Identifying Effects of Air Pollutants in California Bee Population: A Machine Learning Case Study

Julia Roskam

Master of Science in Data Science Program, University of Wisconsin – La Crosse

DS 785 – Capstone

Dr. Tracy Bibelnieks

December 8, 2024

Abstract

Colony collapse disorder (CCD) is a phenomenon affecting honeybee populations globally, causing considerable year-over-year losses of bee colonies and threatening the United States agricultural industry due to the important work honeybees do as pollinators. While the causes of CCD are not yet known definitively, comprehensive machine learning models of honeybee population considering all known stressors and models considering individual stressors like pesticide use and mites exist, yet there is a lack of research that applies machine learning techniques to the relationship between bee population and poor air quality as an effect of climate change and pollution. This project aimed to examine this relationship through evaluating the performance of various machine learning models including linear regression, gradient boosting, k-nearest neighbors (k-NN), and support vector machines. This study utilized air quality data collected from the United States Environmental Protection Agency open-access database for pollutants including CO, Pb, NO_2, ozone, PM_{10}, $PM_{2.5}$, and SO_2 as well as honeybee population data collected every five years from 2002-2022 from the United States Department of Agriculture for all counties in California. The findings indicated that the k-NN model was the most successful model in predicting honeybee population using CO, NO_2, ozone, PM_{10}, $PM_{2.5,}$ year, and county population density as variables. The most important factors in predicting bee population in the selected k-NN model were ozone and PM_{10} concentrations. This project includes a discussion of the implications of these findings including recommendations for managing bee population through policy change and recommendations for future research opportunities.

Keywords: colony collapse disorder, machine learning, California, air quality, honeybee population modeling, linear regression, k-Nearest neighbors, support vector machines, XGBoost

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Project Background

The western honeybee (Apis mellifera) plays a critical role as a pollinator for the diverse array of crops produced in states with thriving agricultural sectors, like California. This role contributes an estimated $18 billion to the United States economy annually (Mayack et al., 2023). Unfortunately, due to a phenomenon now known as colony collapse disorder (CCD), the United States has faced decades of bee colony losses at around a 30% loss a year, a considerable increase from the previous average loss of 10-15% a year since the 1950s (Torres et al., 2015).

There remains to be seen a clear cause for CCD and researchers have determined that bee population decline likely results from several factors affecting honeybees simultaneously. Numerous species of bacteria, fungi, mites and other parasites, and viruses have all been identified as factors in colony collapse disorder alongside environmental factors like pesticides, climate change, and pollution (Watanabe, 2008).

Pollinators like honeybees play a crucial role in growing many of the world's crops and the decline in bee populations raises concerns about food security in the United States. Emerging research on the link between pollinator limitation and crop production shows that while, "global reliance on pollinator-dependent crops has increased over the past several decades," evidence suggests pollinator limitation in crop yields may be an increasing issue worldwide (Reilly et al., 2020). It is difficult to determine the exact proportion of pollination occurring from wild bees versus farmed honeybees, yet, some crops, like the almonds grown in California, are found to be pollinated almost exclusively by honeybees (Reilly et al., 2020). This makes the need for accurate models to predict honeybee population even greater.

Problem Statement

Due to the multifaceted nature of CCD, constructing a model that considers all possible stressors to the bee population is challenging. Factors like disease, mites, and pesticide use were identified early on as significant stressors and since, there have been numerous studies with a machine learning approach to examining how each of these single factors affect population decline (Torres et al., 2015). In the nearly 20 years since colony collapse disorder was first identified, there has been increasing evidence that the effects of climate change and poor air quality have contributed, in part, to the global decline in bee population. Yet, there is a lack of research that applies machine learning techniques to model this relationship (Mayack et al., 2023).

While honeybees have previously been used as indicators of environmental pollution, as Mayack et al. (2023) noted, “little attention has been given to the detrimental impacts of pollution on bees themselves.” The specific goal of this paper was to examine how climate conditions and presence of air pollutants impact honeybee population decline. Using the presence of specific air pollutants as variables, this project aimed to develop predictive models to assess future bee population trends across California.

California is home to more than 68,000 farms and ranches and had an agricultural revenue of nearly $56 billion in 2022, according to the California Agricultural

Statistics Review 2022-2023 (2024). The impact California has on the beekeeping industry is two-fold as California farmers both require honeybees to be brought from beekeepers around the United States for pollinating California crops and they maintain queen honeybees and colonies throughout the year for sale to other regions in warmer months (Mayack et al., 2023). This means that a substantial portion of the United States honeybee population is affected by the air quality in California and by focusing this project on California, larger conclusions can be drawn about the honeybee population nationally.

