UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Predicting Drivers of Change's Impacts on Pollinators Bees Behaviour and Efficiency using Deep Learning

Thiago Joel Angrizanes Rossi

Monograph - MBA in Artificial Intelligence and Big Data

ICMC USP SÃO CARLOS

Thiago Joel Angrizanes Rossi

Predicting Drivers of Change's Impacts on Pollinators Bees Behaviour and Efficiency using Deep Learning

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence and Big Data

Advisor: Prof. Dr. Fernando Santos Osório

Original version

São Carlos

2023

Thiago Joel Angrizanes Rossi

Predicting Drivers of Change's Impacts on Pollinators Bees Behaviour and Efficiency using Deep Learning

Área de concentração: Inteligência Artificial e Big Data

Advisor: Prof. Dr. Fernando Santos Osório

São Carlos

2023

This work is dedicated to all planetary stewards
and guardians of our natural world, for their work of protecting nature's
most valuable individuals and our common home.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special thanks to the entire MBA team and teachers for all the work and effort put in delivering a top quality course and environment for scientific and innovation research.

Besides, I would like to thank my advisor, Professor Fernando Santos Osório, who guided me in doing this research. His valuable advice and help motivated me and contributed tremendously to the successful completion of the research.

Also, I would like to express my deep gratitude for my family, friends and my partner, who supported me. For their patience and honesty, partnership and guidance. Without that support, I wouldn't be able to conclude this course.

At last, but not in least, I would like to thank everyone involved, who helped and motivated me to work on this research.

"Everybody is a genius.
But if you judge a fish by its ability to climb a tree,
it will live its whole life believing that is is stupid."
Albert Einstein

ABSTRACT

Rossi, T.J.A. **Predicting Drivers of Change's Impacts on Pollinators Bees Behaviour and Efficiency using Deep Learning**. 2023. 38p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

In nature, approximately ninety percent of flowering plants rely on pollinators for pollen transfer and sexual reproduction. These plants are crucial for ecosystem functioning, providing food, habitats, and resources for many animal species, including humans. However, human activities are responsible for changes in ecosystem networks that negatively impact the behavior and efficiency of pollinating bees. The direct and indirect consequences surpass human understanding and conventional tools for addressing the functioning and responses of complex adaptive systems, thus requiring new approaches and techniques. Since pollinators are a highly sensitive part of an ecosystem, it was questioned whether Deep Learning algorithms could be an alternative to predict impacts of anthropogenic interventions on pollinator insect populations. Our study evaluated the performance of five algorithms (Linear Regression, Random Forests, Gradient Boosting Machines, Dense Neural Networks, and Long Short-Term Memory) in predicting occurrences of disruptions and anomalies that may impact bees and their hives. The most effective algorithms evaluated were LSTM and GBM, particularly in the dataset relating to the use of neonicotinoid pesticides.

Keywords: Pollinators. Drivers of Change. Prediction. Deep Learning.

RESUMO

Rossi, T.J.A. **Predicting Drivers of Change's Impacts on Pollinators Bees Behaviour and Efficiency using Deep Learning**. 2023. 38p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Na natureza, cerca de noventa por cento das plantas com flores dependem de polinizadores para transferir o pólen para realizar a reprodução sexual. Essas plantas são críticas no funcionamento dos ecossistemas, pois fornecem alimentos, formam habitats e fornecem outros recursos para muitas espécies animais, incluindo humanos. Por outro lado, as atividades humanas são responsáveis por mudanças nas redes ecossistêmicas que afetam negativamente o comportamento e a eficiência das abelhas polinizadoras. Consequências diretas e indiretas superam a compreensão humana e as ferramentas convencionais para abordar o funcionamento e as respostas de sistemas adaptativos complexos, portanto, novas abordagens e técnicas são necessárias. Como os polinizadores são uma parcela muito sensível de um ecossistema, questionou-se se algoritmos de Deep Learning se apresentam como uma alternativa para prever impactos de intervenções antrópicas sobre populações de insetos polinizadores. Nosso estudo avaliou a performance de cinco algoritmos (Linear Regression, Random Forests, Gradient Boosting Machines, Dense Neural Networks e Long Short-Term Memory) em prever a ocorrência de disrupção e anomalias que podem impactar abelhas e suas colméias. Os algoritmos de maior performance avaliados foram o LSTM e o GBM, ao avaliar o conjunto de dados que relaciona o uso de pesticidas neonicotinóides.

Palavras-chave: Polinizadores. Agentes de mudanças. Predição. Deep Learning.

CONTENTS

1 INTRODUCTION

Pollination is an important ecosystem function provided by a variety of arthropod species. The conservation of pollinator habitat can enhance overall biodiversity and the ecosystem services it provides (including pest population reduction), protect soil and water quality by mitigating runoff and protecting against soil erosion, and enhance rural aesthetics (WRATTEN *et al.*, 2012).

To understand the influence of global change on species survival and pollination, we need to understand the fact that all species are connected by ecological interactions (LIBRÁN-EMBID *et al.*, 2021). Plant–pollinator interactions, for example, are mutualistic associations fundamental to the reproductive success of 88% of all flowering plants and consequently to the functioning of natural and agricultural systems (LIBRÁN-EMBID *et al.*, 2021).

The loss of these pollinators is likely to have serious consequences for both general biodiversity and crop productivity (WRATTEN *et al.*, 2012)(Kevan and Phillips, 2001). The survival and development of honey bee colonies is influenced by the regularity, quality and quantity of nectar and pollen (WRATTEN *et al.*, 2012)

Drivers of change in these ecosystem networks are a major concern of today's scientific research, as they are responsible for decreasing biodiversity and the extinction of arthropod species in a habitat (SEIBOLD *et al.*, 2019).

However, assessing sources and outcomes is a difficult task because direct and indirect consequences in surmount human understanding and conventional tools in addressing complex adaptative systems functioning and responses. This brings relevance for novel approaches and techniques empowered by modelling tools, machine learning algorithms, remote sensing and big data(CAPINHA *et al.*, 2021).

