Current Trends in Natural Sciences Vol. 7, Issue 13, pp. 47-53, 2018

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

OSEMN PROCESS FOR WORKING OVER DATA ACQUIRED BY IOT DEVICES MOUNTED IN BEEHIVES

Kristina Dineva [1*], Tatiana Atanasova [1]

[1] Institute of Information and Communication Technologies - Bulgarian Academy of Sciences
Acad. G. Bonchev St., Bl. 2, Sofia, 1113 Bulgaria

Abstract
Approaches for obtaining, clearing, studying, modelling, and interpreting collected IoT data are an important issue and a serious challenge for many researchers. The introduction of a standardized model of work - OSEMN, organizes the process of solving the problems. Beekeeping is a sub-sector in agriculture and it needs a unified process to work with data being obtained from sensors located in beehives. After applying a proper data processing, significant knowledge about the behaviour of individual bee colonies is gained, helping to identify correlations between the different events and the causes that invoke them. The purpose of this article is to describe the OSEMN model and its integration into beekeeping.

Keywords: Beekeeping, internet of things, OSEMN.

1. INTRODUCTION

Technology development is increasingly being used in the environmental care of the planet, for example in the field of agriculture and endangered species such as bee colonies. Beekeeping is one of the agricultural sub-sectors where the new technologies, models and processes can be successfully adapted and implemented. Thanks to their use in beekeeping, the knowledge about the bees and their various conditions is improved with certain parameters.

To obtain up-to-date data on bee families, Internet of things (IoT) sensor devices are built to help to get data on the parameters within and outside the hive over a certain time interval. The data obtained from each beehive needs to go through a number of processing steps before it becomes ready for extracting knowledge by which the future states of the beehive families can be predicted. Integrating different methods, technologies, and processes allows for correct and accurate organization of work with the data gathered from the beehives. Good data organization and the follow-up of standardized processes such as the OSEMN (Mason and Wiggins, 2010) increase the probability of doing accurate analyzes and makes it easy going back to a particular step from the data processing procedure. The information obtained after processing the extracted data, enables beekeepers to be informed in a timely manner about deviations of the beehive parameters and the probability of occurrence of a specific event related to bee behaviour. By integrating new technologies into beekeeping, beekeepers are enabled to make the industry more efficient by reducing costs, increasing and integrating new sources of income.

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

2. MAJOR PROBLEMS

Bees have the biggest role in polluting fruits, vegetables, flowers and farm crops like alfalfa that used for feeding many animals in agriculture. Pollination depends on bees for more than 1/3 of the world's crops. Bee colonies can be considered as a super organism, which consists of 40 to 50 thousand individual bees. They are a very important part of the earth ecosystem.

There is a 300% increase in crops globally, that need bees for pollination. There is also an increasing mortality rate on beehive families on a global scale for three consecutive years. The losses reach 80% in some places. Many factors, independently of each other lead to the death of the bee families.

Researchers in that field highlight three major reasons for the increasing mortality (Figure 1).

