosto-Rivera, J. L. (2020). The Role of Temperature on the Development of Circadian Rhythms in Honey Bee Workers Authors. BioRxiv, 2020.08.17.254557. https://doi.org/10.1101/2020.08.17.254557

Giannoni-Guzmán, M. A., Avalos, A., Perez, J. M., Loperena, E. J. O., Kayým, M., Medina, J. A., … Agosto-Rivera, J. L. **(2014)**. Measuring individual locomotor rhythms in honey bees, paper wasps and other similar-sized insects. Journal of Experimental Biology, 217(8), 1307–1315. https://doi.org/10.1242/jeb.096180

Gibbs, J. (2018). Bees of the genus Lasioglossum (Hymenoptera: Halictidae) from Greater Puerto Rico, West Indies. European Journal of Taxonomy, 0(400). Retrieved from https://europeanjournaloftaxonomy.eu/index.php/ejt/article/view/528/1183

Gibbs, J., Brady, S. G., Kanda, K., & Danforth, B. N. (2012). Phylogeny of halictine bees supports a shared origin of eusociality for Halictus and Lasioglossum (Apoidea: Anthophila: Halictidae). Molecular Phylogenetics and Evolution, 65(3). https://doi.org/10.1016/j.ympev.2012.08.013

Goldin, D. Q., & Kanellakis, P. C. (1995). On similarity queries for time-series data: Constraint specification and implementation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 976, 137–153. https://doi.org/10.1007/3-540-60299-2_9

Gonzalez, V. H., Pascual, C., Burrows, S., Çakmak, I., & Barthell, J. F. (2014). Pollen collecting behavior of Systropha planidens Giraud, 1861 (Hymenoptera:

Halictidae) in Turkey. Pan-Pacific Entomologist, 90(4), 226–230. https://doi.org/10.3956/2014-90.4.226

Grozdanić, S. & Z. Vasić. (1968). Observations on the solitary bees *Systropha planidens* Gir., and *Systropha curvicornis* Scop. Bulletin de Museum d'Historie Naturelle de Belgrade 21b:133–138 [In Serbo-Croatian, with German summary].

Hans W. Borchers (2019). pracma: Practical Numerical Math Functions. R package version 2.2.9.https://CRAN.R-project.org/package=pracma

Helm, B., & Visser, M. E. (2010). Heritable circadian period length in a wild bird population. Proceedings of the Royal Society B: Biological Sciences, 277(1698). https://doi.org/10.1098/rspb.2010.0871

Ingram, K. K., Kutowoi, A., Wurm, Y., Shoemaker, D., Meier, R., & Bloch, G. (2012). The Molecular Clockwork of the Fire Ant Solenopsis invicta. PLoS ONE, 7(11), e45715. https://doi.org/10.1371/journal.pone.0045715

Jeanson, R., Clark, R., Holbrook, C., Bertram, S., Fewell, J., & Kukuk, P. (2008). Division of labour and socially induced changes in response thresholds in associations of solitary halictine bees. https://doi.org/10.1016/j.anbehav.2008.04.007ï

Jud, C., Schmutz, I., Hampp, G., Oster, H., & Albrecht, U. (2005). A guideline for analyzing circadian wheel-running behavior in rodents under different lighting conditions. Biological Procedures Online, 7(1), 101–116. https://doi.org/10.1251/bpo109

Kriegel, H.-P., Schubert, E., & Zimek, · Arthur. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? Knowledge and Information Systems, 52, 341–378. https://doi.org/10.1007/s10115-016-1004-2

Leonard Kaufman Peter J. Rousseeuw.(1990). Partitioning Around Medoids (Program PAM).Wiley Series in Probability and Statistics. https://doi.org/10.1002/9780470316801.ch2

Levine, J. D., Funes, P., Dowse, H. B., & Hall, J. C. (2002). Signal analysis of behavioral and molecular cycles. BMC Neuroscience, 3. https://doi.org/10.1186/1471-2202-3-1

Lin, J., & Li, Y. (2009). Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In LNCS (Vol. 5566). Retrieved from https://pdfs.semanticscholar.org/5b5e/e84ac8ab484c33b3b152c9af5a0715217e53.pdf

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

Michener, C. D. (1969). Comparative Social Behavior of Bees. Annual Review of Entomology, 14(1), 299–342. https://doi.org/10.1146/annurev.en.14.010169.001503

Mistlberger, R. E., & Skene, D. J. **(2004)**. Social influences on mammalian circadian rhythms: Animal and human studies. Biological Reviews of the Cambridge Philosophical Society, Vol. 79, pp. 533–556. https://doi.org/10.1017/S1464793103006353

Monti, S. **(2003)**. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning, 52(1/2), 91-118. doi:10.1023/a:1023949509487