In this context, the use of bioindicators like pollinators is a very common technique for assessing impacts and environmental quality since they are a very sensitive parcel of an ecosystem. Moreover, anthropogenic interventions like land-use intensification and habitat fragmentation are a common subject of research and source of data.

Insects, as stated by (GEROVICHEV *et al.*, 2021) are optimal subjects for ecoinformatic research due to their high abundance, wide distribution and key roles in ecosystem functions. They have crucial impacts on human well-being, both positive (pest control and agricultural pollination) and negative (crop damage and vectoring of disease).

Much research effort is therefore aimed at detecting changes in insect populations, and devising strategies to promote or mitigate these changes. Many important processes

in insect ecology occur over large scales in space (e.g., long-term migrations) or time (e.g., multi-year population cycles), and thus are difficult to study using standard experimental approaches (GEROVICHEV *et al.*, 2021).

The current leading approach to biological problems within Machine Learning is supervised learning using deep neural networks (DNNs), and particularly convolutional neural networks (CNNs), which are able to extract abstract high level features from images (GEROVICHEV *et al.*, 2021).

1.1 Research Questions

This paper evaluates the learned models for predicting the impact of anthropic action, determining whether they present an acceptable and reliable prediction, and determine which model has a better performance than others. In light of challenges and problems currently faced in predicting drivers of change impact in pollinators species, the following research questions were elaborated in order to guide this project:

Q1 “Do Deep Learning algorithms present themselves as an alternative to predict anthropic interventions impacts over insect pollinator populations?"

Q2 "Is it possible to evaluate the learned models for predicting the impact of anthropic action, determining whether they present an acceptable and reliable prediction, and determining which model has a better performance than others?"

Given these research questions, the following objectives for the development of this work are defined:

1. Map the learning algorithms used in the ecological spatio-temporal data analysis. Use the best learning algorithms to develop a simple application, able to classify the impacts and predict the probability of their occurrence. This objective is related to the research question Q1.

2. Analyze the gaps and performance of the available algorithms. The proposed application must attain a performance that is comparable to other prediction approaches to ecological events based on time series. This goal is related to the research question Q2.

2 THEORETICAL FRAME OF REFERENCE

2.1 The Importance of Pollinators

Interactions between plants and pollinators are mutualistic associations fundamental to the reproductive success of 88% of all flowering plants and consequently to the functioning of natural and agricultural systems (OLLERTON; WINFREE; TARRANT, 2011). Understanding the properties of their interactions and networks gives information about their functionality and stability, which ultimately determines species persistence (KLEIN *et al.*, 2007).

Plant-pollinator interaction networks may be particularly susceptible to anthropogenic changes, owing to their sensitivity to the phenology, behavior, physiology, and relative abundances of multiple species (TYLIANAKIS *et al.*, 2008).

Studies have shown (WRATTEN *et al.*, 2012) that habitat loss was the human activity most significantly detrimental to the abundance and diversity of bees, particularly in extremely disturbed landscapes. Also, a growing body of research has demonstrated that farms located in close proximity to natural areas can receive all of their pollination services from wild bees alone (WRATTEN *et al.*, 2012).

2.2 Drivers of Change

Pollination interactions are important as they benefit both biodiversity and humans. A great diversity of plants and animals mainly insects, but also some birds, lizards and mammals depend mutually on each other for pollination and food, and their interactions may influence population persistence (HEGLAND *et al.*, 2009).

As the drivers of change are responsible for affecting the interconnectivity of relationships among species, a metanetwork approach can be used to identify key traits of habitat fragments that are fundamental to maintain metacommunity functionality (LIBRÁN-EMBID *et al.*, 2021).

Moreover, in general, a small proportion of species are structurally important to a network, however, when these are lost, cascades of extinctions might occur, leading to a general collapse of the system (LIBRÁN-EMBID *et al.*, 2021).

Climate change may be cited as an important driver of change as temperature affect the availability of resources for species. However, whether climate warming will affect ecosystem functioning depends on how interactions among species are influenced. Alterations in trophic relationships and energy-flows in both predator-prey and plant-herbivore interactions as a consequence of rising temperatures (STENSETH; MYSTERUD,

2002).

The fragmentation of habitats is another key driver of change in species networks. According to (RIBAS *et al.*, 2005), several ecological processes may occur after a fragmentation event. These events are linked in a network of events that frequently lead to species loss, which, therefore, may determine species richness in each remnant.

One of the main causes of habitat fragmentation is the increasing land-use for agricultural intensification. Although it is unclear (SEIBOLD *et al.*, 2019) whether agricultural intensification is a direct driver of arthropod decline in abundance, it affects indirectly (e.g. fragmentation for agricultural purposes), the insect species that rely on wild resources provided in a spatio-temporal scale.

Along with the agricultural intensification, the increasing use of pesticides for crop protection is also related to a decline in pollinators abundance (WILLIAMS *et al.*, 2010), specially because these arthropods are very sensitive to disturbances, either due to a decline in other insect population or the use of non-specific pesticides.

2.3 Machine Learning Algorithms and Convolutional Neural Networks

Machine Learning Algorithms (ML) and Convolutional Neural Networks (CNNs) have been applied in a variety of purposes and researches. Regarding this work, the pattern recognition are effective examples of their application for time series classification (CAPINHA *et al.*, 2021).

In this context, learning can occur in two different manners: without supervision, where computers automatically discover patterns and similarities in unlabeled data; or, with supervised training, where a labelled dataset is first given to the computer, in order to train and associate the labels to the examples;

With CNNs, automated learning procedures are possible by decomposing the data into multiple layers, each with different levels of abstraction, that allow the algorithm to learn complex features representing the data (CHRISTIN; HERVET; LECOMTE, 2019).

A great deal of ML algorithms have been available for decades, and most notably neural networks. However, until recently, constraints of computational architecture and power have restricted their application, and especially for issues as data-intensive as climate change (BAUER, 2021). As it is a complex scientific and multi-faceted issue, amenable to ML and AI analysis.

CNNs are a family of multilayered neural networks constituting a class of deep, feed-forward artificial neural networks (ANNs) that have been successfully applied to computer vision applications (CHRISTIN; HERVET; LECOMTE, 2019). CNNs typically contain a number of common components, including convolution, pooling and fully connected layers, in different configurations that are connected successively to perform some complex-

learning tasks (YANG; XU, 2021). Therefore, Deep networks have the potential to model the influence of environmental variables on living species, even though they have not yet been applied in this way (CHRISTIN; HERVET; LECOMTE, 2019).