Project Purpose & Significance

This project aimed to further examine air quality as a factor in colony collapse disorder and determine a strategy for using air pollutant data to build predictive models that estimate honeybee population. The results of this project will be used to inform future environmental policy decisions and provide actionable insight and recommendations for honeybee population management in states with high agricultural output.

Project Objectives

Due to the high quantity and diversity of California counties, the first objective of this project was to determine a way to segment the state based on human population. California counties were grouped based on high, medium, or low population density. High population density areas were considered to be greater than 500 persons per square mile, medium was between 499 and 100 persons per square mile, and low was 99 or less persons per square mile.

The second objective was to determine which specific air pollutant or class of pollutants was most greatly linked to bee population decline in California. The pollutants assessed were those associated with combustion of fossil fuels, road transportation, factory and power plant pollution, and natural factors like fires and windblown dust. The pollutants considered as variables for honeybee population decline were particulate matter of various sizes (PM), ozone, sulfur dioxide (SO_2), lead (Pb), carbon monoxide (CO), and nitrogen dioxide (NO_2). The specific pollutant focused approach was expected to provide further insight into successful environmental policy pathways for bee conservation efforts in California.

The third objective was to compare several machine learning techniques to develop a predictive model for bee population in California with the highest accuracy possible. The machine learning methods used were linear regression, gradient boosting, k-nearest neighbors (k-NN), and support vector machines (SVM).

Assumptions & Limitations

A more detailed approach to dividing California counties into groups based on population density could be considered in future projects that includes a larger number of population density groups or even analysis on a county-by-county basis in California. Segmenting counties based on population density rather than population alone was based on findings that link denser urban populations to both higher air pollution levels and higher exposure in an effort to add another potentially explanatory variable to the final machine learning model (Carozzi and Roth, 2023). Individual models for each population density group could be explored in future work as well, however dataset size was a limiting factor in this project.

Conclusion

The intent of chapter 1 was to provide an introductory understanding of the problem surrounding colony collapse disorder and bee population decline, including the greater impacts to crop production and food security in the United States. Also covered were the project purpose and project objectives. The goal of this project was, in short, to examine the role of air quality in bee population decline by determining whether accurate models can be built using machine learning to predict bee population based on presence of pollutants in the air. The results of this project provided insight into this relationship and were used to form actionable recommendations for conservation of honeybee populations in California.

Project Organization

The remainder of the paper will be structured as follows:

Chapter 2: Literature Review. This chapter reviews the current research on factors influencing honeybee population decline, the causes and effects of common air pollutants, and previous efforts using machine learning for population modeling.

Chapter 3: Methodology. This chapter focuses on data collection, cleaning, and preprocessing as well as modeling methods and model selection.

Chapter 4: Results. This chapter reviews model results and evaluates model accuracy. This chapter also comments on the importance of each variable to the final model results and discusses project objectives.

Chapter 5: Discussion. This chapter gives a summary of the findings and conclusions of this project and provides recommendations for future research and future management of honeybee populations.

Chapter 2: Literature Review

Since the term colony collapse disorder (CCD) was coined in 2006 and researchers renewed their focus on pollinator population decline (which had already been occurring regionally since the early 1950s) several new factors have been identified. Globally, the pollinator with the largest contributions to agricultural output and the most widely studied is the honeybee (*Apis mellifera*). While honeybee populations have increased globally across the last few decades, there have been significant declines in population in certain regions in the United States and Europe and a decline in beekeeping as an industry. The most significant potential drivers affecting honeybee population decline and the decline in pollinator diversity are habitat loss, pesticide use and environmental pollution, decreased availability of resources, climate change, and pests and pathogens. (Potts et al., 2010).

Environmental Pollution, Climate Change, and Air Quality

Climate change has been identified as one of the leading drivers of colony collapse disorder and decreasing diversity in bee populations. One of the physical impacts of climate change and global warming on air quality is through climate-driven meteorological changes. Changes in temperature and precipitation due to climate change have downstream effects on the formation of secondary pollutants and lead to "increases in the severity and frequency of pollution periods" due to more stagnant air conditions, according to Zhu et al. (2019).

