Bee Activity Prediction and Pattern Recognition in Environmental Data

Christine Preisach*[1] and Marius Herrmann[2]

[1]University of Applied Sciences Karlsruhe, Germany
[2]Karlsruhe Institute of Technology, Germany

Abstract

As a consequence of climate change, biodiversity is declining rapidly. Many species like insects, especially bees, suffer from changes in temperature and rainfall patterns. Applying machine learning for monitoring and predicting specie's health and life conditions can help understanding and improving biodiversity. In this work we use data collected from cameras and sensors mounted upon beehives together with different other data sources like weather data, information extracted from satellite images and geographical information. We aim at predicting bees' health (measured as their activity) and analyzing influencing environmental conditions. We show that we are able to accurately predict bees' activity and understand their life conditions by using machine learning algorithms and explainable AI. Understanding these conditions can help to make recommendations on good locations for beehives. This work illustrates the potential of applying machine learning on sensor, satellite and weather data for monitoring and predicting species' health and hence shows the ability for adaptation to climate change and a more accurate species monitoring.

1 Introduction

The environmental changes driven by climate change are disturbing natural habitats and species in ways that are still only becoming clear. There are strong signs that rising temperatures, changing rainfall pattern, droughts and extreme weather events are affecting biodiversity [14, 8]. For insects we know that, their number is reduced by two-thirds [15]. The threat posed by climate change to biodiversity is expected to increase, yet thriving ecosystems also have the capacity to help reduce the impacts of climate change [10]. Bees are part of biodiversity, the western honey bee is the most widespread managed pollinator globally, with that it contributes directly to food security, a third of the world's food production depends on bees, they are essential to people and the planet [4]. Bees' health is very much correlated to their environmental conditions. The

*christine.preisach@h-ka.de

Tackling Climate Change with Machine Learning workshop at ICLR 2024

pollen collection activity of bees is directly related to bees' health, since pollen is the only food bees eat. Environmental parameters like temperature, humidity or wind determine the pollen collection activity directly or indirectly, i.e., they either have an impact on the bees' health itself or on its food sources such as plants or rivers [5, 7]. It is therefore important to consider interactions within an ecosystem when analyzing bees' health and pollen collection activity.

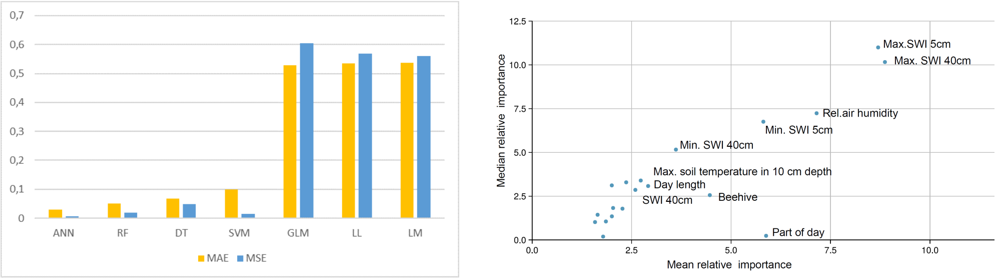
2 Proposal

We propose to analyse environmental factors and verify their influence on bee activity. We are interested in knowing which weather- and time-related conditions, as well as which land cover classes have an impact on bee activity. Measuring pollen activity and health on one side aids monitoring the environmental conditions [6], i.e., bees' health can help monitoring and predicting the environmental conditions and on the other side helps comprehending which conditions are important for the health of a species. This can aid to make recommendations on good locations for beehives and hence improve the protection of species. Our major aim in this work is to predict bee activity and by doing so analysing the influencing environmental factors. Therefore we applied supervised machine learning methods. Since the prediction target is a numeric variable - bee activity, we used regression algorithms. We conducted many experiments in order to analyse the aforementioned objectives, details are explained in the appendix.

3 Results and Discussion

In our experiments we used a 70%/30% split for training and testing our models. We evaluated the prediction quality and analysed the influence of input features on bee activity (see figure 1). For assessing the predictive power of our models, we calculated mean absolute error (MAE) and mean squared error (MSE). Figure 1a depicts *MAE* and *MSE* of the learned models applied to test data. The lowest MAE and MSE values were achieved by *ANN*, *RF*, *DT*, *SVM*. Hence, the relation between input features and bee activity seems not to be linear. Linear models (*GLM*, *LL*, *LM*) didn't perform well.

