Applying Functional Data Clustering for Analyzing Cycles of Periodic Activity of Honeybees

[1]Roberto Trespalacios, [2]Edgar Acuña, [1]Velcy Palomino,
[3]José Agosto, [5]Manuel A. Giannoni-Guzmán, [4]Rémi Mégret

Abstract—In this paper, we analyze the periodic cycle of honeybees when they have between 7 and 9 days of age. The circadian clock of the bees present very erratic behavior that it is a challenge to detect cycles. In signal processing, there are several methods to detect periodic patterns. In here, we will use a well-known test, named periodogram, to evaluate rhythmicity and estimate the period. Besides, to determine whether or no rhythmicity exists, we estimate the time when the bees behavior starts to be rhythmic. Also, it can occur that the bees behavior never gets rhythmic. The test of rhythmicity is applied consecutively until find out periodicity, if this exists. Furthermore, we carry out the periodicity test for the time series obtained from the actogram. We find out that for bees which time series is visually periodic, our method detects correctly the starting time. However, for bees which time series does not show a cyclic pattern our method fails due to a very erratic time series and that the consecutive test results also will show this erratic behavior. Finally, we classify the bees according to theirs beginning of a periodic cycle, using functional data analysis.

Keywords—Honeybees circadian cycle, functional data analysis, hierarchical clustering, periodogram.

I. Introduction

THE circadian clock of honey bees is important in complex physiological processes, such as spatiotemporal learning, time perception and sun-compass navigation. The study of honey bee circadian rhythms is of particular interest because similar to human infants, young honey bees present postembryonic development of circadian rhythms before they forage. Also, it seems that workers will remain arrhythmic performing in-hive tasks and will develop circadian rhythmicity just prior to the onset of foraging behavior. The effect of temperature in the development of circadian rhythms in honey bee workers has yet to be explored. It seems that temperature at the center of the colony is important for the development of circadian rhythms in young honey bee workers. Thus, [1] isolated 1-day-old workers in locomotor activity monitors either at $25°C$ or $35°C$. Previous study indicates that the first 48 hours after emergence are important for the development of circadian rhythms, we examined the effect of colony temperature during these 48 hours by placing some individuals at $35°C$ and then changing the temperature to $25°C$, The remaining individuas are kept at $35°C$. In this paper, we analyze the circadian cycle of honeybees when they have between 7 and 9 days of age. There are several methods to detect periodic patterns, we will use a well-known test, named periodogram, to evaluate rhythmicity and estimate the period. Besides, to determine whether or no rhythmicity exists, we estimate the time when the bee's behavior starts to be rhythmic. Also, it can occurs that the bee's behavior never gets rhythmic. In order to do that, we perform consecutive test of rhythmicity until find out periodicity, if this exists. Finally, we classify the bees according to the beginning of their circadian cycle, using functional data analysis. The first section of this paper deals with data preprocessing, in the second section we explain in detail the methodology used. The last section is about the conclusions obtained from our research work.

[1]R. Trespalacios and V. Palomino are PhD students of Computing and Information Science and Engineering Department, University of Puerto Rico at Mayagüez.

[2]Dr. E. Acuña is profesor of Mathematical Sciences Department, University of Puerto Rico at Mayagüez.

[3]Dr. J. Agosto is profesor of Department of Biology, University of Puerto Rico at Rio Piedras.

[4]Dr. R. Mégret is profesor of Department of Computer Science, University of Puerto Rico at Rio Piedras.