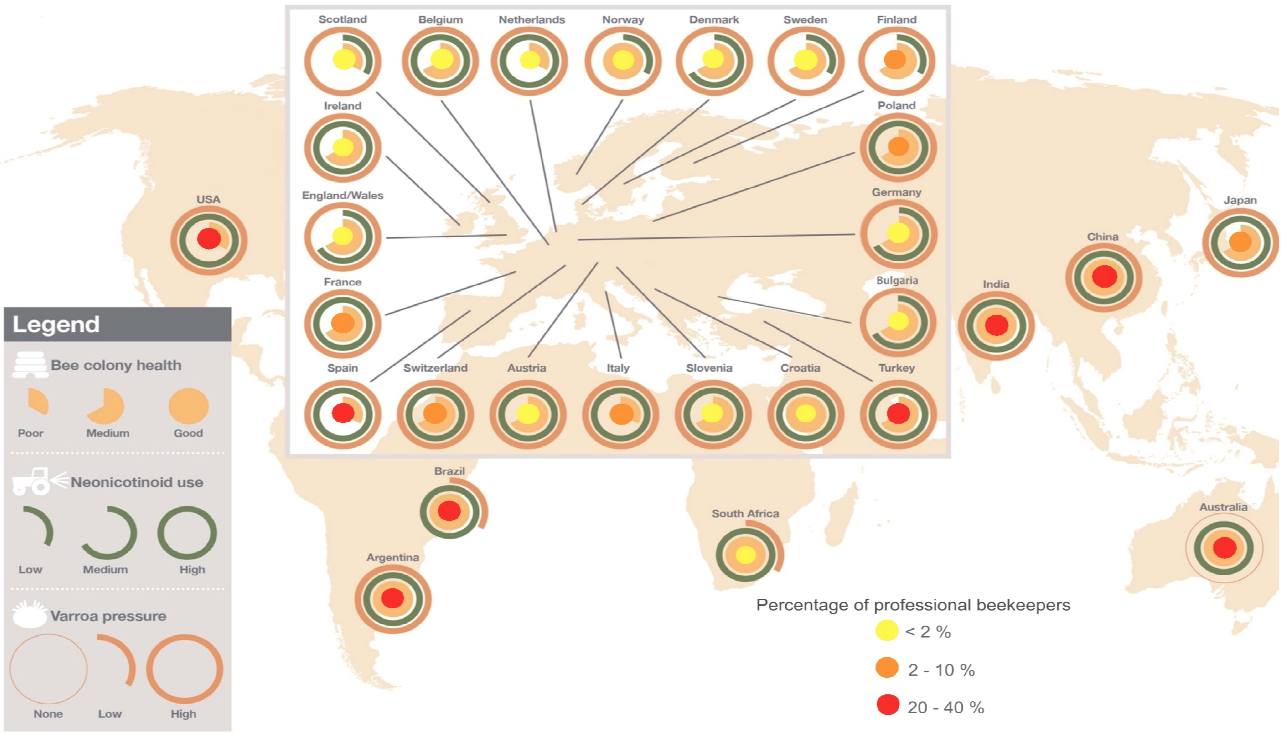

Figure 1. Major problems of bee families in the world, Source: Syngenta 2016 [recasted]

Unregulated pesticide spraying on crops located near beehives during the pollination period is one of the major problems (Syngenta, 2016). More than two-thirds of the pollen that bees collect and carry to their hives is polluted with a cocktail of 17 different toxic pesticides - stated by Greenpeace International. The chemicals found in pollen are from insecticides, acaricides, gungicides and herbicides. After the pollen is brought to the hive, bees that do not go out to collect nectar and pollalso can be poisoned too. Bees can die in the field from fast-acting pesticides, or also in the hive or near it from the slow-acting pesticides. The poisonous substance can be obtained from the bee-visited plants treated with it or can be carried by the wind from non-attractive bee cultures that are in bloom or flowering weed or other honey-growing vegetation (Krupke et al., 2012; Allsopp et al, 2014). This problem exists not only in Europe. This is a global problem for which countries such as England, France and Germany have taken temporary measures that are subject to monitoring and analysis (Neumann and Carreck, 2015).

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

Another major problem directly related to the death of bees is Varroa mite - a parasite that attacks the bees. Varroa mites are external parasites attacking honeybees and they can be seen with naked eye. This mite is found inside honey's cells and the bee's body. Either way, they feed on the hemolymph (blood) of the bees through the bee's body wall. In cases of severe infection, bees, which have defects, will not develop properly and may not be perfectly normal like adult bees. Bees which do not show defects are too weak and so they have shorter lives. Varroa mite lives on the body of the bee but their growth occurs inside the closed cells, especially closed cells of male workers. Nowadays Varroa mite is very important for the beekeeping industry in the world (Mohammadreza et al., 2015). The problem with this parasite is very prevalent in Europe and North America and has led to catastrophic losses since 2006 (Conte et al., 2010).

Besides these two factors, there is also a third one that most strongly affects the health and productivity of the bee families – it is the beekeeper's competence and skills as well as the time and resources invested in growing the bee colonies. The regular inspection of each beehive is of particular importance. Beekeeping inspections are often one of the most complex tasks even for professional beekeepers.

Studies show that more than half of the world's beehives are located outside urban areas. For this reason, it is difficult to carry out the required number of inspections of each beehive for a variety of reasons, such as bad meteorological conditions, poor infrastructure or the increasing transportation costs, which in turn leads to reduced profitability. On the other hand, carrying out inspections has an adverse effect on the overall condition of the bee family. Studies have shown that bees need three to four days to stabilize and restore the ideal temperature and humidity parameters in the hive (Verboven et al., 2014) after inspections.

