An Autonomic, Adaptive and High-Precision Statistical Model to Determine Bee Colonies Well-Being Scenarios

Daniel de Amaral da Silva [1,2], Ícaro de Lima Rodrigues [2], Antonio Rafael Braga [2,3], Juvêncio S. Nobre [1], Breno M. Freitas [4], Danielo G. Gomes [2]

[1]Departamento de Estatística e Matemática Aplicada (DEMA), Centro de Ciências, Universidade Federal do Ceará, Fortaleza - CE, CEP 60.440-900, Brasil.

[2]GREat, Departamento de Engenharia de Teleinformática (DETI), Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza - CE, CEP 60.455-970, Brasil.

[3]Sistemas de Informação, Campus Quixadá, Universidade Federal do Ceará, Quixadá - CE, CEP 63.902-580, Brasil.

[4]Setor de Abelhas, Departamento de Zootecnia, Centro de Ciências Agrárias, Universidade Federal do Ceará, Fortaleza - CE, CEP 60.356-000, Brasil.

{danielamaral, icarodelima}@alu.ufc.br

{rafaelbraga, juvencio, freitas, danielo}@ufc.br

Abstract. *Honey bees, important pollinators, are threatened by a variety of pests, pathogens and extreme climatic events, such as the winter period. This paper proposes a two-stages model that seeks to define and predict evolutionary scenarios for improving the bee colonies' well-being. The used dataset has data from both internal and external beehive sensors, and on-site inspection of beekeepers from six apiaries between the years 2016-2018. In the first stage, three evolutionary scenarios were obtained (pessimistic, conservative and optimistic) through the clustering technique. In the second one, aiming to classify these scenarios, an elastic net penalty logistic regression model was obtained with an accuracy of* $\approx 99.5\%$.

1. Introduction

Honey bees (Apis mellifera) are responsible for pollinating about 15$ billion worth of food only in the US every year [Braga et al. 2020]. However about 40% of the world's bee species are dying [Sánchez-Bayo and Wyckhuys 2019]. According to *Bee Informed Partnership* (BIP) [1], the total annual loss of bee hives in the 2018/19 season was above average, at almost 40%, the biggest loss in 13 years. In particular, honeybees populations have suffered mass deaths in some European regions and in North America due to Colony Collapse Disorder (CCD) and severe winters [Barron 2015, Gil-Lebrero et al. 2017].

Today, thanks to the sensor networks and Internet of Things (IoT) paradigms, beekeepers and researchers can remotely monitor bee colonies [Meikle and Holst 2015, Kridi et al. 2016, Zogovic et al. 2017]. Remote monitoring via wireless sensors is one of the most important characteristics of the precision beekeeping [Zacepins et al. 2015] which basically involves beehives data collection, data analysis and support decision making in an apiary management context

[1]https://research.beeinformed.org

[Dineva and Atanasova 2018, Braga et al. 2019]. Once the sensors are installed in the hives, the apiary can be monitored without disturbance, even during periods when invasive inspections of the hives are contraindicated, such as during the winter [Meikle et al. 2017]. However, little is yet known about the semantics of the data collected from the hives [Zacepins et al. 2015, Jacobs et al. 2017], such as which physical variables most affect the bees behavior. Such knowledge would help to improve for instance the bee colonies' well-being.

In this context, this paper presents a solution for the definition and classification of bee colonies' well-being states. Some of these states could lead the hive to an irreversible imbalance. Our proposal uses clustering (unsupervised learning), with internal, external sensors and inspection data, and classification (supervised learning), with internal, external sensor and clustering data. By not including inspection data during the classification step, we reinforce the idea of not needing invasive inspections in the hives to know the current status of the colonies.

2. Material and Methods

2.1. Dataset

We used a dataset composed of internal data from 27 colonies, distributed in 6 apiaries in the USA as well meteorological data[2] from the apiaries monitored from January 2016 to December 2018. Raw dataset has 731,654 observations. Table 1 shows the data distribution per hives/apiaries and monitoring periods.