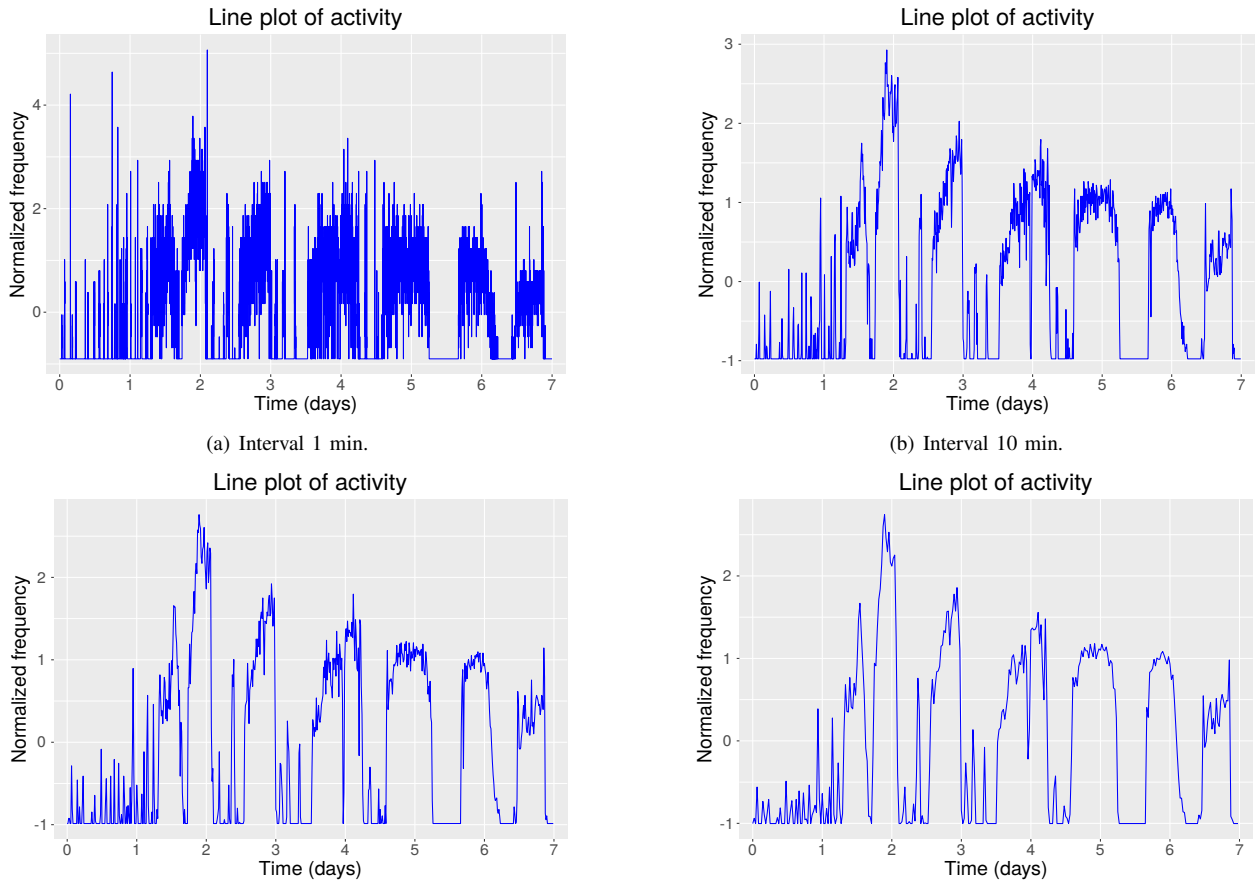
[5]Dr. M. Giannoni-Guzmán is profesor of Department of Biological Sciences, University Vanderbilt Nashville, TN, USA.

Manuscript received February 10, 2018; revised March 20, 2018.

II. Data Preprocessing

The data used in this paper comes from two experimental groups, each of them with four monitors. There are 32 bees in each monitor. The first group, including monitors 53, 54, 55 and 56, was set to a temperature of $35°C$ for two days (48 hours), the next 7 days (168 hours) the temperature was decreased to $25°C$. The second group, including monitors 23, 24, 40 and 41, was maintained to a constant temperature of $35°C$ during 7 days (168 hours). The activity of each bee was taken in intervals of one minute and it was considered as a time series. These measurements do not allow to observe clearly trends o possible cycles, since the period can be hidden when there are a large amount of data points [2]. For this reason, we grouped the data in a larger time period and averaged them in each interval. In this way, short term variations are smoothed and trends and large terms cycles are emphasized. Also, we only use the original data from the time series up to zero values starting to appear, since the average is sensible to a large amount of zeros appearing at the end of some time series and they can affect greatly the average of each interval.

Looking at the Fig. 1, we notice that the grouping in 15-minutes intervals is the best one to show trends in the behavior of the bees (cycles) and the loss of information is minimized (Fig. 1(c)). Grouping in 30 minutes time intervals had been suggested in [1]. In Fig. 2, we show the actogram and the corresponding time series for the activity of two bees; number 8 and 31 from monitors 41 and 24 respectively.