Studies on air quality and climate change in California have shown that increased temperatures in southern California, due to climate change, can lead to increased concentrations of ozone and $PM_{2.5}$. There is also evidence to support the claim that increases in surface temperature in California due to climate change correlate to increased vehicle emissions and higher concentrations of pollutants like CO, NO_2, SO_2, and particulate matter in the atmosphere (Zhu et el., 2019).

Particulate matter is a type of pollutant that can consist of many different chemicals and is classified instead by particle size. In humans, exposure to fine particulate matter is associated with premature death and cardiovascular disease (Jerrett et al., 2013). In honeybees, the effects of airborne particulate matter exposure are not as well documented. With regular foraging activity, bees are encountering and collecting particulate matter daily, and studies show that deposits of these particles are found on the heads, wings, and hind legs of bees in areas with high particulate matter pollution (Pellecchia & Negri, 2018). Forager bees also ingest particulate matter and other pollutants like ozone that can later be found in the gut and other organs of the bees. Along with ingestion, the pollutants that accumulate on forager bees are then brought back to the hive where they may eventually end up in the wax and honey produced by the hive, having long term effects on the health of the entire colony (Negri et al., 2015).

Similar to particulate matter, ozone and other concerning pollutants like those that enter the atmosphere via vehicle emissions NO_2, SO_2, and CO, to name a few, are becoming increasingly concerning as a threat to the honeybee population. Recent research done by Feldhaar & Otti (2020) has established the interaction between pollutants and social insects like honeybees. Their study finds negative effects like increased susceptibility to disease and a weakened immune system, jeopardizing honeybee survival and contributing to population decline (Feldhaar & Otti, 2020). While the link between pollution and honeybee population decline has been established, there is still a lack of research on how air pollution can be used to model population in bees that this paper seeks to address.

Challenges in Traditional Population Forecasting Systems

The majority of research surrounding honeybee modeling is done at the colony level rather than population level. As explained by Russell et al. (2013), honeybee population is influenced in two ways, "the number of colonies in the population and the number of bees in a colony." In a comprehensive review of honeybee models by Becher et al. (2013), they analyzed models of single colonies that were "potentially relevant for understanding honeybee colony survival and death" and found that most models fell into three categories. The categories included models that address "within-hive colony dynamics," foraging models that address honeybee interaction with the landscape around them, and models that address the effects of mites on honeybees. They also searched for models that focused on known stressors like pesticides use or other pathogens and pests (Becher et al., 2013).

Colony-level models seek to address the process of colony failure by attempting to use the social dynamics of the hives and populations of bees within the hives stratified by age, gender, position (working inside or outside the hive), or health to predict at what point a hive will fail (Khoury et al., 2011). While this work is important, part of the reason researchers have neglected to build larger scale or population level models is due to the lack of data gathered at the national or state level. Colony-level data such as health metrics or reproduction rates are more readily available than consistent larger scale or national datasets.

Machine Learning Methods in Honeybee Population Modeling

A systematic review of existing models of honeybee population decline done by Becher et al. (2013) found that machine learning models can effectively model honeybee behavior and population dynamics. Some notable early population models at the colony level include BEEPOP, HoPoMo (Honey Population Model), and BEEHAVE.

BEEPOP was the first age-structured model, using population dynamics of an individual colony to build a model that would simulate the various life cycle stages within a colony. The aim of the BEEPOP model was to use mechanistic modeling to create a model colony that accurately represented a real-life colony and then further develop models using predictive analytics to determine changes to the colony based on future influences and stressors (DeGrandi-Hoffman et al., 1989).

HoPoMo is a model that was later developed and remains one of the most complex honeybee population models. HoPoMo uses 60 equations to track the day-to-day life of a bee from adolescence to the start of adulthood. This model uses machine learning techniques to optimize parameters and perform sensitivity analysis. Part of the framework of the HoPoMo model includes different regression techniques and neural networks (Schmickl & Crailsheim, 2007). Later, HoPoMo went on to inspire the development of similar models that continued to track day-to-day honeybee life into adulthood and incorporate the use of differential equations formulas (Torres et al., 2015).

Becher et al. (2013) constructed an extensive model, BEEHAVE, that was another age-structured model incorporating colony dynamics and focusing on the effects of mites and viruses on honeybees as these were some of the earliest identified causes of CCD. The BEEHAVE model uses regression models, artificial neural networks, and Bayesian optimization as well to create a model that simulates the behavior of individual bees within a colony (Becher et al., 2013).

