2ND CONFERENCE ON
STATISTICS AND
DATA SCIENCE
Salvador, Brazil, November 18-20, 2019

Clustering and Elastic Net Logistic Regression as Support Tools for Honeybee (*Apis mellifera*) Colonies Health Diagnosis

Daniel A. Silva, Antonio Rafael Braga, Juvêncio S. Nobre, Danielo G. Gomes

UFC DEMA PPGETI sm@rtBee bee care BAYER

Bee studied

Figure: European honey bee (*Apis mellifera*)
Available in: http://apicultura.to.gov.br/wp-content/uploads/2018/04/Apis-mellifera_1860.jpg

Why are bees important?

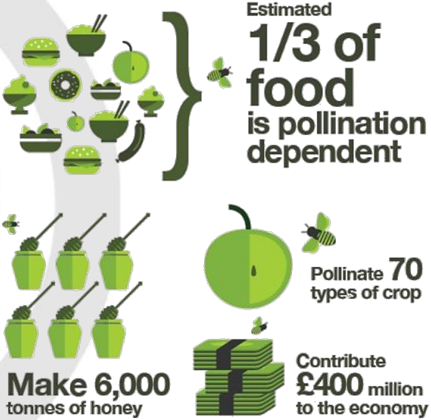

Estimated
1/3 of food
is pollination dependent
Pollinate 70 types of crop
Make 6,000 tonnes of honey
Contribute £400 million to the economy

Figure: Bee benefits
Available in: https://ichef.bbci.co.uk/images/ic/1280xn/p07crcfm.jpg

Inside the Hive

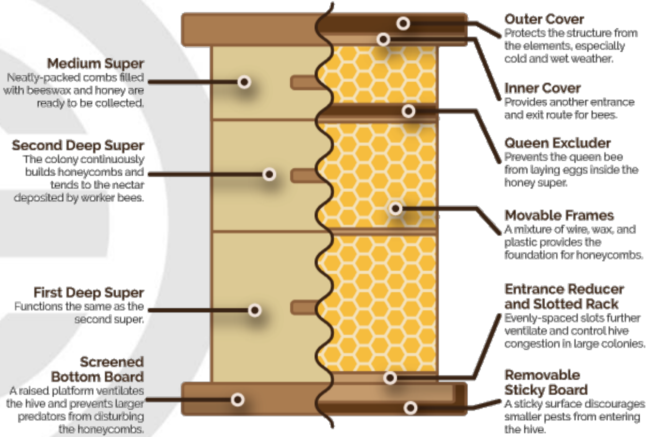

Medium Super
Neatly-packed combs filled with beeswax and honey are ready to be collected.
Second Deep Super
The colony continuously builds honeycombs and tends to the nectar deposited by worker bees.
First Deep Super
Functions the same as the second super.
Screened Bottom Board
A raised platform ventilates the hive and prevents larger predators from disturbing the honeycombs.
Outer Cover
Protects the structure from the elements, especially cold and wet weather.
Inner Cover
Provides another entrance and exit route for bees.
Queen Excluder
Prevents the queen bee from laying eggs inside the honey super.
Movable Frames
A mixture of wire, wax, and plastic provides the foundation for honeycombs.
Entrance Reducer and Slotted Rack
Evenly-spaced slots further ventilate and control hive congestion in large colonies.
Removable Sticky Board
A sticky surface discourages smaller pests from entering the hive.

Figure: Hive structure
Available in: https://ichef.bbci.co.uk/images/ic/1280xn/p07crcfm.jpg

In loco inspection causes stress to bees

Figure: Inspection in a Bee Colony
Available in: https://static.independent.co.uk/s3fs-public/thumbnails/image/2018/07/26/13/british-beekeepers-1.jpg

Beehive fragility in winter

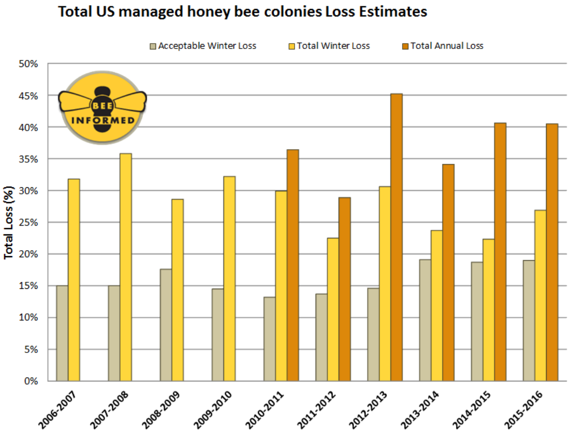


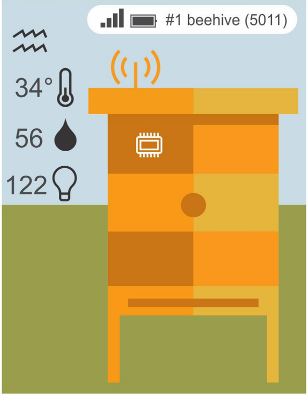
Figure: Annual loss in (%) of colonies in the United States in 2006-2016
Available in: https://beeinformed.org/wp-content/uploads/2016/11/BIP-2015-2016-Loss-Chart.png

Data collect

1. Internal sensors via the IoT (Internet of Things);
2. External sensors via US National Weather Service data;
3. *In-loco* Inspection via Healthy Colonie Checklist (HCC).