(a) Interval 1 min.

(b) Interval 10 min.

(c) Interval 15 min.

(d) Interval 30 min.

Fig. 1. Time series plot of the average activity for the bee number 8, from monitor 41, using different time intervals (1, 10, 15 and 30 minutes), at 35°C of temperature.

III. Methodology

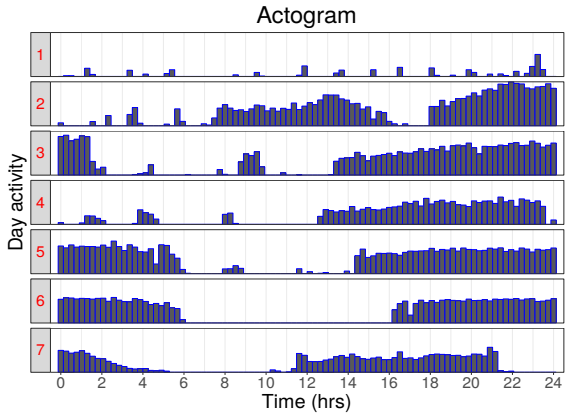
A. Analysis of p-values series

The main idea in our method is to compare time intervals recursively, by increasing the number of period of the series and simultaneously find out the periodograms using the Lomb-Scargle method implemented in the `R` package `Lomb` [3]. After that the *p-values* for the highest peak in the periodogram, which are computed from the exponential distribution, are compared to decide starting in which time intervals there exists a potential cycle. In the sequel, our procedure is described in detail for a single bee.

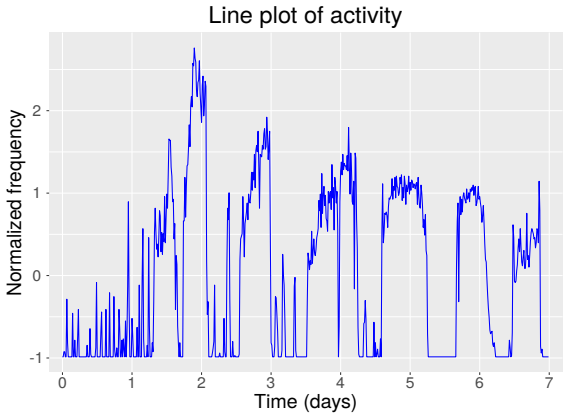
Let $X_t = (x_1, x_2, \ldots, x_n)$ be the activity series of a bee at intervals of 15 minutes. The initial 6-hours interval of the series X_t, is divided in $X_{1 \leqslant t \leqslant k \times 24} = (x_1, x_2, \ldots, x_{k \times 24})$ with $k = 1$, to determine if there is periodicity. Then, we add a new group of observations (increment of 6 hours), and the periodicity is analyzed again; Thus, successively for $k = 2, 3, \ldots$, up to $k = 28$, the last increment of 6 hours (there are 28×24 observations each at 15 minutes). The time were the series of *p-values* is stable is represented by t_{stable}, and it corresponds to the first time t, where the *p-value* is less than the last non-significant *p-value*. That is,

$$t_{stable} = \max_{1 \leqslant t \leqslant n} \{t \mid p_{t-1} > p_t \ \ and \ \ p_t < 0.05\} \qquad (1)$$