2.4 Making Predictions in Biology with Deep Learning

In ecology, time series classification is generally approached by processing the time series data into a set of summary variables and then using these variables as predictors in 'classical' supervised classification algorithms, such as logistic or multinomial regressions or random forests (CAPINHA *et al.*, 2021).

With traditional machine learning algorithms, feature extraction requires human supervision, whereas deep learning tools can learn by themselves very complex representations of data due to their multilayered nature (CHRISTIN; HERVET; LECOMTE, 2019). This feature is one of the main reasons that makes Deep Learning algorithms a better and easier approach for ecological spatio-temporal researches: performance.

Using general learning procedures, deep learning algorithms are able to automatically detect and extract features from data. However, these results depend on the existence of a sizeable labelled dataset that can be used to train the algorithms to extract the desired features from the data. Deep learning may be considered especially appropriate when analyzing large amounts of data, and it performs particularly well for complex tasks such as image classification or speech/sound recognition (CHRISTIN; HERVET; LECOMTE, 2019).

Deep learning and neural network approaches avoid specifying a process-based model, which makes it more data-led, improving the understanding of multivariate relationships in nonlinear systems (BAUER, 2021). These approaches allow classifying phenomena directly from raw time series data, a characteristic that requires ecologists to think more critically about the temporal component of the phenomena being classified (CAPINHA *et al.*, 2021). Algorithms used in the recent literature for plant pollinator interactions (PICHLER *et al.*, 2019), are assessing predictive and inferential performance of the models, creating a minimal simulation model.

The functioning and stability of ecosystems can then be monitored by converting all these species data and interactions into food web models and/or focusing on indicator species, which are very sensitive to habitat and climate changes (Mac Aodha *et al.*, 2018). Beyond that, the same sort of 'fully' temporally explicit approach can be exploited for virtually any ecological or biological entity or state, as long as the putative drivers have a temporal dimension (CAPINHA *et al.*, 2021).

2.5 Datasets Availability and Related Studies

Deep Learning algorithms, as stated before, rely on great amounts of data. Their availability is a recurrent issue in biological studies and prediction algorithms because time series classification is generally approached by processing the time series data into a set of summary variables (CAPINHA *et al.*, 2021), an approached known as feature-based.

However, some limitations still undermine their predictive performance and scalability (e.g. the need for domain-specific knowledge about the phenomenon that is being classified) (CAPINHA *et al.*, 2021). When considering the ever-growing body of knowledge in the ecological literature, few, if any, ecological phenomena are fully understood (CURRIE, 2019).

Recent publications in the literature have started to use Creative Commons Licenses to share datasets and codes, like (CAPINHA *et al.*, 2020), where the datasets represent classification tasks that are predominantly approached by ecologists through feature-based approaches. Beyond that, public datasets from Official Sources like the IPBES (POTTS,) and others, are key publications that may increase data availability along with other relevant options. Ecology scientists and researchers might also benefit from the upcoming Data Markets that use Blockchain for tokenizing Data Assets like the Ocean Market Protocol.

3 MATERIAL AND METHODS

In this study, I conducted two distinct experiments, each employing a unique set of variables. Initially, the Turkish Beekeeper's Monitoring System Dataset was analyzed, aiming to predict temperature and humidity anomalies. This analysis was intended to inform the approach to the subsequent evaluation of the HOneyBee Online Studies (HOBOS) dataset. In the examination of the HOBOS dataset, the focus was on assessing the predictability of the impact that 'Temperature' and 'Humidity' exert on the 'Flow' variable. Specifically, the goal was to identify patterns that would indicate increases or decreases in bee traffic, defined as the movement of bees into or out of a hive. This dual-experiment approach allowed for a comprehensive understanding of the environmental factors affecting bee behavior in apiaries.

The second experiment of the study focused on investigating the relationship between the use of neonicotinoid pesticides and honey production in the United States. This experiment was designed to evaluate the correlation among key factors: the number of bee colonies, yield per colony, total honey production, and the extent of neonicotinoid pesticide usage. By examining these variables, the aim was to gain insights and predict the impact of these pesticides on bee populations and honey production, thereby contributing valuable data to the ongoing discourse on the environmental effects of neonicotinoids.

3.1 Data Search and Collection

The present study employed a multi-faceted approach combining Machine Learning and Deep Learning algorithms to analyze and predict the impacts of environmental changes on pollinator bees and hives. Data were searched using Capes repository for scientific articles, Google Dataset Search, The Center for Plant Conservation, The Database of Pollinator Interactions and Kaggle.

The methodology is structured into distinct phases: initially, data collection focused on bees and their environmental drivers, followed by employing Machine Learning models for time series forecasting. Subsequently, Random Forests were utilized for predictions, with thorough data cleaning and normalization processes to ensure data integrity. The final phase involved rigorous training and comparative analysis of these models to ascertain their predictive accuracy.

3.2 Preprocessing

3.2.1 HOneyBee Online Studies - HOBOS

This data set, referred to as "HOBOS," contains records of beehive metrics, including temperature, humidity, and bee flow. The dataset was collected from the HOBOS Project, a network of beehive monitoring stations equipped with HOBOS sensors. These sensors continuously recorded environmental conditions and bee activity over a specified period.

The data set provided 50,724 rows and 5 columns of data. The columns include the following variables: timestamp, weight, temperature, humidity, and flow. There were 22,972 missing values in the columns weight and humidity, and 3,768 in the temperature column. The flow column has only one missing value. The column weight was withdrawn from the analysis as the purpose of the study was to understand the correlation between climate variables and flow. For the remaining columns, all missing values were imputed with the median for each column.

For this data set, an Exploratory Data Analysis (EDA) and a correlation analysis (fig. 1) was conducted to understand the strength and direction of the relationship between the variables. The analysis helped in identifying patterns, before conducting a predictive analytics, and data exploration.