Table 1. Distribution of data between beehives and apiaries in the United States.

State	Apiary	Monitoring Period	#Samples
North Carolina	BBCC	06/2016 - 10/2018	200,544
	Juniper Level	04/2017 - 12/2018	90,613
	Beesboro	01/2018 - 12/2018	132,552
Pennsylvania	BBTS	01/2016 - 12/2016	27,673
Indiana	The Bee Hive	01/2016 - 12/2017	160,381
Utah	Lakeview	02/2017 - 12/2018	118,891

The collection of the internal data was carried out through the *SolutionBee*[3] system, using as internal variables the temperature of the bees cluster (i.e. the central part of the hive, where the bees gather) and the hive weight.

The used meteorological data were sampled each 1 hour (sampling period) and the weather station was chosed based on the shortest distance between the station and the apiary. In short, the internal sensor data variables were: temperature ($^{\circ}C$) and weight (kg) whereas those from external sensors were temperature ($^{\circ}C$), humidity (%), pressure (hPa), dew point ($^{\circ}C$), precipitation (mm) and wind speed (km/h).

Inspection data *in loco* were obtained using a standardized inspection form in a weekly basis. The form here used, called Healthy Colony Checklist (HCC), was

[2] www.worldweatheronline.com
[3] www.solutionbee.com

proposed by the Bayer Bee Care Center (BBCC) [4]. The HCC consists of 6 binary levels (yes/no questions) to define the colony health: 1 - brood - all stages of brood and instars (i.e. the larvae growth stages) present in appropriate amounts? 2 - adult bees (Sufficient adult bees and age structure to care for brood and perform all tasks of the colony?), 3 - queen (a young (<1 year old), productive, laying queen present?), 4 - food (Sufficient nutritious water, forage, and food stores available (inside and/or outside the hive), and young brood being fed?), 5 - no stressors (no (apparent) stressors present that would lead to reduced colony survival and/or growth potential?; and 6 - suitable space (suitable space for current near-term expected colony size that is sanitary, defendable, and room for egg laying?). Thus, if all these 6 questions are marked "no problem", then the colony is considered 100 % healthy. On the other hand, each item marked as 'with problem', represents a theoretically decrease of 1/6 of the colony health level.

2.2. Pre-processing

The raw data set, described in the subsection 2.1, has been merged with the inspection data from the HCC form, with a 7-day limit of the module difference between the collection dates of the inspection data sets and sensors. Observations from sensors that did not correspond to any inspection observations were excluded.

After obtaining the sensor and inspection data set, a "filter" was performed to make the values consistent and as a way of controlling the variability of the phenomenon. In the weight variable (kg), removing observations from the BBCC apiary with weights above $100kg$, as they presented high variation in a short time, what was considered to be *measurement error*. Subsequently we removed observations with internal temperatures greater than $40°C$, due to the thermal protection characteristics of the hives, the bees tend to control the temperature and therefore internal temperatures above $40°C$ already are considered abnormal. Two variables were also extracted from the variable "date of collection", the categorical variable "season of the year" (spring, summer, autumn and winter) and the binary variable "day shift" (day and night).

Thus, a dataset of 661,025 observations has 16 variables: day shift, season, internal temperature ($°C$), weight (kg), external temperature ($°C$), humidity (%), pressure (hPa), dew point ($°C$), precipitation (mm), speed wind (km/h) and the six binary inspection factors (brood, adult bees, queen, food, no stressors and suitable space).

2.3. Elastic Net Logistic Regression

Regression analysis is a statistical technique to investigate and model the relationship between variables, these variables can be dependent or independent according to the modeling purpose.