Moore, D., Angel, J. E., Cheeseman, I. M., Fahrbach, S. E., & Robinson, G. E. (1998). Timekeeping in the honey bee colony: Integration of circadian rhythms and division of labor. Behavioral Ecology and Sociobiology, 43(3), 147–160. https://doi.org/10.1007/s002650050476

Ortiz-Alvarado, Carlos A,. Ambrose, Alexandria F., Cordero-Martinez, Claudia S., Silva-Echeandia, Sebastian A,. Petanidou, Theodora, Tscheulin Thomas, Gonzalez, Victor H., Giray, Tugrul, Agosto-Rivera, Jose A., Barthell, John F. (in rev) Daily foraging activity and circadian rhythms of a guild of three large carpenter bee species in Chasteberry (Vitex agnus-castus L.) on a Greek island (Lesvos).

Patiny, S., & Michez, D. (2007). New insights on the distribution and floral choices of Systropha Illiger, 1806 in Africa (Hymenoptera, Apoidea), with description of a new species from Sudan. Zootaxa, (1461), 59–68. https://doi.org/10.11646/zootaxa.1461.1.6

Patiny, S., Michez, D., & Danforth, B. N. **(2008)**. Phylogenetic relationships and host-plant evolution within the basal clade of Halictidae (Hymenoptera, Apoidea). Cladistics, 24(3), 255–269. https://doi.org/10.1111/j.1096-0031.2007.00182.x

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., … Niehuis, O. **(2017)**. Evolutionary History of the Hymenoptera. Current Biology, 27(7). https://doi.org/10.1016/j.cub.2017.01.027

Radmacher, S., & Strohm, E. (2010). Factors affecting offspring body size in the solitary bee Osmia bicornis (Hymenoptera, Megachilidae) Factors affecting offspring body size in the solitary bee Osmia bicornis (Hymenoptera, Megachilidae)*. Apidologie, 41, 169–177. https://doi.org/10.1051/apido/2009064ï

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Refinetti, Roberto & Lissen, Germaine & Halberg, Franz. (2007). Procedures for numerical analysis of circadian rhythms. Biological rhythm research. 38. 275-325. 10.1080/09291010600903692.

Richards, M. H. (2000). Evidence for geographic variation in colony social organization in an obligately social sweat bee, Lasioglossum malachurum Kirby (Hymenoptera; Halictidae). Canadian Journal of Zoology, 78(7). https://doi.org/10.1139/z00-064

Richards, M. H., von Wettberg, E. J. & Rutgers, A. C. (2003). A novel social polymorphism in a primitively eusocial bee. Proc. Natl Acad. Sci. USA 100, 7175–7180. doi:10.1073//pnas.1030738100

Roenneberg, T., Daan, S., and Merrow, M. (2003). The art of entrainment. J. Biol. Rhythms 18, 183–194.

Roenneberg, T., Wirz-Justice, A., & Merrow, M. (2003). Life between Clocks: Daily Temporal Patterns of Human Chronotypes. Journal of Biological Rhythms, 18(1). https://doi.org/10.1177/0748730402239679

Rubin, E. B., Shemesh, Y., Cohen, M., Elgavish, S., Robertson, H. M., & Bloch, G. (2006). Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (Apis mellifera) and shed new light on the molecular evolution of the circadian clock. Genome Research, 16(11), 1352–1365. https://doi.org/10.1101/gr.5094806

Schwarz, M. P., Richards, M. H., & Danforth, B. N. (2006). Changing Paradigms in Insect Social Evolution: Insights from Halictine and Allodapine Bees. https://doi.org/10.1146/annurev.ento.51.110104.150950

Schwarz, M. P., Richards, M. H., & Danforth, B. N. (2006). Changing Paradigms in Insect Social Evolution: Insights from Halictine and Allodapine Bees. https://doi.org/10.1146/annurev.ento.51.110104.150950

Steitz, I., Kingwell, C., Paxton, R. J., & Ayasse, M. (2018). Evolution of Caste-Specific Chemical Profiles in Halictid Bees. Journal of Chemical Ecology, 44(9), 827–837. https://doi.org/10.1007/s10886-018-0991-8

Toth, A. L., & Rehan, S. M. (2017). Molecular Evolution of Insect Sociality: An Eco-Evo-Devo Perspective. Annual Review of Entomology, 62(1). https://doi.org/10.1146/annurev-ento-031616-035601

VanderPlas, J. T. **(2017)**. Understanding the lomb-scargle periodogram. ArXiv, Vol. 236, p. 16. https://doi.org/10.3847/1538-4365/aab766

West-Eberhard MJ. (**2003)**. Developmental Plasticity and Evolution. Oxford, UK: Oxford Univ. Press

Westrich, P. (1989). Die Wildbienen Baden-Wu¨rttembergs. Ulmer Verlag, Stuttgart, Germany. 972 pp

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. In Media (Vol. 35).