More recent models are incorporating deep learning and AI to predict health and colony sizes in regional honeybee populations. Researchers like Pinto et al. (2024) use a “recurrent neural network with long short-term memory” to predict pesticide presence in bees using daily flight activity to estimate health of honeybee populations. While others, like Apis-Prime, a model developed by Anwar et al. (2023) use a hybrid deep learning model and time series forecasting to estimate the daily weight of honeybee hives, allowing them to draw conclusions about the size, health, strength, and productivity of a bee colony.

Gaps in Current Research

While there have been many exciting advancements recently using machine learning and other advanced data science techniques to model and predict honeybee population, there are still gaps in the current research. Many of the current models that incorporate pollutants as parameters in building their models are focused on chemicals found in pesticides and overlook air pollutants that are due to climate change, vehicle emissions, and industrial emissions. There is also a notable lack in population level studies that seek to build models to estimate population on a large scale and not at the individual or colony level.

Conclusion

The literature demonstrates the effectiveness of machine learning and other advanced data science techniques in honeybee population modeling and supports the need for predictive models that address the effects of air pollution on the honeybee population. To address the previously identified gap in the literature, this project sought to determine whether accurate models can be built using California as a case study and data that tracks the concentration of pollutants in the atmosphere as well as yearly honeybee population.

Chapter 3: Methodology

The primary goal of this project was to determine whether machine learning models could accurately predict honeybee populations using air pollution data in California and to compare the efficacy of various modeling techniques. This chapter will outline the data collection and preparation methods used in preprocessing the data as well as feature engineering methods and model selection. This chapter also discusses what modeling techniques were explored, what techniques were chosen, and why those machine learning models were the best choice for this data.

Data Collection

The air quality data collected for this analysis was collected from the United States Environmental Protection Agency (EPA) open-access database. The data was collected daily from monitors in California and included separate datasets for each pollutant including CO, Pb, NO_2, ozone, PM_{10}, $PM_{2.5}$, and SO_2. CO and ozone were measured in parts per million (ppm) as daily max 8-hour concentration, NO_2 and SO_2 in parts per billion (ppb) as daily max 1-hour concentration, and finally, $PM_{2.5}$, PM_{10}, and Pb were measured as daily mean concentration in ug/m^3 LC, ug/m^3 SC, and ug/m^3 SC respectively.

The honeybee population data was collected from the United States Department of Agriculture National Agricultural Statistics Service, an open-access database that provides honeybee population data collected every five years by county in the state of California. The bee population data used for analysis was collected from 2002-2022 at five-year intervals and the air quality data used was collected from corresponding years.

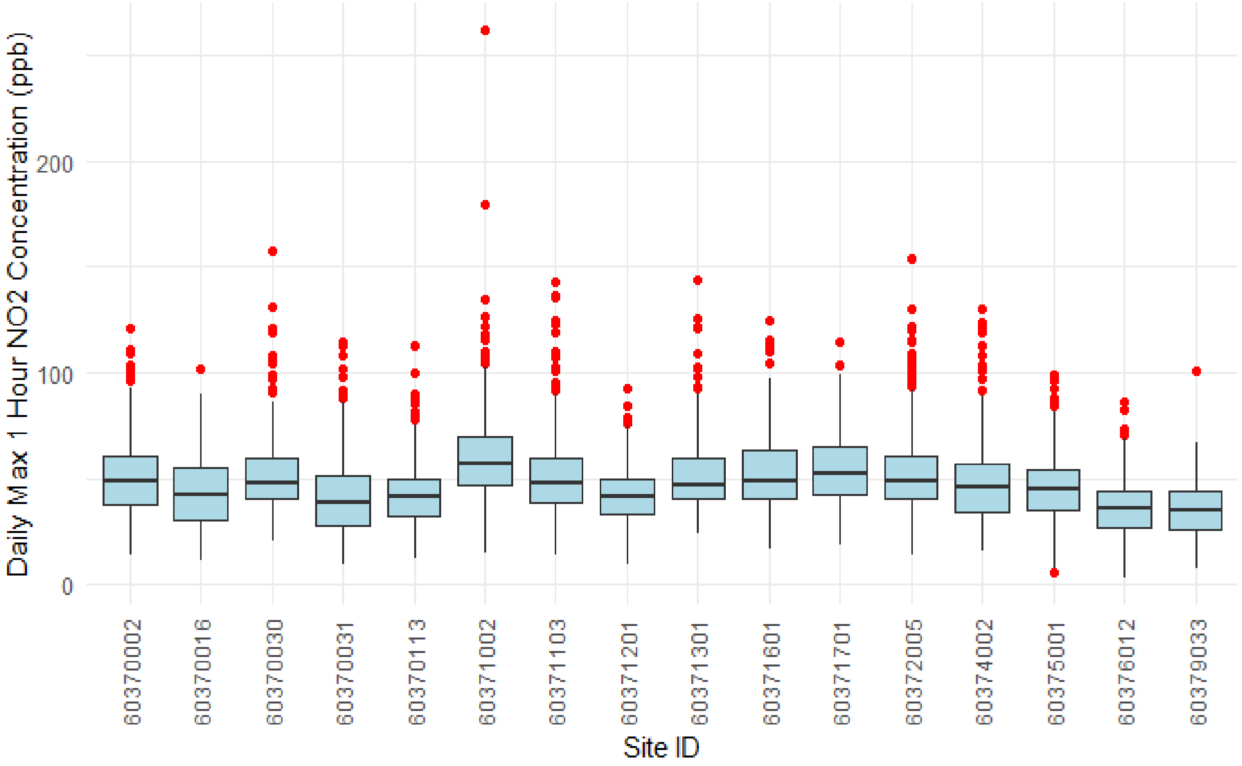
The datasets for each air pollutant were combined into five datasets, one for each year corresponding to the years bee population data was collected, containing all seven pollutants for which data was collected that year, grouped by county. Each county had between one and sixteen sites that collected air quality data for each pollutant and the data for each site was aggregated using the mean daily value of all sites within a county to represent said county. Mean was chosen as the best method for aggregating daily air quality data as the aim was to determine the average conditions of the county and to mitigate the influence of outliers or anomalies specific to certain sites within a county.