Logistic regression is a statistical method for modeling a classification problem. The logistic regression response variable has K levels $Y = \{1, 2, \ldots, K\}$, these levels will be obtained through clustering, section 2.4, and will be described and commented in the Results section. Assuming you have n observations and p independent variables. Let $y_i \in \{0, 1, \ldots, K\}$ be the value of the response variable for the observation

[4] https://beehealth.bayer.us/bayer-news-and-resources/setting-the-standard-for-managing-healthy-honey-bee-colonies

$i, i = 1,2,\ldots,n$ e $\mathbf{x_i} = (x_{i1},\ldots,x_{ip},1)^\top$ the i -th vector of observations from the $\mathbf{X}$ specification matrix. Then, the response variable is related to the explanatory variables according to Equation 1.

$$\mathbb{P}(Y = k|X = x) = \frac{e^{\beta_{0k}+\beta_k^\top x_k}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_l^\top x_l}}, i \in \{1,2,\ldots,n\}; k \in \{1,2,\ldots,K\}. \quad (1)$$

Logistic regression with *Elastic Net* or Elastic Net penalty is a mixture of the penalty terms *Lasso* [Tibshirani 1996] and *Ridge* [Hoerl and Kennard 1970], introduced by Zou and Hastie [Zou and Hastie 2005] to handle highly correlated variables and perform variable selection simultaneously. A non-negative regulatory term is added to the negative log-likelihood function, $-l(\beta; x_i)$ in order to control the "size" of the β vector coefficients, using the term *Ridge* , and select variables "resetting" their coefficients, using the term *Lasso*. The negative value of the log-likelihood function of the logistic regression with penalty *Elastic Net* is given by

$$-l(\beta; x_i) = -l(\beta) + \lambda g(\beta). \quad (2)$$

Let $\mathbf{Y}_{N\times 3}$ be the design matrix and $y_{il} = \mathbb{I}(y_i = l)$, then we have

$$-l(\beta; x_i) = -\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - log\left(\sum_{k=1}^{K} e^{\beta_{0k}+x_i^T\beta_k}\right)\right)\right] + \lambda\left[(1-\alpha)||\beta||_F^2/2 + \alpha\sum_{j=1}^{p}||\beta_j||_q\right]. \quad (3)$$

Note that the Equation 3 has two parameters: α and λ, the parameter α controls the level of mixing between the two penalty methods, for $\alpha = 0$ we have the regularization *Ridge* and for $\alpha = 1$ we have regularization *Lasso*, for values of α between 0 and 1 a mixture of the two methods is incorporated. The parameter λ controls the impact of the penalty on the adjustment of the model and its coefficients. In general cases the values of these parameters are obtained via cross-validation experiments.

The solution of the logistic regression equation with *Elastic Net* penalty, Equation 3, is found using numerical methods, more specifically by the method of descending coordinates [Friedman et al. 2010].

2.4. k-Prototypes

We used the statistical tool of clustering k-prototypes for the variable definition that characterize the evolutionary scenarios and or working in environments with mixed types of variables (continuous and categorical). The k-prototype clustering algorithm is based on a mixture of k-means and k-modes methods.

The k-means algorithm, one of the most used for data clustering, is classified with a method of partial clustering. Given a $\mathbf{X}$ array of observations and a k ($\leq n$) number, the algorithm seeks to minimize the quadratic sum of intragroup errors by partitioning the $\mathbf{X}$ matrix into k groups.

One of the main concern with k-means is its limitation to solve cases in which the dataset is categorical since it was designed for environments where the data is con-

tinuous/numerical even converting categorical columns to *dummys* variables. To mitigate this problem, Huang (1998) proposed two methods (Equation 4), which both were used in this paper.

The dissimilarity function (distance) between two objects A and B, both with p attributes / variables, being v attributes of numerical character and c nominal / categorical attributes, such that $p = v + c$, is given by the Equation 4.

$$d_{proto}(A,B) = \sum_{j=1}^{v} (a_j - b_j)^2 + \gamma \sum_{j=v+1}^{c} \delta(a_j, b_j), \tag{4}$$

where the first term is the Euclidean distance measure squared in the numeric attributes and the second term is the dissimilarity function (distance) used in the k-modes method in the nominal / categorical attributes. The γ constant is used to avoid any "favoring" of one or more attributes in the cluster. In the k-prototypes algorithm, to obtain clusters, it minimizes a cost function based on the k-means algorithm [Huang 1998], modified to attend data with mixed type variables.