Wilkerson, D. M, Hayes, N. D. (2010). “ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.” Bioinformatics, 26(12), 1572-1573. http://bioinformatics.oxfordjournals.org/content/26/12/1572.abstract.

Wyman, L. M., & Richards, M. H. (2003). Colony social organization of Lasioglossum malachurum Kirby (Hymenoptera, Halictidae) in southern Greece. Insectes Sociaux, 50(3). https://doi.org/10.1007/s00040-003-0647-7

Yuan Tang, Masaaki Horikoshi, and Wenxuan Li.**(2016)**. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 : 478-489.

Appendix

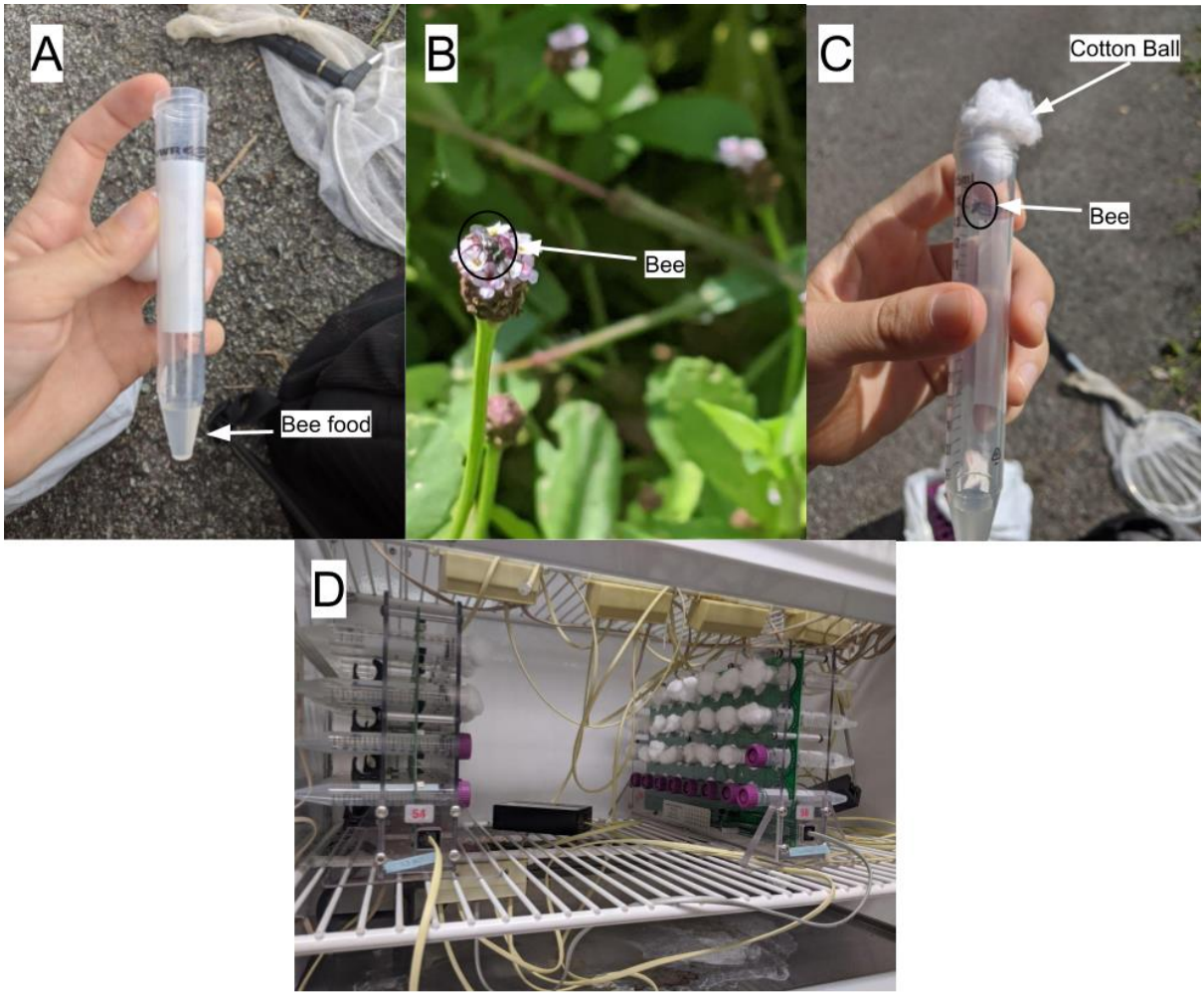


Figure A.1: Methods for capturing and husbandry of the bees. A.1.A) Shows a tube that will house an individual bee, with the agar/sucrose food gel. A.1.B) A bee specimen visiting Phyla nodiflora, where they will be captured as illustrated in A.1.C) with cotton instead of a cap. A.1.D) The bees in their final destination in the incubator placed in their monitors.