Internal sensors

1. Cluster and hive temperature
2. Cluster and beehive humidity
3. Beehive Weight



External sensors

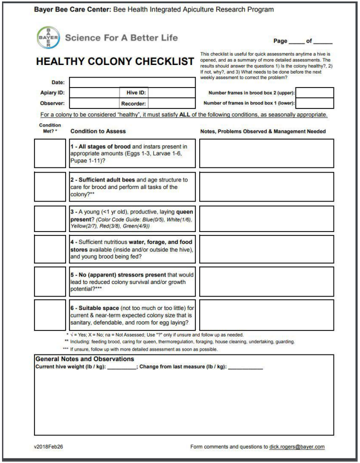
1. Temperature
2. Dew Point
3. Sky Condition
4. Pressure
5. Precipitacion
6. Wind Speed and Direction

Figure: Illustration of a weather station.

Inspection Factors

In-loco Inspection via HCC

1. Brood
2. Bees
3. Queen
4. Food
5. Stressors
6. Space



Bayer Bee Care Center: Bee Health Integrated Apiculture Research Program

Science For A Better Life

Page _____ of ______

HEALTHY COLONY CHECKLIST

This checklist is useful for quick assessments anytime a hive is opened, and as a summary of more detailed assessments. The results should answer the questions 1) Is the colony healthy?, 2) If not, why?, and 3) What needs to be done before the next weekly assesment to correct the problem?

Date:			
Apiary ID:		**Hive ID:**	
Observer:		**Recorder:**	

Number frames in brood box 2 (upper): □

Number of frames in brood box 1 (lower): □

For a colony to be considered "healthy", it must satisfy **ALL** of the following conditions, as seasonally appropriate.

Condition Met? *	Condition to Assess	Notes, Problems Observed & Management Needed
□	1 - **All stages of brood** and instars present in appropriate amounts (Eggs 1-3, Larvae 1-6, Pupae 1-11)?	
□	2 - **Sufficient adult bees** and age structure to care for brood and perform all tasks of the colony?**	
□	3 - A young (<1 yr old), productive, laying **queen present**? *(Color Code Guide: Blue(0/5), White(1/6), Yellow(2/7), Red(3/8), Green(4/9))*	
□	4 - Sufficient nutritious **water, forage, and food stores** available (inside and/or outside the hive), and young brood being fed?	
□	5 - **No (apparent) stressors present** that would lead to reduced colony survival and/or growth potential?***	
□	6 - **Suitable space** (not too much or too little) for current & near-term expected colony size that is sanitary, defendable, and room for egg laying?	

* √ = Yes; X = No; na = Not Assessed; Use "?" only if unsure and follow up as needed.

** Including: feeding brood, caring for queen, thermoregulation, foraging, house cleaning, undertaking, guarding.

*** If unsure, follow up with more detailed assessment as soon as possible.

General Notes and Observations

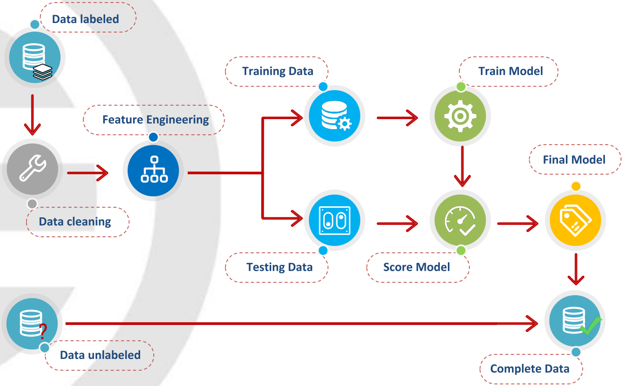
Current hive weight (lb / kg): __________; Change from last measure (lb / kg): ____________

v2018Feb26

Form comments and questions to dick.rogers@bayer.com

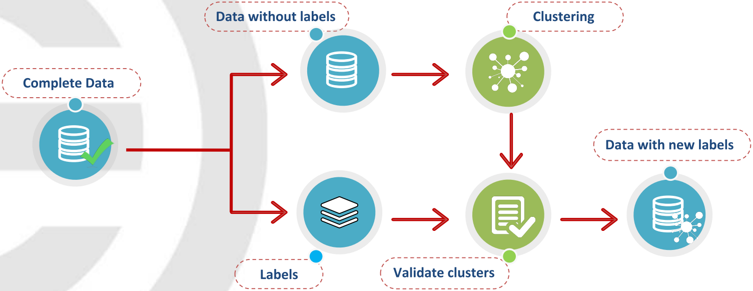
Figure: Healthy Colony Checklist (HCC).