3. Results

3.1. Clustering

The number of classes was defined by analyzing the *scree-plot* graph. A grid search was performed with k, in which $k \in \{1,2,\ldots,10\}$, through the application of the k-prototype algorithm whose results can be seen in Figure 1.

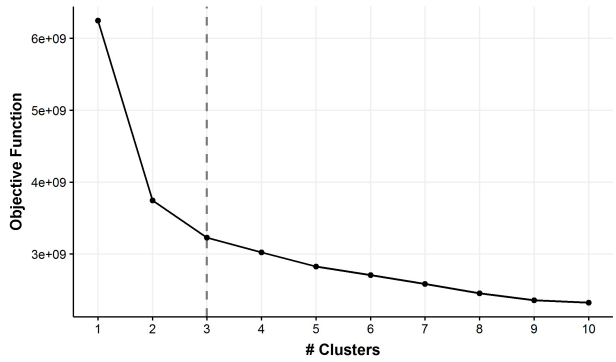

Figure 1. Objective function vs number of clusters using the k -prototypes algorithm.

After choosing the ideal number of classes, $k = 3$ (i.e. 3 clusters), we applied the k-prototype algorithm to the pre-processed data set and its prototypes were evaluated. Next, the interpretations of each obtained cluster are made and, in addition, Tables 2, 3 and 4 present basic statistics of the variables in each group.

3.1.1. Cluster #1

Regarding the moment of first cluster sample observation, we have a predominance (about 91%) of daytime samples and the majority observed during the summer (63.8%). The internal temperature ranged from approximately $11°C$ to almost $39°C$.

However, half of the samples are located between 23.89°C and 32.22°C, which is a good range for the colony. With respect to humidity, values below the ideal were recorded, with the first quartile equals to 50% and only a quarter of the samples with humidity above 75%. Finally, regarding the *in loco* inspections, we have satisfactory results for most items. However, it is important to note that 76.87% of the samples certified for a young laying queen. Furthermore, 37.02% of the samples registered the presence of stressors.

Table 2. Basic statistics for Cluster #1 variables with 277,063 samples.

(a) Continuous Variables

Statistics	Temperature Internal (ºC)	Weight (kg)	Temperature External (ºC)	Humidity (%)	Pressure (%)	Point of Dew (ºC)	Precipitation (mm)	Speed of Wind (km/h)
Minimum	11.11	0.01	8.00	11.00	995.0	-11.00	0.00	0.00
1st Quartile	23.89	23.72	22.00	50.00	1013.0	13.00	0.00	6.00
Median	27.78	29.16	25.00	63.00	1016.0	19.00	0.00	9.00
Mean	28.11	32.76	25.08	61.39	1016.0	16.67	0.0564	9.82
3rd Quartile	32.22	39.84	28.00	75.00	1019.0	22.00	0.00	12.00
Maximum	38.89	108.35	37.00	99.00	1033.0	26.00	4.7000	41.00

(b) Nominal / Categorical Variables

Season Year	Frequency
Summer	176751
Autumn	51291
Winter	486
Spring	48535

Shift of Day	Frequency
Day	252467
Night	24596

Valor	Brood	Adult Bees	Queen	Food	No Stressors	Suitable Space
Presence (1)	243485	263817	212982	247878	174497	241544
Frequency	87.88 %	95.22 %	76.87 %	89.47 %	62.98 %	87.18 %

3.1.2. Cluster #2

As noted in Table 3, cluster #2 presents a day/night sample ratio more balanced than Cluster #1, with a slight majority (about 60%) of night samples.

Table 3. Basic statistics for Cluster #2 variables with 158,252 samples.