We are interested in understanding which weather- and time-related conditions show a positive or negative effect on bee activity, therefore we analyzed the importance of features on bee activity. Since not all algorithms used, are explainable and some of the explainable ones are less accurate, we decided to analyze the effect of input features on bee activity with shapley values. However, when comparing the absolute shapely values, they differ from model to model. Therefore it is not possible to decide on the size of an effect over all models. For this reason, we calculated the relative importance (influence) of a feature in a model by dividing the absolute mean shapley value of a feature by the sum of all absolute shapley values of a model. This enables the comparison of feature importance among different models. Figure 1b illustrates the mean and median



(a) Prediction quality

(b) Relative feature importance

Figure 1: Evaluation results

relative importance of features over all models. Only the features with a mean relative importance of more than 2.5% are labeled which their names. The most influential features are related to *soil water index (SWI)*, humidity in the air, *soil temperature*, *day length*, *part of the day* and the *beehive* itself. The beehive is an important input feature in our models, it seems that each bee colony is somewhat different in terms of its collection activity. We inspected the absolute shapley values in order to extract the direction of the effects, i.e., whether a feature has a positive or negative influence on bee activity. It is known that bees do not collect food before and during rainfall [7, 3], this is probably the reason for the negative effect of *relative air humidity* and the *soil humidity* (SWI at 5 cm depth) and *soil temperature* at 10 cm. While it rains the soil humidity in 5 cm depth increases and the soil temperature decreases. For lower soil layers (depth of 40 cm) we see a positive effect of *SWI* on bee activity, this might be because the water has seeped away after rainfall and helps plants to stay healthy [1]. *Part of day* has a negative effect too, since bees do only collect pollen during daytime. Day length has a positive impact, since bees have more time to collect pollen if there are more hours with day light. In addition to the most influential features shown in figure 1b, we found out that strong winds lead to less activity because of the additional effort bees have to put in flying and collecting pollen, this corresponds to the findings in [16, 7]. *Air pressure* and *sunshine hours* have a positive effect, high air pressure and many hours of sunshine indicate low rainfall and leads to more collection activity. Moreover, we saw that higher methane concentration has a negative effect, since higher methane concentrations are related to higher air temperatures (due to back reflection as well as absorption of sunlight and heat radiation) and therefore more cooling efforts in the beehive are needed by the bees, which means that bees are less active. It is surprising that some environmental factors like air temperature and precipitation play a smaller role in our models, which contradicts the findings in the literature. Our assumption is, that this is caused by *multicollinearity* in the data, we see highly correlated input features (see figure 3 in the appendix), this is natural since there are many interactions between various environmental parameters. Furthermore, we are interested to know which effect land cover classes like pastures, urban areas and others have on bee activity. In order analyze these effects, we had to conduct an additional experiment where we only used the land cover classes in a linear regression model. Reason for this is, that land cover class values (portion of each land cover class value and area) of a specific beehive do not vary over time, i.e., there is not much variance in the data, therefore machine learning models do not regard these features as important compared to others. We analyzed the coefficients of the linear regression model and the shapley values (see figure 4 in the appendix). We found significant positive effects of *mixed* and *broad-leaf forest*, as well as of *orchards*, weaker were the effects of *pastures*, *discontinued urban areas* and *water bodies.* We have observed strong negative effects for *water courses*, *shrubs*, *urban green.* When comparing the average activity of our ten beehives, we found least activity in locations with a large proportion of urban areas. During our project we faced two main challenges: One related to multicollinearity (highly correlated input features) which hampers the interpretation of the results. We addressed this issue by applying algorithms that are less sensitive to the problem (like RF and ANN) or use regularization. Moreover, we performed feature selection which removed some of the highly correlated features. The second challenge was around the availability of bee activity data, as mentioned in the data section, we had only observations for August and September, that means that data from spring and early summer was not available and hence we cannot draw conclusions for those seasons.

4 Conclusion

We showed that nonlinear machine learning algorithms are able to accurately predict bee activity and although most well performing algorithms are not explainable, we were still able to extract the impact of environmental conditions on be health (activity). We computed shapley values and the relative features importance. We found several environmental conditions that are important for bees' health. Monitoring bee activity and understanding these conditions can help to make recommendations on good locations for beehives. We found strong effects (negative and positive) of environmental conditions (weather-, time- and land cover class related features) on bee activity (especially variables around soil water index at different depths were important and not yet analysed), hence we conclude that it is important where beehives are being located in order to adapt to climate change and to keep them healthy. In future work we plan to conduct further experiments on additional data that includes recordings from spring and early summer, in order to make our findings more robust and generalisable.

References

[1] A. Amin, G. Zuecco, J. Geris, L. Schwendenmann, J. McDonnell, M. Borga, and D. Penna. Depth distribution of soil water sourced by plants at the

global scale: A new direct inference approach. *Ecohydrology*, 13:e2177, 03 2020.

[2] J. Cavanaugh and A. Neath. The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11, 03 2019.