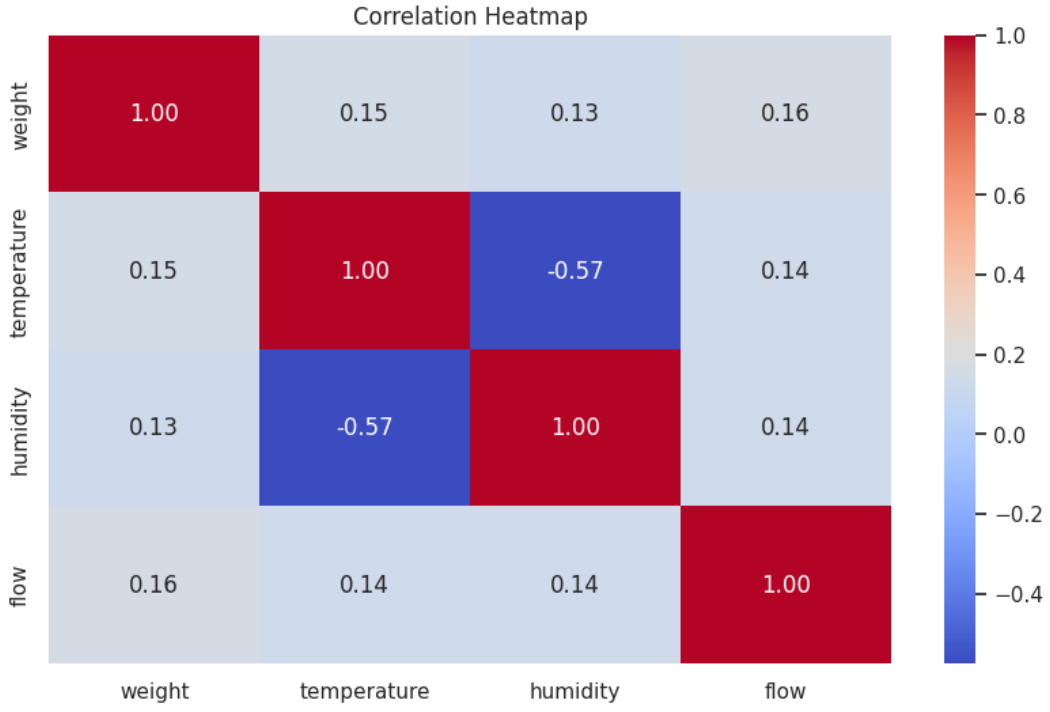

Figure 1 – HOBOS Correlation Heatmap.

3.2.1.1 Data Cleaning

The initial dataset underwent a data cleaning process to ensure data quality and consistency. This process involved the following steps:

1. Removal of the 'weight' column: The 'weight' column was removed as it was deemed irrelevant for the analysis.
2. Imputation of Missing Values: Missing values in the 'temperature,' 'humidity,' and 'flow' columns were imputed with the median value of each respective column.

3.2.1.2 Feature Engineering

Feature engineering was performed to create new variables that could potentially capture the influence of ideal temperature and humidity conditions on bee flow. Two new binary variables were introduced:

- 'temp_in_ideal_range': This binary variable indicates whether the temperature falls within the ideal range of 21 to 35 degrees Celsius.
- 'humidity_in_ideal_range': This binary variable indicates whether the humidity falls within the ideal range of 50% to 70%.

3.2.2 Turkish Beekeeper's Monitoring

This dataset was collected by a beekeeper in an apiary located in Çanakkale, Turkey. The dataset includes detailed records of various beehive parameters such as temperature ('T_Hive') and humidity ('RH_Hive'). This dataset was compiled from advanced monitoring systems installed in beehives.

The data set provided 12,147 rows and 15 columns. The dataset contains columns like Hour, DateTime, T_Hive (hive temperature), RH_Hive (hive relative humidity), AT_Hive (apparent hive temperature), Tamb (ambient temperature), RHamb (ambient relative humidity), ATamb (apparent ambient temperature), and several columns comparing temperature and apparent temperature differences between different hives and ambient conditions with no missing values. An EDA and correlation analysis (fig. 2) were conducted, evaluating the correlation between the variables and how each variable relate to each other.

3.2.2.1 Data Cleaning

The initial data cleaning steps for the dataset involved two key processes:

1. Removal of Extraneous Columns: Columns that were not relevant to the study's objectives were identified and removed. This step was crucial to focus the analysis on meaningful data and reduce computational complexity.

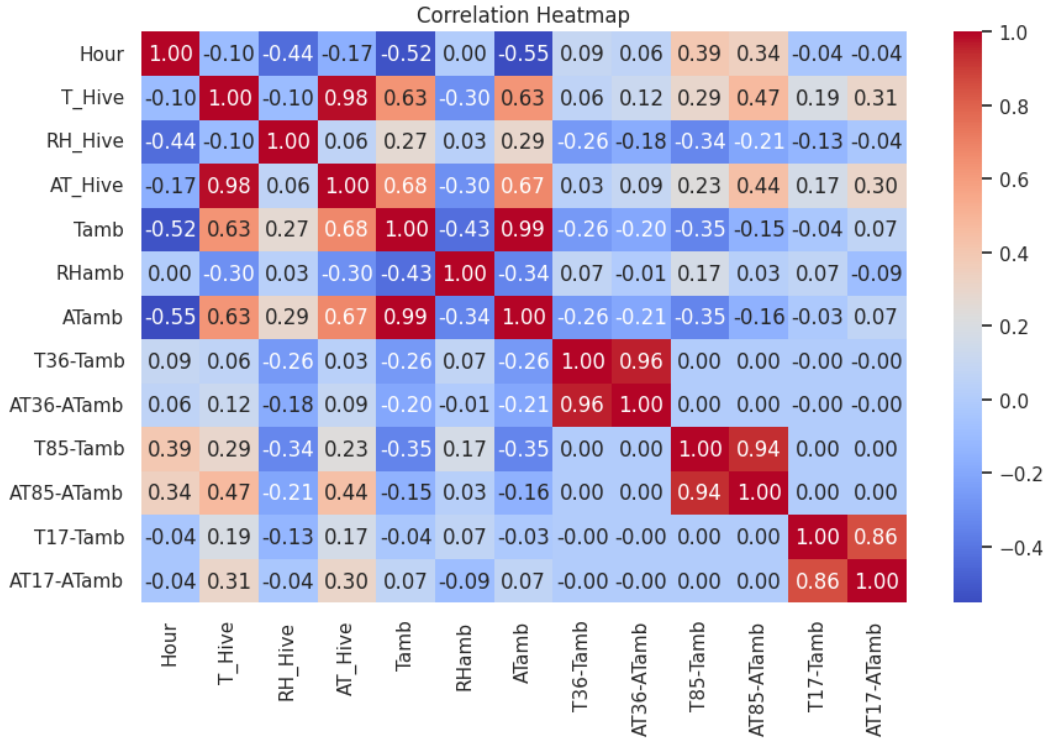

Figure 2 – Turkish Beehives Correlation Heatmap.