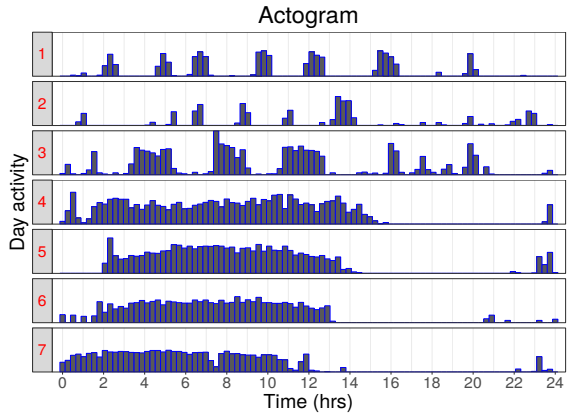
The sub-sequence $x_1, x_2, \ldots, x_{t_{stable}}$ is the initial part of original series $x_1, x_2, \ldots, x_n$, and $x_{t_{stable}}$, is the point where stability was found in the series of the corresponding in *p-values*. In the neighborhoods of time t_{stable} (given by the equation (1)), it follows that the bee starts its periodic cycle. In Fig. 3(a), the time series of *p-values* and its corresponding spline for the bee number 8 from the monitor 41, is shown. It can be noticed that the series of *p-values* stabilizes in the time $t_{stable} = 6$. From this is inferred that bee started her periodic cycle at approximately 36 hours (1.5 days). In Fig. 3(b), we can see the series of *p-values* of bee 31 from monitor 24. This series starts at values greater than 0.05, then descends and after that there is local stabilization of the series; but again the values of the series increase, and descend again, to finally stabilize in time $t_{stable} = 10$. This means that bee 31 begins its periodic cycle at approximately 60 hours (2.5 days). Later, we will describe a method to calculate the start of periodic cycles in bee clusters, using the instantaneous speed at the t_{stable} projection point on the curve described by the *p-values* series.



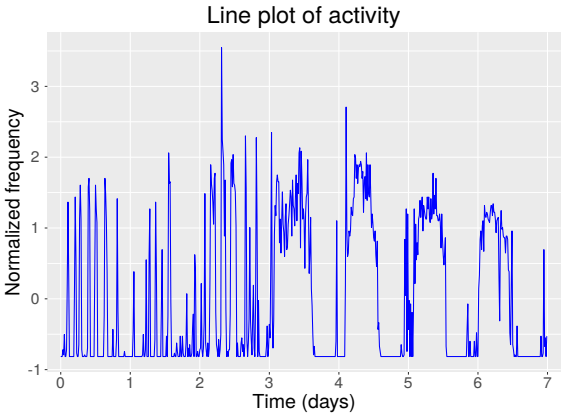
(a) Bee 8, monitor 41.



(b) Bee 8, monitor 41.



(c) Bee 31, monitor 24.



(d) Bee 31, monitor 24.

Fig. 2. Actograms and time series plot for bee 8(plots a and b) from monitor 41 and bee 31(plots c and d) from monitor 24, using time intervals of 15 minutes at 35°C of temperature.

B. Periodogram analysis

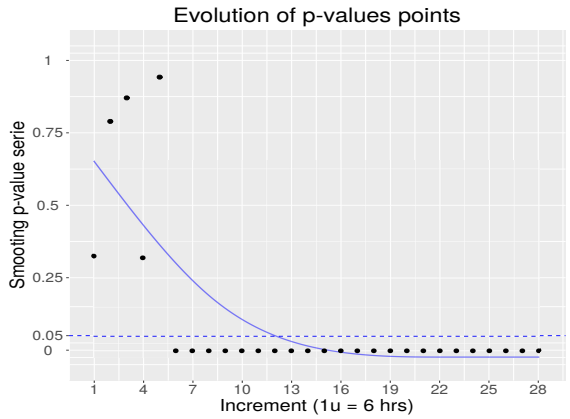
We then perform an analysis of the periods of bees 8 and 31 of monitors 41 and 24 respectively, in their stability time in order to find their period of activity (periodic cycle). In the Fig. 4(a), we show the graph of the periodogram for the bee 8 of monitor 41. In this graph, we find that the periodicity detection is significant at peak value 80.41, which occurs at the sub-sequence $x_{t_{stable}}, x_{t_{stable+1}}, \ldots, x_{n-1}, x_n$ this is the final part of original series $x_1, x_2, \ldots, x_n$. The highest significant value of the periodogram (80.41), corresponds to periodic cycle of 26.4 hours ≈ 1.1 days. This means, it takes a $26.4 \times 15 = 105.6$ time periods (each one of 15 minutes) for a complete cycle. In the same way, in the Fig. 4(b), we have the periodogram for time $t_{stable} = 10$; time at which the detection of periodicity is significant. Carrying out an analysis similar to the previous one, we have that the period for the bee 31 of monitor 24 is 21.6 hours ≈ 0.9 days.