(a) Continuous Variables

Statistics	Temperature Internal (ºC)	Weight (kg)	Temperature External (ºC)	Humidity (%)	Pressure (%)	Point of Dew (ºC)	Precipitation (mm)	Speed of Wind (km/h)
Minimum	-8.89	0.01	-11.00	11.00	996.0	-20.00	0.00	0.00
1st Quartile	7.78	25.43	6.00	46.00	1013.0	-2.00	0.00	6.00
Median	12.78	29.96	10.00	60.00	1018.0	2.00	0.00	9.00
Mean	12.57	33.52	9.487	59.90	1018.0	1.78	0.02	10.31
3rd Quartile	17.22	40.07	13.00	74.00	1023.0	6.00	0.00	13.00
Maximum	37.22	89.88	24.00	100.00	1040.0	16.00	3.30	55.00

(b) Nominal / Categorical Variables

Season Year	Frequency
Summer	13091
Autumn	90095
Winter	6045
Spring	49021

Shift of Day	Frequency
Dia	62592
Night	95660

Valor	Brood	Adult Bees	Queen	Food	No Stressors	Suitable Space
Presence (1)	141500	144831	101515	143465	130340	147698
Frequency	89.41 %	91.52 %	64.15 %	90.66 %	82.36 %	93.33 %

We also have an interesting number of autumn samples (almost 57%). By

checking continuous variables, the internal temperature changes from approximately $-9°C$ to just over $37°C$. It is a clearly decrease in temperatures, as result of the decrease in number of summer samples. Only 25% of the samples had temperatures equal to or greater than $17.22°C$. Regarding humidity, there was a quarter of the samples between 74% and 100%. Following the *in loco* inspections, 64.15% of the observations stated that there was a young laying queen and 82.36% of the samples showed no stressors.

3.1.3. Cluster #3

For the Cluster #3, the number of samples observed during the night is the majority, about 81.5%. With respect to the seasons, the majority is distributed between spring (39.7%) and summer (36.8%). In the description of the continuous variables, we have an internal temperature range of approximately $4°C$ to $35°C$. Examining the first quartile of the internal temperature, less than 25% of the samples recorded temperatures below $19°C$. Next, we have the humidity ranging from 44% to 100% in this Cluster #3. In addition, 75% of the samples recorded humidity above 82%. Moving on to the *in loco* inspections, we have ideal values. It is worth mentioning that almost 80% of the samples recorded the presence of a young laying queen. Another highlight is the stressors, just over 60% of the samples indicated no apparent stressors.

Table 4. Basic statistics for Cluster #3 variables with 225,710 samples.

(a) Continuous Variables

Statistics	Temperature Internal (ºC)	Weight (kg)	Temperature External (ºC)	Humidity (%)	Pressure (%)	Point of Dew (ºC)	Precipitation (mm)	Speed of Wind (km/h)
Minimum	3.89	0.01	3.00	44.00	995.0	2.00	0.00	0.00
1st Quartile	18.89	21.00	17.00	82.00	1013.0	14.00	0.00	6.00
Median	22,22	26,13	20,00	89,00	1016,0	18,00	0.00	9.00
Mean	21.02	26.39	19.32	86.87	1016.0	17.29	0.11	9.86
3rd Quartile	23.89	30.52	22.00	93.00	1019.0	21.00	0.00	12.00
Maximum	35.00	82.56	29.00	100.00	1032.0	25.00	11.90	39.00

(b) Nominal / Categorical Variables

Season Year	Frequency
Summer	83003
Autumn	52660
Winter	410
Spring	89585

Shift of Day	Frequency
Day	41661
Night	184049

Valor	Brood	Adult Bees	Queen	Food	No Stressors	Suitable Space
Presence (1)	183586	199778	180032	197344	135969	189375
Frequency	81.33 %	88.51 %	79.76 %	87.43 %	60.24 %	83.90 %

3.2. Classification

The parameters defined in the cross-validation experiment eliminated any need for the *Elastic Net* penalty, since the λ parameter that provided greater accuracy was $\lambda = 0$. Thus, with the current data and the value of $\lambda = 0$, the usual logistic regression was used. The results of applying the *Elastic Net* logistic regression to the test set are shown in Table 5.