2. Imputation of Missing Values: For essential variables like temperature and humidity, which had missing values, a median imputation technique was employed. This involved replacing missing data with the median value of each respective column. The choice of median (over mean, for example) helped in mitigating the impact of potential outliers in the data, ensuring a more robust approach to handling missing values.

3.2.2.2 Feature Engineering

Feature engineering was conducted to establish binary indicators for temperature and humidity anomalies. These were based on specific threshold values, identifying significant deviations from normal patterns. Temporal attributes were extracted from the 'DateTime' column, facilitating the analysis based on time-series data.

3.2.3 HoneyBees and Neonic Pesticides

This dataset utilizes Honey Production in the USA, extended to the period 1998-2017. The dataset also includes data from USGS's Pesticide National Synthesis Project, which allows evaluation of the statistical connections between Honey Production and the use of Neonicotinoid (neonic) pesticides.

The data set provided 1,957 rows and 17 columns of data. The columns include state, numcol (number of colonies), yieldpercol, totalprod (total production), stocks, priceperlb

(price per pound), prodvalue (production value), year, StateName, Region, FIPS (Federal Information Processing Standards code), and various neonicotinoid pesticide measurements (nCLOTHIANIDIN, nIMIDACLOPRID, nTHIAMETHOXAM, nACETAMIPRID, nTHIACLOPRID, nAllNeonic).

There were 825 missing values in the 'FIPS' column and 301 missing values in the 'nCLOTHIANIDIN', 'nIMIDACLOPRID', 'nTHIAMETHOXAM', 'nACETAMIPRID', 'nTHIACLOPRID', 'nAllNeonic' columns each. The columns stocks, priceperlb, prodvalue, StateName, Region and FIPS were excluded from the analysis. The median values were imputed in the neonic columns with missing values and the correlation between the remain columns is highlighted in Figure 3.

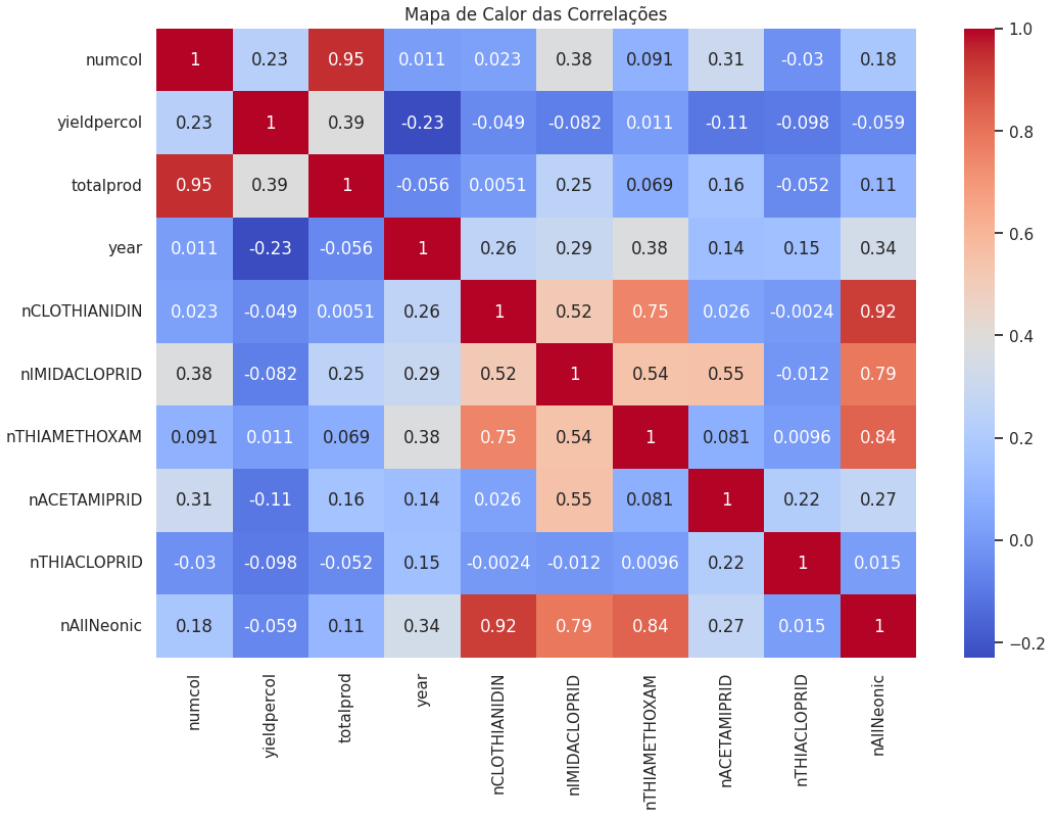

Figure 3 – Neonic Correlation Heatmap.

3.2.3.1 Data Cleaning and Feature Engineering

The dataset was cleansed of any irrelevant data. Columns not pertinent to the study, such as economic values and regional identifiers, were removed. Missing values within the dataset were addressed by imputing them with the median values of their respective columns.

To ensure consistency in data scale and range, a normalization process was applied to all numerical variables within the dataset. This step was crucial to prepare the data for effective analysis and modeling. A new feature representing the cumulative exposure to

various neonicotinoids was engineered. This feature was calculated by summing the quantities of different types of neonicotinoids present in the dataset, providing a comprehensive measure of neonicotinoid exposure.