C. Functional splines curves of p-values

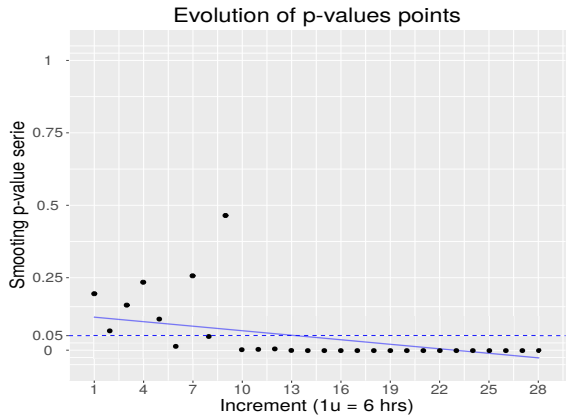
It is clear that to make increases on the series, and assuming that each bee stabilizes its rhythm of activity as time passes, we can expect the convergence of the series of *p-values*. According to this, the interpolation problem of pvalues points described by the *p-values* of the periodogram of each subseries, is guaranteed by the cubic interpolation; but it is not enough to obtain a smooth monotone decreasing curve; therefore, monotonous cubic spline interpolation [4] is used, and in this way ensure that the obtained curve fits the phenomenon described in the problem. The sequence of 28 *p-values*, corresponds to the periods of increments in the series of X_t, modeled by monotonic cubic splines with 5 nodes. Each curve ϕ_i, shows the stability of the activity rhythm of the bee $i = 1, 2, \ldots, 128$, in all monitors through time. This is for both temperatures i.e, for all 256 bees of the 8 monitors (4 monitors at 25°C and 4 at 35°C). Fig. 5 shows all 32 bees of monitor 41 with their respective spline curves for the *p-values* series.

D. Periodicity start time

Each curve, in Fig. 5, were modeled using monotonic cubic splines with 5 nodes using the series of *p-values*. Note that curves with a greater slope (speed) at each time (as bee 8 of monitor 41 in Fig. 3(a)), characterize to bees that start their periodic cycle earlier than bees whose slope of the curve is smaller; as is the case with the bee 31 of monitor 24 (see Fig. 3(b)). Let's see the idea in more detail. Be the curve ϕ_i, which models the series $p_t(i) = (p_1, p_2, \ldots, p_k)$ of *p-values* of the bee i; then,

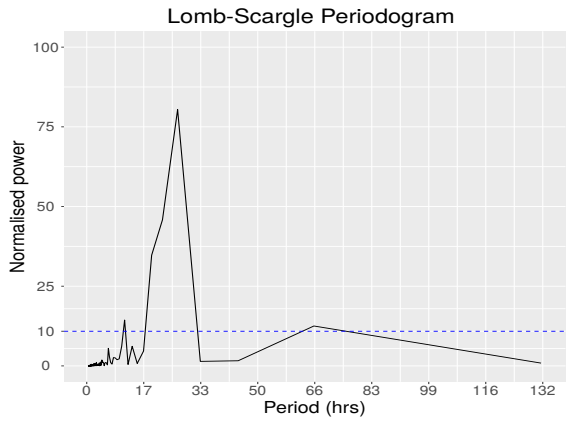


(a) Evolution of *p-values* series and $t_{stable} = 6$, increase of 6 hours (bee 8, monitor 41).

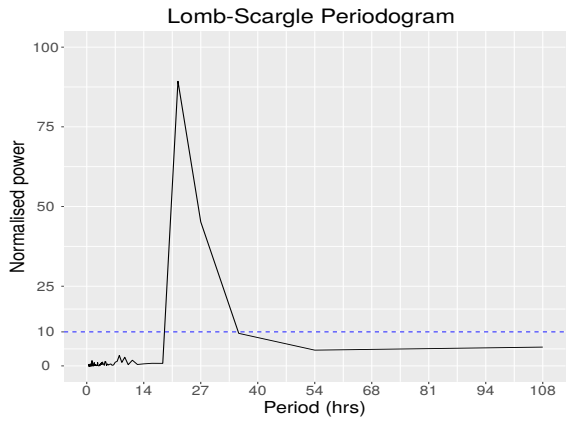


(b) Evolution of *p-values* series and $t_{stable} = 10$, increase of 6 hours (bee 31, monitor 24).