Algorithm 1 - Getting Complete Data



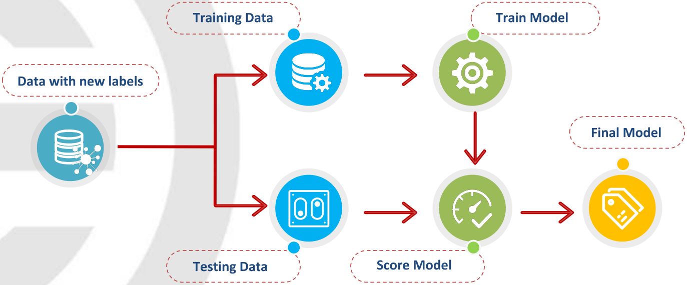
The model chosen for algorithm 1 was the Random Forest, as it is robust in class imbalance problems.

Algorithm 2 - Getting new labels with Clustering



The clustering method chosen was CLARA (Clustering Large Applications), as it works with large data sets using resampling.

Algorithm 3 - Final Model



The model chosen to classify the new data set was the Elastic Net Logistic Regression, dealing with multicollinearity and performing feature selection.

Random Forest Model - Algorithm 1

- Hyperparameters by 10 fold 5-repeatedCV
  1. mtry (#Randomly Selected Predictors) = 5
  2. splitrule (Splitting Rule) = gini
  3. min.node.size (Minimal Node Size) = 1
- Test Data Confusion Matrix

		Predicted					
		1	2	3	4	5	6
Real	1	82	1	4	0	1	0
	2	4	171	2	1	4	0
	3	0	1	88	2	5	0
	4	0	7	10	837	40	24
	5	1	13	25	140	2368	187
	6	2	3	5	44	104	1053

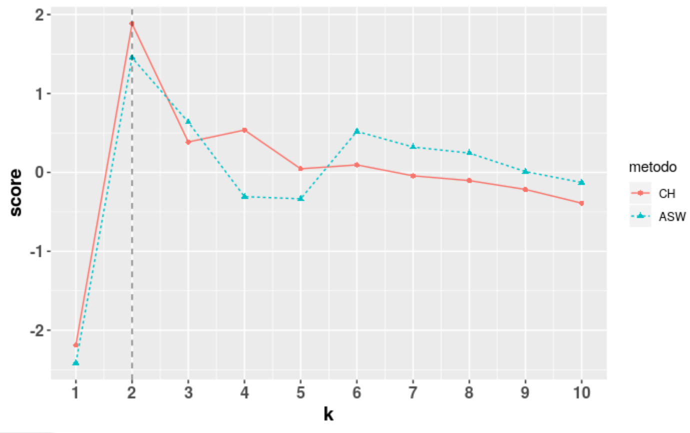
Random Forest Model - Algorithm 1

- Accuracy = 0,8795
- Others statistics

	Precision	Recall	F1
Class: 1	0,93	0,92	0,93
Class: 2	0,94	0,87	0,90
Class: 3	0,92	0,66	0,77
Class: 4	0,91	0,82	0,86
Class: 5	0,87	0,94	0,90
Class: 6	0,87	0,83	0,85

CLARA Clustering - Algorithm 2

- Choosing the best number of clusters by the Silhouett and Calinski-Harabasz indices



CLARA Clustering - Algorithm 2

- Validation of cluster medoids

	Turn Day	Brood Temperature	Brood Humidity	Hive Temperature	Hive Humidity
1º Cluster	day	29,04	62,00	27,71	61,00
2º Cluster	day	33,99	69,00	34,14	69,00

	Weight	External Temperature	Dew Point	Wind Direction	Wind Speed
1º Cluster	22,66	2,06	1,83	150,00	2,60
2º Cluster	33,90	23,90	17,80	90,00	26,00

Elastic Net Logistic Regression - Algorithm 3

- Hyperparameters by 10 fold 5-repeatedCV
  1. alpha (Mixing Parameter) = 0,27
  2. lambda (Regularization Parameter) = 0,04
- Test Data Confusion Matrix

		Predicted	
		0	1
Real	0	2221	5
	1	4	3000

Elastic Net Logistic Regression - Algorithm 3

- Accuracy = 0,9983
- Precision = 0,9978
- Recall = 0,9982
- F1 = 0,9980

Conclusions

1. Real-time monitoring proposal.
2. Uses data in the winter period.
3. It can avoid "unnecessary" inspections.
4. Scalable application by resampling clustering.
5. Highly discriminative model (acc = 99,83%)

Future works

1. Use of semi-supervised techniques;
2. Application to a larger data set;
3. Application in Brazilian Beehives;

References I

- Calinski, Tadeusz and Jerzy Harabasz (1974). “A dendrite method for cluster analysis”. In: *Communications in Statistics-theory and Methods* 3.1, pp. 1–27.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*.
- Kaufman, Leonard and Peter Rousseeuw (1990). “Finding Groups in Data: An Introduction to Cluster Analysis”. In:
- Rousseeuw, Peter (Nov. 1987). “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *J. Comput. Appl. Math.* 20.1, pp. 53–65. ISSN: 0377-0427.
- Zou, Hui and Trevor Hastie (2005). “Regularization and variable selection via the Elastic Net”. In: *Journal of the Royal Statistical Society, Series B* 67, pp. 301–320.

Thank you