High discrimination is observed between the scenarios, given that the correct answers (elements of the main diagonal) are greater than the errors (other elements). This result shows that the data from internal and external sensors are sufficient to explain almost all the variability of the defined variable, which is showed in Table 5(b).

Table 5. Basic Statistics Metrics from the model

(a) Confusion Matrix for Logistic Regression *Elastic Net* with *repeated cross-validation*.

		Predictions		
	Scenario	**1**	**2**	**3**
	1	82645	86	139
Real	**2**	81	47239	108
	3	392	150	67466

(b) Metrics obtained from the combination of the 3 scenarios.

Metrics					
Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**	**Log loss**
0.9952	0.9952	0.9951	0.9952	0.9999	0.0342

The high discrimination of the 3 clusters (subsections 3.1.1, 3.1.2 and 3.1.3) by the classifier is partly due to to the pattern found during the annual cycle and its high correlation with the seasons, as shown in Figure 2. The colors black, gray and light gray represent Clusters 1, 2 and 3, respectively. At the summer, Cluster 1 has the higher proportion, this rate decreases as summer ends and increases when summer approaches again. Conversely, Cluster 2 has the lowest rate in the summer, furthermore in the beginning of autumn it increases and it is the highest in winter and the first month of spring, but starts to decrease as summer approaches. Moreover, regarding Cluster 3, its rate is higher only in May, but it maintains a substantial rate from mid-spring to mid-autumn, i.e. from April to October.

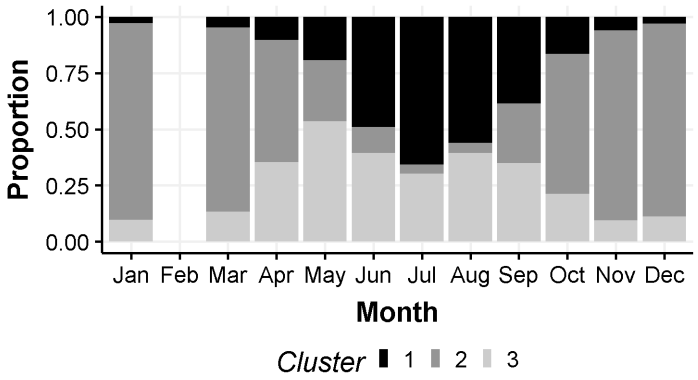

Figure 2. Scenarios along the months of the year. The month of February was hidden due to few observations in the period, all from the same apiary.

4. Discussion

We can highlight the difference between Cluster #1, Cluster #2, and Cluster #3 with respect to (i) an undesirable state, which represents risk to the colony health; (ii) a relatively good but not ideal state; and (iii) a state very close to the ideal, which represents the best colony healthy state. Thus, ordered by increasing health and development potential of the colony, according to section 3.1 we have Cluster #2 as being

a worrying state, Cluster #1 as being a good state and Cluster #3 as being the optimal state.

The internal temperatures of Cluster #2 present a risk to the colony health, as shown in Table 3(a). While Clusters #1 and #3 have adequate internal temperature ranges, the first one are even better though (Table 2(a) and 4(a)). Regarding humidity, the Cluster #3 stands out positively from the others, about 75% of the samples with humidity equal to or greater than 82%, the range of 90 to 95% being ideal for the brood rearing [Abou-Shaara et al. 2017]. Thus, the Cluster #3 represents the best level of colony health. Similarly, for nominal variables, in relation to the presence of a young queen ($age < 1year$), according to Table 3(b) we have in Cluster #2 the lowest rate whereas Cluster #3 has the best (almost 80%), as shown in Table 4(b). However, regarding the absence of stressors, we have the highest rate (82.36%) in Cluster #2, since this cluster has fewer samples during the summer, the time of year when there are more stress factors for the colony.

Furthermore, regarding the parameter λ which defines the degree of penalty applied to the model and which was defined as 0 (zero), in the case of the usual logistic regression, it does not rule out the possibility that the addition of more variables, e.g. more humidity sensors or temperature, problems of multicollinearity or adding variables that do not add information appear. On the other hand, based on sensor data, our proposal is able to determine whether there are problems, whether these problems are relevant or not and, through diagnosis, readjust the model, all based on its precision.