The rapid development and the ease of integration of new technologies in beekeeping enable remote monitoring of both the internal condition of hives and outside also the conditions. The deployment of IoT devices in each individual hive allows collection by sensors of large volumes of data about the hive’s parameters, which in turn, by analyzing them, makes it possible to detect and classify the reasons explaining the increasing bee mortality. However, working with these large data volumes is often a very hard process (Balabanov, 2016). With the use of standardized workflow processes such as OSEMN, it is easier to apply accurate models and the results of the analysis of large data volumes could predict deviations in the behaviour of the bee family at a much earlier stage, allowing the beekeepers to take the right measures on time.

3. OSEMN WORKFLOW

The OSEMN process is a standardized and widely accepted model of organization of research in the field of Data Science. The OSEMN process solves the problems with Data Science/Analytics (Byrne, 2017) at a large scale.

The process of retrieving and manipulating beehive data needs to be organized, well prepared and pre-processed. The use of the OSEMN process provides a clear order of activities - Obtain, Scrub, Explore, Model the data and iNterpret the data (Figure 2). By following these steps, the entire process can be well planned and organized – starting with data acquisition to the analyzed data results visualization in specially developed software platform such as www.smartbeehives.eu (Dineva and Atanasova, 2017a) for the honey beehives monitoring.

The process of work is well developed and organized. It consists of several logical consecutive activities through which the original goals are achieved.

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

Obtain data

Data is collected from sensors that are located inside and outside hives. The sensors located inside hives collect data about temperature, humidity, weight, noise levels, and more. These data are used to monitor the condition of the bee families. External sensors are situated in different locations in the apiary and collect environmental information (temperature, humidity and CO_2) that gives a clear and accurate idea of the particular weather, air pollution, and so on.

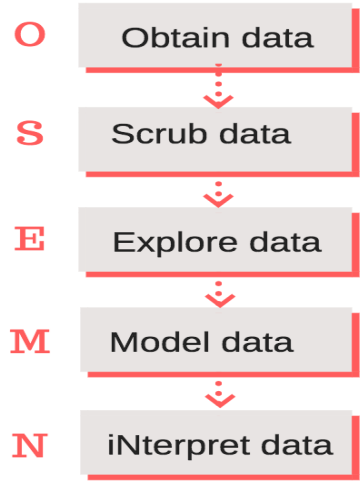

Figure 2. OSEMN workflow

Sensors are grouped according to the specific data to the user needs. A group of sensors are connected to a common microcontroller, thus forming a node. One node collects specific data types, allowing the simultaneous operation of different nodes, with virtually unlimited number of nodes. Depending on the system load and the size of the apiary, the microcontrollers can be several types - *arduino, msp430 launchpad, nanode, pinguino pic32, stm32 discovery* and others. The most popular is *arduino* because it offers the necessary quality at a very good price. The honeybee monitoring system (Dineva and Atanasova, 2017b) is designed in a way that it can easily integrate different types of microcontrollers, making it extremely flexible and adaptable for the different use cases that arise according the needs of the end customer.

It is desirable the used programming language to be a scripting one that can help automating data extraction and allowing asynchronous handling of the received data. The modern dev world provides several programming languages, and Python is a prominent preferred among others.

Scrub data

Before processing and analyzing the obtained from the honey bee monitoring system data, several actions need to be carried out: merging the individual data columns into a single table, clearing the data from invalid values, normalizing the data and processing the extreme values (Vander, 2016).

- *Merging the individual data columns into a single table* - with the help of the intelligent beehive monitoring system, data for different parameters (temperature, humidity, weight, sound, etc.) is collected and then it needs to be merged into a single common table where the parameters remain as column variables.

- *Clearing data from invalid values* - When collecting real-world data, there are often different reasons why there are null, NaN or NA values. As a result the analysis may be incorrect and the data models might be created the wrong way for the reason that most of the predictive

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

algorithms can not cope well with missing or invalid data. The most common approaches to solving this missing data problem are to replace them with average values or to directly delete these values. If the data array is stacked then the missing data can be replaced with the next closest value in ascending or descending order. For the purposes of the bee-monitoring project, data sequence is important and array sorting is not performed, so the missing values are not replaced with nothing.