Fig. 3. Plot of *p-values* series and its spline modeling for the bee 8 of monitor 41 and the bee 31 of monitor 24; at 15 minute intervals and at a temperature of $35°C$.



(a) Periodogram of $t_{stable} = 6$, from hour 36 to hour 168 for the series of activities of the bee 8, monitor 41 (period 26.4 hours).



(b) Periodogram of $t_{stable} = 10$, from hour 90 to hour 168 for the series of activities of the bee 31, monitor 24 (period 21.6 hours).

Fig. 4. Plots of periodograms for bees 8 of monitor 41 and bee 31 of monitor 24; at 15 minute intervals and at a temperature of $35°C$.

$$\left| \frac{d\phi_i(t)}{dt} \Big|_{t=t_{stable}} \right| = speed(t_{stable}) \tag{2}$$

This means that the slope (speed) of the curve that models *p-values*, gives us a characterization of the bees, with respect to the time in which they could start their periodic cycle. Since the threshold (distance in the probability interval $[0,1]$), that these should reach is a *p-value*< 0.05. In other words, using the equation (2), which is the slope (speed), in the times t_{stable} for each curve and assuming they all (if there are outliers, they will be excluded from the analysis) reach the threshold, we can find the estimated time of the exact value at which they start their periodic cycle a group of bees.

E. Functional data clustering for smooting p-values series

With the analysis done so far we conclude that the organization of bees by the beginning of the periodic cycle, is a problem of classification of curves. Therefore, it is more efficient, robust and accurate, to use an appropriate classification method of these curves. All this would be within the scope of functional data clustering analysis [5].

Functional data clustering is used when data are curves and the aim is grouping this curves. This grouping allow to identify patterns in data. By exploring the observations in each cluster, we may be able to describe common characteristics.

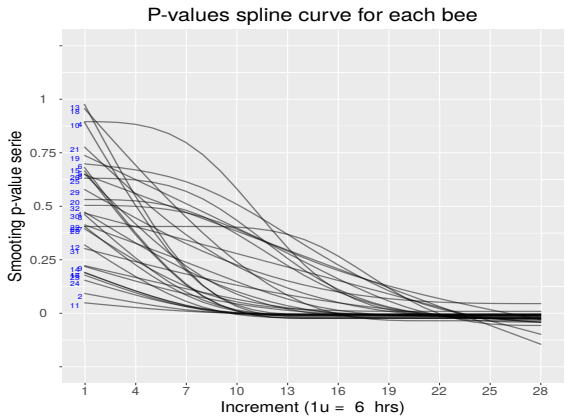
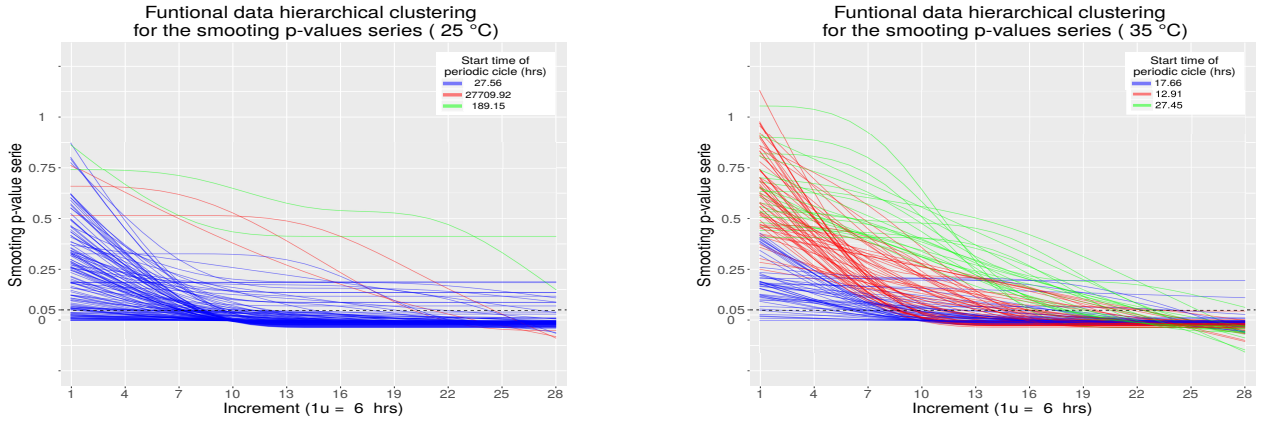


Fig. 5. Functional monotonic cubic spline curves for the *p-values* series of all 32 bees, monitor 41 at $35°C$.