5. Conclusion

Here we propose a two-fold statistical model with i) a clustering method used to determine and characterize the evolutionary states of the colonies through sensor measurements and inspection, and ii) a classification method to obtain a model with a very high precision (99.5%) to classify the clusters/scenarios without inspection data, since we aim to diagnose the colony without having to open it. The main contribution of this paper is a high precision model that can adapt to new data naturally and making it possible to determine with a high degree of certainty when the colony needs immediate intervention (Cluster #2), when it may need some attention (Cluster #1) and when it is well and healthy, without the need for external assistance (Cluster #3).

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. Danielo G. Gomes and Breno Freitas thanks the financial support of the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) [grant numbers #302934/2010-3, #310317/2019-3, #432585/2016-8, #129426/2018-0].

References

Abou-Shaara, H., Owayss, A., Ibrahim, Y., and Basuny, N. (2017). A review of impacts of temperature and relative humidity on various activities of honey bees. *Insectes Sociaux*, 64(4):455–463.

Barron, A. B. (2015). Death of the bee hive: understanding the failure of an insect society. *Current Opinion in Insect Science*, 10:45 – 50. Social Insects * Vectors and Medical and Veterinary Entomology.

Braga, A. R., Furtado, L., Bezerra, A. D., Freitas, B., Cazier, J., and Gomes, D. G. (2019). Applying the long-term memory algorithm to forecast thermoregulation capacity loss in honeybee colonies. In *CSBC 2019 - 10º Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA)*, pages 1–14.

Braga, A. R., Gomes, D. G., Rogers, R., Hassler, E. E., Freitas, B. M., and Cazier, J. A. (2020). A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies. *Computers and Electronics in Agriculture*, 169:105161.

Dineva, K. and Atanasova, T. (2018). Osemn process for working over data acquired by iot devices mounted in beehives. *Current Trends in Natural Sciences*, 7:47–53.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, page 1–22.

Gil-Lebrero, S., Quiles-Latorre, F. J., Ortiz-López, M., Sánchez-Ruiz, V., Gámiz-López, V., and Luna-Rodríguez, J. J. (2017). Honey bee colonies remote monitoring system. *Sensors*, 17(1).

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283–304.

Jacobs, M., Cazier, J. A., Wilkes, J. T., Rogers, R., and Hassler, E. E. (2017). Building a business analytics platform for enhancing commercial beekeepers' performance: Descriptive validation of a data framework for widespread adoption by citizen scientists. In *23rd Americas Conference on Information Systems, AMCIS 2017, Boston, MA, USA, August 10-12, 2017*, pages 611 – 620.

Kridi, D. S., de Carvalho, C. G. N., and Gomes, D. G. (2016). Application of wireless sensor networks for beehive monitoring and in-hive thermal patterns detection. *Computers and Electronics in Agriculture*, 127:221 – 235.

Meikle, W. G. and Holst, N. (2015). Application of continuous monitoring of honeybee colonies. *Apidologie*, 46(1):10–22.

Meikle, W. G., Weiss, M., Maes, P. W., Fitz, W., Snyder, L. A., Sheehan, T., Mott, B. M., and Anderson, K. E. (2017). Internal hive temperature as a means of monitoring honey bee colony health in a migratory beekeeping operation before and during winter. *Apidologie*, 48(5):666–680.

Sánchez-Bayo, F. and Wyckhuys, K. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, 232.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Zacepins, A., Brusbardis, V., Meitalovs, J., and Stalidzans, E. (2015). Challenges in the development of precision beekeeping. *Biosystems Engineering*, 130:60 – 71.

Zogovic, N., Mladenovic, M., and Rasic, S. (2017). From primitive to cyber-physical beekeeping. In *Zdravkovic, M., Konjovic, Z., Trajanovic, M. (Eds.) 7th International Conference on Information Society and Technology ICIST 2017 Proceedings Vol.1*, pages 38–43.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.