[3] D. Clarke and D. Robert. Predictive modelling of honey bee foraging activity using local weather conditions. *Apidologie*, 49, 02 2018.

[4] IPBES. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Apr. 2022.

[5] K. Jiang, Z. Pan, F. Pan, J. Wang, G. Han, Y. Song, Z. Zhang, N. Huang, S. Ma, and X. Chen. Influence patterns of soil moisture change on surface-air temperature difference under different climatic background. *Science of The Total Environment*, 822:153607, 2022.

[6] S. Khalifa, E. Elshafiey, A. Shetaia, A. Ali, A. Algethami, S. Musharraf, M. Alajmi, C. Zhao, S. Masry, M. Abdel Daim, M. Halabi, G. Kai, Y. Al Naggar, M. Bishr, M. Diab, H. El-Seedi, and M. Pascaul. Overview of bee pollination and its economic value for crop production. *Insects*, 12:688, 07 2021.

[7] O. Komasilova, V. Komasilovs, A. Kviesis, and A. Zacepins. Modeling of the potential honey bee colony foraging activity based on the agrometeorological factors. *Baltic Journal of Modern Computing*, 9, 01 2021.

[8] T. Lovejoy, L. Hannah, and E. Wilson. *Biodiversity and Climate Change: Transforming the Biosphere*. Yale University Press, 2019.

[9] Q. A. Mendoza, L. Pordesimo, M. Neilsen, P. Armstrong, J. Campbell, and P. T. Mendoza. Application of machine learning for insect monitoring in grain facilities. *AI*, 4:348–360, 03 2023.

[10] T. Newbold. Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proceedings of the Royal Society B: Biological Sciences*, 285:20180792, 06 2018.

[11] R. Rodríguez-Pérez and J. Bajorath. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34, 10 2020.

[12] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[13] M. Tschaikner, D. Brandt, H. Schmidt, F. Biessmann, T. Chiaburu, I. Schrimpf, T. Schrimpf, A. Stadel, F. Haußer, and I. Beckers. Multisensor data fusion for automatized insect monitoring (kinsecta). page 1, 10 2023.

[14] M. C. Urban, G. Bocedi, A. P. Hendry, J.-B. Mihoub, G. Pe'er, A. Singer, J. R. Bridle, L. G. Crozier, L. D. Meester, W. Godsoe, A. Gonzalez, J. J. Hellmann, R. D. Holt, A. Huth, K. Johst, C. B. Krug, P. W. Leadley, S. C. F. Palmer, J. H. Pantel, A. Schmitz, P. A. Zollner, and J. M. J. Travis. Improving the forecast for biodiversity under climate change. *Science*, 353(6304):aad8466, 2016.

[15] R. Warren, J. Price, E. Graham, N. Forstenhäusler, and J. Vanderwal. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°c rather than 2°c. *Science*, 360:791–795, 05 2018.

[16] H. Zhao, G. Li, D. Guo, H. Li, Q. Liu, B. Xu, and X. Guo. Response mechanisms to heat stress in bees. *Apidologie*, 52(2):388–399, Apr. 2021.

A Data

In general monitoring species and quantifying health of a specific species is still difficult. Nowadays different camera systems for insect detection have been developed [9, 13]. To determine bees' health we are extracting the pollen collection activity (activity for short) of bees from videos captured with intelligent camera- and sensor-technology. The number of bees flying into the beehive and the amount of pollen collected by an individual bee is recognized using AI. The activity of a beehive at a specific point in time (every five minutes during the day) is calculated as the ratio of pollen input (amount of collected pollen) and the number of bees flying into the beehive. In this study we are using data collected during August and September 2021 from ten beehives in five different locations in Germany. Additionally, different other data sources are considered in our study, like:

- Weather data[1] including information on dew point, humidity, precipitation, air pressure, ground temperature, sunshine duration, air temperature and wind).
- Data on greenhouse gases at different altitudes (CO,CO_2,N_2O, CH_4)[2].
- Information on land cover classes (CLC)[3], they represent spatial information on different types (classes) of physical coverage of the Earth's surface, e.g. forests, grasslands, croplands, lakes, wetlands (we considered those land cover classed which are 3 kms around the beehives).

[1] https://www.dwd.de/

[2] https://www.icos-cp.eu/

[3] https://land.copernicus.eu/content/corine-land-cover-nomenclature-guidelines/html/

- Information on soil properties like soil temperature and humidity. We used the soil water index (SWI) which is determined by soil temperature and precipitation (at eight different depths - from 2 to 100 cm depth, near the beehives). SWI is extracted from satellite data (from the Copernicus project[4]).