- *Data normalization* - the data obtained from the monitoring system needs to be normalized because of the different types and ranges. The type of normalization used is *min-max*. Min-max normalization performs a linear transformation to the initial data (Kantardzic, 2011), where min_a and max_a are the minimum and maximum values for the attribute *a*. Min-max normalization maps values *v* of the range [min_a , max_a] into a particular range [*new* min_a , *new* max_a] by computing:

$$v' = [(v\text{-}min_a)/(max_a - min_a)] * (new\ max_a - new\ min_a) + new\ min_a \quad (1)$$

- *Extreme value processing* - A *RANSAC* (RANdom SAmple Consensus) algorithm is used to determine the extremes in a dataset. This algorithm provides a statistical estimation of the probability of obtaining reliable forecasts, i.e. probability within a predetermined number of standard deviations from the true values. Also, the algorithm can be interpreted as a method of detecting emergency situations. It is a non-deterministic algorithm - it produces a reasonable result with only a certain probability and this probability can increase further because replications are allowed. The algorithm produces and validates a linear QSAR (Quantitative Structure–Activity Relationships) model based on the Minimum Least Square (LMS) method by (Kaspi et al., 2017):

- filtering noisy samples (i.e., outliers);
- selecting the best features (i.e., descriptors);
- deriving a QSAR model from training set samples;
- predicting the activity of test set samples while invoking the concept of applicability domain, all in a single process without the need of complementary processes.

Explore data (EDA)

Finding, structuring, and enriching are operations that are extremely useful for exploring the gathered data. Observing the raw data set helps choosing the best approach for conducting analytical research. This allows the discovery and understanding of unique data elements, such as extreme or unordinary values. This approach is used to generalize the data obtained from beehives and to summarize their main characteristics. The main objectives for applying this approach to the sensor readings are:

- Creating hypotheses about the causes of observed phenomena;
- Assisting for the selection of appropriate statistical tools and techniques;
- Build a basis for future data collection through surveys and experiments;
- Identifying relationships between variables.

The perspective of exploratory data analysis (EDA) is described in a simple formula:

$$Data = Smooth + Rough \quad (2)$$

This means that data should be divided into two parts. The first part is called “smooth” and refers to models that can be extracted from raw data using different techniques. EDA techniques focus on extracting the "smooth" of each set of data. The smooth, whatever, comes from the data and does not stem from our expectations or our assumptions about the data. This means that the first step in

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

the EDA process is to extract the data smoothly. A variable may have more than one template or smooth. Retrieving smooth from raw data may require more than one pass through it and may have more than one template that the data contains.
The second part of the formula, "rough" is the remaining leftovers that do not have a template at all. Leftovers are what are left after all models are extracted from a dataset. However, it is very important to look closely at the "rough" because this set of values may contain additional models that need to be considered (Borcard et al., 2011; Waltenburg and McLauchlan, 2012).

Model data

In the machine learning paradigm, the model refers to a mathematical expression of the model parameters, along with the inserted substitutes for each prediction, class and action for regression, classification, and reinforcement categories, respectively.
Modelling in the bee monitoring system is used to predict, hence it requires good theoretical and mathematical knowledge. Models can range from classical logistic regression, to a more complex state machine or random forest to classify something, if a prediction or establishing trends are aimed.
The generated model receives inputs with predefined structure (prepared, cleaned, normalized, etc.). The python module *scikit-learn* (Hackeling, 2014) was used to easily create and use patterns in the bee-monitoring project.
Scikit-learn is used for data modelling in the beehive monitoring system as a module that integrate classical machine learning algorithms into several scientific Python packages:

- *NumPy* – Base n-dimensional array package;
- *SciPy* – Fundamental library for scientific computing;
- *Matplotlib* – 2D – 3D plotting;
- *Seaborn* – Visualization of statistical models
- *Ipython* – Enhanced interactive console;
- *Sympy* – Symbolic Mathematics;
- *Pandas* – Structures and data analysis.

Interpret the data

The last and perhaps the most important step in the OSEMN model is the interpretation of the data. The examination phase must answer completely or partially to the questions that provoke the data modelling processes and needs. This is the stage during where it is used everything learned from the collected and processed data from the beehive monitoring system.
This step includes:

- Drawing conclusions from the data;
- Evaluation of results;
- Propagation / reporting of results.

Interpretation of research results is important for understanding the effectiveness of the study. It is necessary to clearly describe the results in a way in which other researchers can compare with their own results. Proper understanding of the methodology and survey statistics is necessary for the correct interpretation. The results are analyzed using appropriate statistical methods to determine the probability that the results were not random and can be reproduced in larger studies.
The results should be interpreted in an objective and critical way before assessing the implications and drawing conclusions.