In the analysis of the dataset, a critical aspect was to quantify the cumulative exposure to various neonicotinoids, which are a class of pesticides. This cumulative exposure feature aimed to provide a comprehensive measure of the total neonicotinoid exposure across different types of neonicotinoid pesticides present in the dataset.

The resulting cumulative exposure value provided a comprehensive measure of the total neonicotinoid exposure associated with each record. It effectively aggregated the contributions of different neonicotinoid compounds, reflecting the combined impact of various pesticides on the environment.

The feature engineering process involved the creation of a new feature that quantified the cumulative exposure to different neonicotinoid compounds within the dataset. This feature provides a comprehensive measure of neonicotinoid pesticide exposure, considering all identified neonicotinoid types in the dataset.

An extensive EDA was conducted to uncover underlying patterns, relationships, and distributions within the data. This step included examining variable correlations and distributions to better understand the dataset's characteristics.

3.3 Data Transformation

In the HOBOS dataset, temporal information extraction and transformation were meticulously conducted. The key steps included:

- **Timestamp Conversion**: The 'timestamp' column underwent conversion into a standard datetime format for accurate temporal analysis.
- **Date Feature Engineering**: Essential date features, namely the day ('day'), month ('month'), and year ('year'), were derived from the datetime format and represented as individual columns for enhanced granularity in analysis.
- **Original Timestamp Removal**: Post-extraction of relevant date features, the original 'timestamp' column was removed to streamline the dataset.

For the Neonics dataset, data normalization was executed. This involved:

- **Normalization of Numerical Variables**: All numerical variables were normalized to standardize data scales and ranges, a critical step for ensuring homogeneity in subsequent analytical and modeling processes.

The Turkish Beehives dataset underwent a similar temporal transformation:

- **'DateTime' Column Transformation**: The 'DateTime' column was processed to extract and structure vital temporal attributes. This included conversion into a structured format and segmentation into separate columns for day, month, and year, paralleling the approach used in the HOBOS dataset.

3.4 Modeling

3.4.1 Machine Learning

3.4.1.1 Linear Regression and Random Forests

In our study, Linear Regression and Random Forest algorithms were applied to the three distinct datasets. Specifically, for the HOBOS dataset, a Linear Regression model was developed to predict bee flow, utilizing environmental factors and date-related features. Conversely, the Random Forest Regressor was strategically employed to investigate the effects of neonicotinoids on hive health, primarily indicated by honey production metrics. Lastly, for the Turkish Beehives dataset, the goal was to predict the occurrence of out of range temperatures and humidity.

To ascertain the models' efficacy and reliability, a cross-validation method was employed. Additionally, the models' predictive capabilities were further evaluated using a separate test data sets with an 80% - 20% parametrisation. This approach ensured the robustness and generalizability of the findings, providing a comprehensive understanding of the models' performance in real-world scenarios.

3.4.2 Deep Learning

Two types of neural network models were developed: LSTM (Long Short-Term Memory) and DNN (Dense Neural Network). The LSTM model was designed to capture temporal dependencies in the time-series data, while the DNN focused on identifying complex patterns in the dataset. The LSTM model was structured with LSTM layers, dropout for regularization, and a dense output layer. The DNN model comprised multiple dense layers with dropout layers to prevent overfitting.

3.4.2.1 Model Training and Evaluation

Both models were compiled and trained on the dataset, with specific loss functions tailored to each model's architecture and objectives. The LSTM model focused on minimizing the mean squared error (MSE), suitable for continuous data prediction. In contrast, the DNN was evaluated on its classification accuracy.

3.5 Results

3.5.1 HOneyBee Online Studies - HOBOS

The results of the Linear Regression model, the Random Forest model (executed separately), and GBM were compared and analyzed to assess their effectiveness in predicting bee flow based on environmental factors and date features.

The performance of three ML Algorithms was evaluated: Linear Regression, Random Forest, and Gradient Boosting Machine (GBM). The assessment of the dataset using showed the Linear Regression model with a notably higher MSE of 394094.57 and a substantially lower R^2 of 0.0492, highlighting a weaker predictive performance. The Random Forest algorithm, however, demonstrated enhanced efficiency with an MSE of 235963.91 and an R^2 of 0.4307. GBM, with an MSE of 294627.85 and an R^2 of 0.2892, indicated a better fit than Linear Regression but was outperformed by Random Forest. This comparison underscores the varying effectiveness of these algorithms depending on the dataset characteristics.

Algorithm	**MSE**	**R^2**
Linear Regression	394094.5696341925	0.049192539343682395
Random Forest	235963.91034882396	0.430704547353817
GBM	294627.8503151977	0.2891697075226136

Table 1 – ML Algorithms Evaluation - MSE and R^2 - HOBOS

Two deep learning algorithms were also evaluated: a Deep Neural Network (DNN) and a Long Short-Term Memory network (LSTM). In the comparative analysis of LSTM and DNN models for predicting, the LSTM model exhibited a superior performance. The LSTM achieved a Mean Squared Error (MSE) of 23.7944, which was lower than the DNN's MSE of 28.5150, indicating a higher precision in predictions. Furthermore, the LSTM model demonstrated a more robust fit to the data with a Coefficient of Determination (R2) of 0.7749, compared to the DNN's R2 of 0.7303. This indicates that the LSTM model could explain approximately 77.49% of the variance in the dependent variable, surpassing the DNN's capability, which accounts for about 73.03% of the variance (Table 2). These findings suggest that LSTM models may provide enhanced predictability in similar scenarios where capturing long-range dependencies and sequential patterns in data is crucial.

Algorithm	**MSE**	**R^2**
DNN	28.514998018314134	0.7302561946626089
LSTM	23.7943699800498	0.7749119989451823

Table 2 – Deep Learning Algorithms Evaluation - MSE and R^2 - HOBOS

3.5.2 Turkish Beekeeper's Monitoring

For the Turkish Beehives' Dataset, the Linear Regression model exhibited a Mean Squared Error (MSE) of 25.274 and an R-squared (R^2) of 0.409, indicating moderate predictive accuracy. The Random Forest model resulted in a slightly higher MSE of 27.060 and a lower R^2 of 0.368, suggesting less predictive efficiency compared to Linear Regression. Conversely, the GBM demonstrated superior performance with the lowest MSE of 23.424 and the highest R^2 of 0.453. These results illustrate that while Linear Regression provides a reasonable baseline, GBM offers a more precise predictive capability in this context, outperforming both Linear Regression and Random Forest.