Initially, daily concentrations for each pollutant and daily air quality index (AQI) values corresponding to each pollutant were included, however daily AQI values were later removed from the dataset for each pollutant. This was done in an effort to reduce multicollinearity as each daily AQI variable showed high correlation values to the corresponding daily concentration variable for each pollutant.

Figure 1 shows a boxplot of daily max 1-hour NO_2 concentration measured in ppb by site in Los Angeles County in 2002. The median values for each site were similar across the county and each site had a similar range and distribution of values across the year. While aggregating this data to the county level meant possibly losing some spatial variation, the intention of this paper was to focus on county and state level effects. Figure 1 serves as an example of the considerations that were taken across all pollutants, counties, and years during the data collection and preprocessing steps.

Figure 1

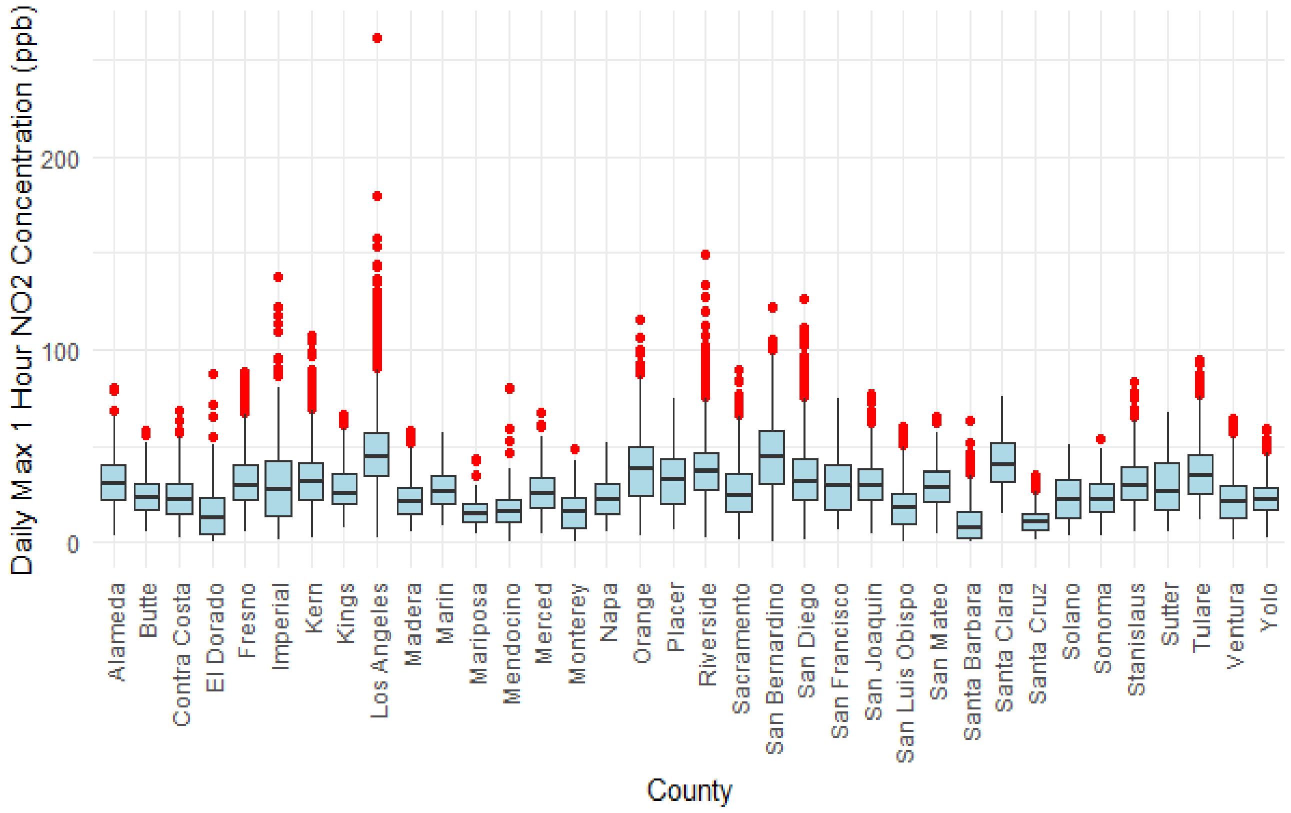
Daily Max 1-Hour NO_2 Concentration by Site in Los Angeles County in 2002