(a) Spline curves for all bees of the monitors at 25°C.

(b) Spline curves for all bees of the monitors at 35°C.

Fig. 6. Graph of natural cubic spline curves and their respective funtional data clustering for the smooting *p-values* series of the bees at intervals of 15 minutes.

These characteristics may allowing us establish convenient interpretations [6]. In our study patterns enable us to determine the start cycle of bees. In the practice, functional data are discrete measurement points for each individual. Naturally, each of these discrete measurement functional would be modeled by a continuous function using some smoothing method. In sub-section III-D, we have the *p-values* series for each bee, which were smoothing using natural cubic spline.

Classification of curves (bees), is carried out using Ward's hierarchical agglomerative clustering method [7]. For this purpose, a distance is used, which will give support for the minimum variance (distance) criterion. Next, we define the distance, by the equation (4) proposed by Ferreira [8] as dissimilarity measure between two curves. Ferreira estimates the distance between two curves using a trapezoidal-rule approximation equation (3), it is:

$$I_k = \frac{t_k - t_0}{2k}[h(t_0) + 2h(t_1) + \cdots + 2h(t_{k-1}) + h(t_k)], \quad (3)$$

where k is the number of measurement points used in the approximation in the interval $[t_0, t_k]$. The distance is defined by:

$$h(t) = [y_i(t) - y_j(t)]^2, \quad (4)$$

here, $y_i(t)$ is the *i-th* smoothing *p-value* in the time t.

Next, in Fig. 6, we can see in detail the bees classification into three clusters at 25°C and 35°C. This classification shows the start time of periodicity for each cluster of curves. In each cluster of curves we take their respective projection of t_{stable} on curve and calcule the average of all speeds and them using their to find the start time stimated. The classification of the bees at 25°C, show that for each group the time of beginning of the periodic cycles, they are later to the groups of the classification of the bees to 35°C. The group of bees at 25°C, show that for each group the time of beginning of the periodic cycles, they are later to the group of the classification of bees to 35°C. On the other hand, we observe in Table I, that there are individuals (bees) whose estimated curves are autlier.

TABLE I
START TIME OF PERIODICITY BY CLUSTER

	Temperature					
	25°C			35°C		
Class	1	2	3	1	2	3
Periodicity start time (hrs)	27.58	27708.92	189.15	17.66	12.91	27.45
Number of outliers	5			3		

Finally, we can see, the classification of those starting their cycle period earlier. In general, bees that were subjected to a temperature of 35°C, began their periodic cycles in a shorter time than those bees subjected to 25°C. It should be said that the times of the groups of bees at 35°C are homogeneous and their range is 14.54 hours; while for bees at 25°C, the range is 161.57 hours.

It should be taken into account, that these times were calculated exactly at the initial points where the series of *p-values* was estimated to be stable. It is possible to give a greater tolerance, and to take the stability times that are beyond (one, two or even three increments ahead) to the time of stability t_{stable}. On the other hand, it would also be a good strategy to vary the size of each increment; as well as the amplitude of the intervals with which the data is filtered.

Note: All the algorithms of this work were made using the statistical software R. For more details, you can see: [9] and [10].

IV. Conclusions

In this work, after an exploration of different methods and strategies for the analysis and classification of the period cycle of the bees, the following conclusions were obtained:

i. The Analysis of the period cycle of bees is a complex problem due to the particularity and uncertainty of activity and behavior in each bee.

ii. Measurements of the data should not be very close, because the time series may lose its tendency when there are too many samples in a very short period of time.
iii. It was observed that the curves modeled by monotone cubic splines and using hierarchical agglomerative clustering, are a powerful tool to determine the different groups the curves. In addition, these same curves serve as a basis for calculating the beginning of periodic cycles, by means of the projection of the t_{stable} times on each curve.
iv. The use of periodograms is beneficial; however, it has its constrains. For this reason, the consecutive analysis of periodograms until reaching the periodicity of the series, can lead to erroneous conclusions, To minimize them it is important to refine the calculations and method of speeds and then compare the results.