Algorithm	**MSE**	**R^2**
Linear Regression	25.274020998625446	0.40933201231207206
Random Forest	27.060282014467006	0.3675860946454724
GBM	23.42408968062116	0.45256594050362864

Table 3 – ML Algorithms Evaluation - MSE and R^2 - Turkish

The Deep Learning algorithms presented higher accuracy than the other datasets as the DNN achieved an MSE of 25.274 and an R^2 of 0.4093, indicating a moderate level of predictive accuracy. The LSTM, on the other hand, showed a slightly higher MSE of 27.060 and a lower R^2 of 0.3676. These results suggest that while both models have a fair predictive capability, the DNN slightly outperforms the LSTM in terms of both error minimization and variance explanation in this particular scenario.

Algorithm	**MSE**	**R^2**
DNN	25.274020998625446	0.40933201231207206
LSTM	27.060282014467006	0.3675860946454724

Table 4 – ML Algorithms Evaluation - MSE and R^2 - Turkish

3.5.3 HoneyBees and Neonic Pesticides

The assessment of a different dataset using the same algorithms yielded contrasting results. The Mean Squared Error (MSE) and R-squared (R^2) values for the Linear Regression exhibited an MSE of 0.001116621 and an R^2 of 0.942457, indicating a high degree of model fit. Random Forest, with an MSE of 7.78e-05 and an R^2 of 0.995989, showed superior predictive accuracy. GBM outperformed both with the lowest MSE of 5.43e-05 and the highest R^2 of 0.997199, suggesting exceptional model precision and predictive capability. These results demonstrate the effectiveness of ensemble methods like Random Forest and GBM in capturing complex patterns in data.

For the Neonics dataset models, The DNN achieved an MSE of approximately 6.71e-05 and an R^2 of 0.0492 (Table 6), indicating a low error but also a low proportion of

Algorithm	MSE	R^2
Linear Regression	0.0011166210745929578	0.9424572362099335
Random Forest	7.781883991855942e-05	0.9959897666041427
GBM	5.434662973006844e-05	0.9971993585393474

Table 5 – ML Algorithms Evaluation - MSE and R^2 - Neonics

variance explained in the dependent variable. In contrast, the LSTM showed significantly better performance with an MSE of 0.000285 and an impressive R^2 of 0.9872, demonstrating a high degree of predictive accuracy and the ability to capture complex patterns in the data. These results highlight the LSTM's superior capability in modeling sequential data compared to the DNN in this specific context.

Algorithm	MSE	R^2
DNN	6.712691060459702e-05	0.049192539343682395
LSTM	0.00028471731527383485	0.9872039415434115

Table 6 – Deep Learning Algorithms Evaluation - MSE and R^2 - Neonics

The LSTM and DNN models demonstrated high predictive accuracy, indicated by low MSE values and high accuracy percentages, respectively. This suggests a strong capability in forecasting anomalies in beehives.

3.6 Discussion

The present study applied advanced machine learning techniques to three different datasets to predict environmental anomalies in beehives and their impact on bees. The approach involved data preprocessing, cleaning and transformation, to ensure data quality. The LSTM and Dense Neural Network models demonstrated predictive capabilities, as evidenced by low MSE and high accuracy metrics. However, the possibility of overfitting warrants further investigation. This study highlights the potential of deep learning in ecological monitoring, emphasizing the need for robust validation and the exploration of model applicability in diverse ecological scenarios. Future work should focus on model refinement and testing the models' generalizability across various ecological settings.

The findings must be contextualized within the constraints of data availability and study design. Data Challenges: One of the primary challenges faced was the limited availability of comprehensive datasets. The existing data on pollinator behavior and environmental factors was fragmented and often lacked the granularity necessary for deep learning algorithms to achieve optimal accuracy. This limitation necessitated a reliance on interpolations and assumptions that may have introduced biases or oversimplifications into the models.

A Lack of Longer Time Series: Another significant challenge was the absence of long-term data. Pollinator behaviors and environmental changes are phenomena that unfold over extended periods, often spanning several years to decades. The lack of longer time series limited the ability to comprehensively model and predict the long-term impacts of environmental changes on pollinator efficiency and behavior. This temporal limitation restricts a deep understanding of more subtle or delayed effects that might only become apparent over extended periods.

Generalizability of Studies: Additionally, the generalizability of the findings is constrained due to the limitation on specific regions and bee species, which may not accurately represent global patterns. The ecological dynamics in different geographical areas can vary significantly, thereby affecting the applicability of results to other contexts. This limitation underscores the need for more generalized, globally inclusive studies to enhance the understanding of environmental impacts on pollinators at a broader scale.

Broader Studies and Variable Inclusion: Furthermore, the research highlights a notable gap in broader studies that simultaneously incorporate multiple variables within the same region. Most existing research tends to focus on isolated variables or specific aspects of pollinator behavior. This narrow focus limits the understanding of how multiple factors synergistically affect pollinators. Additionally, there is a distinct lack of global studies evaluating consistent metrics across different regions. Such studies are crucial for developing a holistic understanding of pollinators' responses to environmental changes on a global scale. Addressing these gaps would significantly enhance the ability to devise effective conservation strategies and predict future ecological scenarios with greater accuracy.

Despite these challenges, the study contributes valuable insights into the complex interplay between environmental factors and pollinator behavior. Future research should aim to address these limitations by securing more comprehensive datasets, extending the duration of observational studies, and focusing on a broader range of species and geographical areas. Such efforts will not only refine the understanding of these ecological dynamics but also enhance the predictive capabilities of deep learning models in ecological research.