Figure 2 shows the five locations (each beehive is depicted as a black dot) and the surrounding area in terms of the land cover classes, we see that some locations are mainly surrounded by grasslands and broad-leaf forest and others are located in urban areas or are placed near a water body. We computed the area and the proportion of each land cover class surrounding the beehive. We included time related information by extracting features like day length and part of day (we divided the day into eight 3-hours segments). In order to integrate all data sources, we had to aggregate the data from the original granularity of five minutes to three hours (we computed the minimum, maximum and mean of each numeric input variable). After combining all data sources, we had 130 input features and around 5,000 observations (an observation is identified by the beehive, part of the day and date).

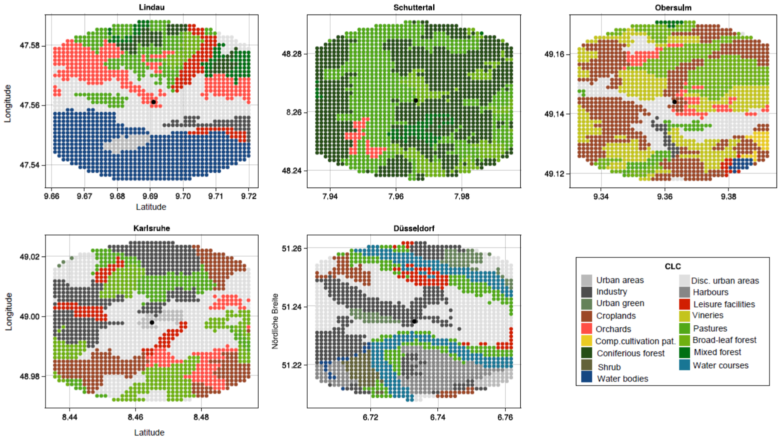


Figure 2: Land cover classes around the beehives

B Exploratory Analysis and Methods

We performed a correlation analysis (using Pearson-Bravais correlation) on different feature groups and activity, e.g., we computed the correlation of weather features and activity. Figure 3 shows several weather features that are highly correlated to the target variable but also to themselves. Other features e.g., the SWI features show high correlation to bee activity too. We applied filter-

[4] Copernicus Project https://www.copernicus.eu/

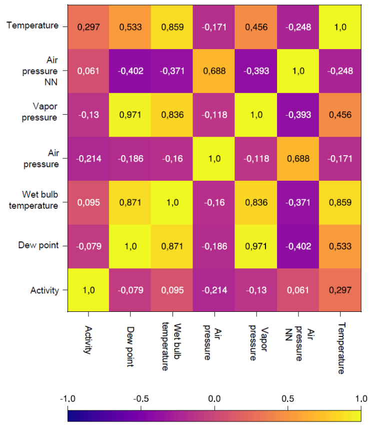


Figure 3: Correlation of weather features and activity

based feature selection using the Akaike Information Criterion (AIC) [2] and used the resulting 55 features for learning. Since we aimed at understanding the impact of the environment onto bee activity, we first learned several explainable regression models like linear regression (LR), generalized linear model (GLM), lasso lars, i.e., a lasso model with least angle regression (LL), decision trees (DT) and random forest (RF). We analyzed the coefficients of the linear models (LR, GLM, LL) as well as the rules learned by the decision tree algorithms and the feature importance of the random forest model. But we assumed

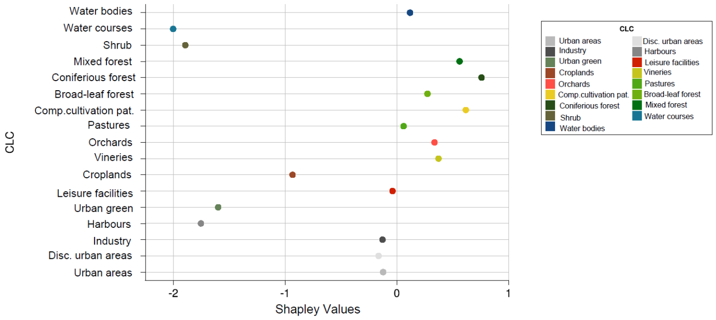


Figure 4: Absolute shapley vaues of land cover classes

that some other so called black box algorithms, which are not easily usable for explaining the impact of features on predictions, might perform better in terms of prediction accuracy, hence we used algorithms like support vector machines for regression (SVM) and artificial neural networks (ANN). Since these models are not interpretable [11] and different model types are hardly comparable in terms of their explainable effects, we additionally computed so called shapley values [12]. Shapley values, a method from coalitional game theory, they help to explain a prediction by assuming that each feature value of an instance is a “player” in a game where the prediction is the payout. Or in other words, a shapley value corresponds to the mean marginal contribution of each feature value across all possible values in the feature space i.e., they describe how much features contribute to the prediction. Features with positive shapely values, positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.