*Corresponding author, E-mail address: k.dineva@iit.bas.bg

Current Trends in Natural Sciences (on-line)
ISSN: 2284-953X
ISSN-L: 2284-9521

Current Trends in Natural Sciences (CD-Rom)
ISSN: 2284-9521
ISSN-L: 2284-9521

4. CONCLUSIONS

Choosing a standardized data processing workflow is a combination of expected functionality and ease of implementation. The orientation towards standardized processes is a philosophy of work that shifts focus from the activities to the results because activities are dealt within their coordinated aggregation in creating value for the end result. The end result of the obtaining, clearing, studying, modelling, and interpreting collected from the beehives IoT data is to predict events and to find a correlation between the analyzed data and events occurring in beehives. The easily understand and logically consistent steps of the data processing workflow (OSEMN), enriched with additional instructions, notes and sample documents, ensure the performance of the activities and the achievement of the results in the same way by the different participants. A higher maturity of standardization and harmonization of practices at different stages is achieved. Errors due to insufficient awareness are avoided. One of the most important features of standardized work processes is that they unequivocally describe not only the sequence but also the responsibilities. At each stage of the process, it is clear what is expected, who will receive a request for a particular activity and to whom the outcome should be provided. OSEMN can significantly facilitate the identification of existing good practices in the beekeeping.

Future investigation are directed to achieving transparency at each stage of the process which will allow easy detection of errors and quick step back to a certain stage of process if needed.

5. REFERENCES

Allsopp, M., Tirado, R., Johnston, P. (2014). Plan bee – living without pesticides, Greenpeace, pp. 56.

Balabanov, T., Zankinski, I., Barova, M. (2016). Strategy for individuals distribution by incident nodes participation in star topology of distributed evolutionary algorithms. *Cybernetics and Information Technologies*, *16*(1), 80-88.

Borcard, D., Gillet, F., Legendre, P. (2011). Numerical Ecology with R, Springer, pp. 9 – 30.

Byrne, C. (2017). Development workflows for Data Scientists. O'Reilly, pp.28.

Dineva, K., Atanasova, T. (2017). Computer System Using Internet of Things for Monitoring of Bee Hives, Vienna, SGEM.GeoConference (Vol. 17, Issue 63, pp. 169-176).

Dineva, K., Atanasova, T. (2017). Model of Modular IoT-based Bee-Keeping System. ESM'2017, Lisbon, EUROSIS-ETI, 404-406.

Hackeling, G. (2014). Mastering Machine Learning with scikit-learn, PACKT, pp. 238.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms, IEEE, pp. 552.

Kaspi, O., Yosipof, A., Senderowitz, H. (2017). Random Sample Consensus (RANSAC) algorithm for material – informatics: application to photovoltaic solar cells, Springer, 6, 2 – 15.

Krupke, Christian H., Hunt, Greg J., Eitzer, Brian D. (2012). Multiple Routes of pesticide exposure for honey bees living near agricultural fields, PLOS.

Le Conte, Y., Ellis, M., Ritter, W. (2010). Varroa mites and honey bee health: can Varroa explain part of the colony losses?. *Apidologie*, *41*(3), 353-363.

Maadinia, M., Shabestari, B., Mahmoudi, R., Kaboudari, A., Rahimi Pir-Mahalleh, S. F. (2015). Evaluation of Varroa Mites in the Apiaries from Iran. *International Journal of Food Nutrition and Safety*, *6*(2), 74-81.

Mason, H., Wiggins, C. H. (2010). A Taxonomy of Data Science. Retrieved November 2017, from http://www.dataists.com/2010/09/a-taxonomy-of-data-science.

Neumann, P., Carreck, N. L. (2015). Honey bee colony losses, Journal of Apicultural Research.

Syngenta, (2016). Financial Report 2016, Retrieved February 2017 from https://www.syngenta.com/~/media/Files/S/Syngenta/ar-2016/syngenta-financial-report-2016.pdf

Vander, J. (2016). Python data science handbook, O'Reilly, pp.541.

Verboven, H. A., Uyttenbroeck, R., Brys, R., Hermy, M. (2014). Different responses of bees and hoverflies to land use in an urban–rural gradient show the importance of the nature of the rural land use. *Landscape and Urban Planning*, *126*, 31-41.

Waltenburg, E., McLauchlan, W. (2012). Exploratory Data Analysis: A primer for undergraduates, Purdue e-Pubs, pp.81.

*Corresponding author, E-mail address: k.dineva@iit.bas.bg