As a future work, we will try to apply other filters or smoothing technique to detect periodicity in these series, use enhanced statistical methods of rhythm detection for genome data[11], and to improve the proposed method and compare its results with another existent methods.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1707355 and 1633184. BIGDATA:Collaborative Research: Large-scale multi-parameter analysis of honeybee behavior in their natural habitat.

REFERENCES

[1] M. Giannoni-Guzmán, *Individual Differences in Circadian and Behavioral Rhythms of Honey Bee Workers (Apis Mellifera L.).* PhD thesis, 2016.
[2] H. B. Dowse, *Statistical Analysis of Biological Rhythm Data*, pp. 29–45. Totowa, NJ: Humana Press, 2007.
[3] T. Ruf, "The lomb-scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series," *Biological Rhythm Research*, vol. 30, no. 2, pp. 178–201, 1999.
[4] G. Wolberg and I. Alfy, "Monotonic cubic spline interpolation," in *cgi*, p. 188, IEEE, 1999.
[5] J. O. Ramsay, *Functional data analysis*. Wiley Online Library, 2006.
[6] J. Jacques and C. Preda, "Functional data clustering: a survey," *Advances in Data Analysis and Classification*, vol. 8, no. 3, pp. 231–255, 2014.
[7] F. Murtagh and P. Legendre, "Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm," *arXiv preprint arXiv:1111.6285*, 2011.
[8] L. Ferreira and D. B. Hitchcock, "A comparison of hierarchical methods for clustering functional data.," *Communications in Statistics - Simulation and Computation*, vol. 38, no. 9, pp. 1925–1949, 2009.
[9] N. Matloff, *The art of R programming: A tour of statistical software design*. No Starch Press, 2011.
[10] T. Mailund, *Functional Data Structures in R: Advanced Statistical Programming in R*. Apress, 2017.
[11] A. L. Hutchison, M. Maienschein-Cline, A. H. Chiang, S. A. Tabei, H. Gudjonson, N. Bahroos, R. Allada, and A. R. Dinner, "Improved statistical methods enable greater sensitivity in rhythm detection for genome-wide data," *PLoS Comput Biol*, vol. 11, no. 3, p. e1004094, 2015.

Roberto Trespalacios received his B.S degree in Mathematics from Universidad de Cartagena, Colombia in 2005, and a M.S. In Mathematical Statistics from the University of Puerto Rico at Mayaguez Campus (UPRM) in 2013. Currently, he is a PhD student in Computing and Information Sciences and Engineering at the UPRM. His research interests include computational statistics, Functional data Analysis and Data Mining.

Edgar Acuna received his PhD in Statisics n 1989 from University of Rochester, New York, USA. Currently he is Professor with the Department of Mathematical Sciences at the University of Puerto Rico, Mayaguez Campus, His research interests are in data preprocessing for Big Data, functional data analysis, and scalale machine learning

Velcy Palomino received her BS degree in Mathematics from the National University of Cusco, Peru in 2005 and a M.S. In Mathematics and Statistics from the University of Puerto Rico at Mayaguez Campus (UPRM) in 2012. Currently, she is a PhD student in Computing and Information Sciences and Engineering at the UPRM. Her research interests are in Data Mining, Data Streams, Big Data

Jose Agosto received his PhD in Biology in 2008 from Brandeis University, USA. Currently he is Professor with the Department of Biology at the Universioty of Puerto Rico, Rio Piedras Campus. His research interest are Molecular genetics, Circadium Rithms ans Neurosciences.

Remi Megret received his PhD degree in Computer Science from INSA Lyon, France, in 2003. In 2004, he joined the ENSEIRB-MATMECA at Bordeaux Institute of Technology and the Signal and Image Processing Research Group at IMS Laboratory. Currently, he is Professor the Computer Science Department of the University of Puerto Rico at Rio Piedras . His research of interest are within the fields of Computer Science, Signal and Image Processing and Mathematical modeling.

Manuel Giannoni-Guzman received his PHD degree in Biology from the University of Puerto Rico, Rio Piedras Campus, in 2015. Currently, he ia a Postodoctoral Scholar at the Department of Biological Sciences, Vanderbilt University. His reserach interest is in Neuroscience.