4 CONCLUSIONS

In this study, machine and deep learning algorithms were utilized to unravel the intricate effects of environmental changes on the behavior and efficiency of pollinator bees. Despite confronting challenges like limited datasets and the lack of extensive time series, research offers significant insights into these vital ecological components. Our findings stress the imperative need for comprehensive, long-term, and globally inclusive research to effectively steward pollinator bees against the intensifying environmental alterations. Machine Learning and Deep Learning algorithms rely greatly in good data, as highlighted by the performance differences among the algorithms.

The results underscore the importance of selecting an appropriate algorithm based on the specific characteristics of the dataset and the objectives of the model, highlighting the efficacy of each algorithm in handling complex patterns in sequential data accordingly to each data collection and dataset framework. Moving forward, this study not only contributes to the academic understanding of pollinator dynamics but also for understanding how different variables and data structures require different algorithmic and learning approaches. Moreover, the expectations are that the study may serve as an incentive for policy formulation, long term monitoring and research of conservation strategies. Such efforts are vital for preserving the health of pollinator populations, which are indispensable for ecosystem sustainability and agricultural productivity.

REFERENCES

BAUER, P. The digital revolution of Earth-system science. **Nat. Comput. Sci.**, Springer Science and Business Media LLC, v. 1, n. 2, p. 104–113, feb 2021.

CAPINHA, C. *et al.* **Code & Data: Deep learning for time series classification in ecology.** Zenodo, 2020. CC and ACH were supported by Portuguese National Funds through Fundação para a Ciência e a Tecnologia (CC: CEECIND/02037/2017, UIDB/00295/2020 and UIDP/00295/2020; ACH: PTDC /SAU-PUB/30089/2017 and GHTM-UID/Multi/04413/2013). Available at: https://doi.org/10.5281/zenodo.4017750.

CAPINHA, C. *et al.* Deep learning for supervised classification of temporal data in ecology. **Ecological Informatics**, Elsevier, v. 61, p. 101252, mar 2021. ISSN 1574-9541.

CHRISTIN, S.; HERVET, É.; LECOMTE, N. Applications for deep learning in ecology. **Methods in Ecology and Evolution**, John Wiley & Sons, Ltd, v. 10, n. 10, p. 1632–1644, oct 2019. ISSN 2041-210X. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13256https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13256https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13256.

CURRIE, D. J. Where Newton might have taken ecology. **Global Ecology and Biogeography**, v. 28, n. 1, p. 18–27, 2019. ISSN 14668238.

GEROVICHEV, A. *et al.* High Throughput Data Acquisition and Deep Learning for Insect Ecoinformatics. **Frontiers in Ecology and Evolution**, Frontiers, v. 0, p. 309, may 2021. ISSN 2296-701X.

HEGLAND, S. J. *et al.* How does climate warming affect plant-pollinator interactions? **Ecology Letters**, v. 12, n. 2, p. 184–195, 2009. ISSN 1461023X.

KLEIN, A. M. *et al.* Importance of pollinators in changing landscapes for world crops. **Proceedings of the Royal Society B: Biological Sciences**, v. 274, n. 1608, p. 303–313, 2007. ISSN 14712970.

LIBRÁN-EMBID, F. *et al.* A plant–pollinator metanetwork along a habitat fragmentation gradient. **Ecology Letters**, John Wiley & Sons, Ltd, oct 2021. ISSN 1461-0248. Available at: https://onlinelibrary-wiley.ez26.periodicos.capes.gov.br/doi/full/10.1111/ele.13892https://onlinelibrary-wiley.ez26.periodicos.capes.gov.br/doi/abs/10.1111/ele.13892https://onlinelibrary-wiley.ez26.periodicos.capes.gov.br/doi/10.1111/ele.13892.

Mac Aodha, O. *et al.* Bat detective—Deep learning tools for bat acoustic signal detection. **PLoS Computational Biology**, v. 14, n. 3, p. 1–19, 2018. ISSN 15537358.

OLLERTON, J.; WINFREE, R.; TARRANT, S. How many flowering plants are pollinated by animals? **Oikos**, v. 120, n. 3, p. 321–326, 2011. ISSN 00301299.

PICHLER, M. *et al.* Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. **Methods in Ecology and Evolution**, British Ecological Society, v. 11, n. 2, p. 281–293, aug 2019. Available at: https://arxiv.org/abs/1908.09853v2.

POTTS, S. **The assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production.** Available at: https://www.ipbes.net/assessment-reports/pollinators.

RIBAS, C. R. *et al.* How large is large enough for insects? Forest fragmentation effects at three spatial scales. **Acta Oecologica**, Elsevier Masson, v. 27, n. 1, p. 31–41, jan 2005. ISSN 1146609X.

SEIBOLD, S. *et al.* Arthropod decline in grasslands and forests is associated with landscape-level drivers. **Nature 2019 574:7780**, Nature Publishing Group, v. 574, n. 7780, p. 671–674, oct 2019. ISSN 1476-4687. Available at: https://www.nature.com/articles/s41586-019-1684-3.

STENSETH, N. C.; MYSTERUD, A. Climate, changing phenology, and other life history traits: Nonlinearity and match-mismatch to the environment. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 21, p. 13379–13381, 2002. ISSN 00278424.

TYLIANAKIS, J. M. *et al.* Global change and species interactions in terrestrial ecosystems. **Ecology letters**, Wiley Online Library, v. 11, n. 12, p. 1351–1363, 2008.

WILLIAMS, N. M. *et al.* Ecological and life-history traits predict bee species responses to environmental disturbances. **Biological Conservation**, Elsevier Ltd, v. 143, n. 10, p. 2280–2291, 2010. ISSN 00063207. Available at: http://dx.doi.org/10.1016/j.biocon.2010.03.024.

WRATTEN, S. D. *et al.* Pollinator habitat enhancement: Benefits to other ecosystem services. **Agriculture, Ecosystems & Environment**, Elsevier, v. 159, p. 112–122, sep 2012. ISSN 0167-8809.

YANG, B.; XU, Y. Applications of deep-learning approaches in horticultural research: a review. **Horticulture Research 2021 8:1**, Nature Publishing Group, v. 8, n. 1, p. 1–31, jun 2021. ISSN 2052-7276. Available at: https://www.nature.com/articles/s41438-021-00560-9.