For reference, the EPA determined a standard level of NO_2 for 1 hour of exposure to be less than 100 ppb and considers the annual average NO_2 standard to be 53 ppb (“Air Quality Guide for Nitrogen Dioxide,” 2011). Figure 2 shows the distribution of daily max 1-hour NO_2 concentration by county in California, again using data from 2002 as an example. Figure 2 shows that the median value for each county is below the annual average NO_2 and nearly all the data is well below the standard level of 100 ppb with the exceptions of a handful of outliers across the state and year.

Figure 2

Daily Max 1-Hour NO_2 Concentration by California County in 2002



Using the data aggregated to the county level, daily air quality values were then aggregated using the annual mean to reflect the average conditions experienced by honeybee colonies in each county across the year. This final aggregation was necessary to ensure that our predictor variables (air quality values) and target variable (honeybee population) were on the same time scale before assessing various machine learning model approaches to ensure valid predictions would be made.

Data Preparation

After binding the individual air quality datasets with the bee population data by both county and year and aggregating the daily air quality data as necessary, several important steps still remained in the data preparation process. Many of the sites collecting air quality data had a significant number of missing values for Pb and SO_2 concentrations so those columns were removed from the dataset along with any other rows with missing values. Due to the small size of the dataset, data imputation through interpolation of surrounding values had a negative effect on the final models while removing rows with missing values maintained the integrity of the data and led to more successful outcomes.

Beyond determining whether a specific air pollutant or class of pollutants was most greatly linked to bee population decline in California a secondary objective of this paper was to investigate whether counties with higher, medium, or lower population densities had different relationships with honeybee populations and air quality. To further investigate this link, a new variable, population density, was introduced that assigned a value of high, medium, or low to each California county included in the final dataset. High population density areas were considered to be greater than 500 persons per square mile, medium was between 499 and 100 persons per square mile, and low was 99 or less persons per square mile ("California Counties by Population," 2024). Population density was used to segment counties in place of population as the link between population density and increased environmental pollutants like particulate matter and nitrogen oxides has been well established and is more relevant than population data without the added context of county area (mi^2) (Borck and Schrauth, 2019).

Lastly, before training a machine learning model to predict honeybee population, the data for the columns containing yearly mean values for CO, NO_2, ozone, $PM_{2.5}$, and PM_{10} were all normalized using z-scale normalization. Z-scale normalization was chosen to scale these features to a common range due to the varying ranges and units of each air quality value. The goal of normalization was to enhance the accuracy of the resulting machine learning models, and the z-scale method was chosen specifically for its robustness when dealing with outliers.

Modeling Methodology

The machine learning techniques chosen for this project were linear regression, gradient boosting, k-nearest neighbors (KNN), and support vector machines (SVM). Linear regression was chosen to investigate whether the relationship between the various predictors and honeybee population was linear and later SVM was chosen to investigate whether the relationship was nonlinear. Due to the small number of rows in the dataset and smaller number of features, KNN was also chosen as a modeling technique. KNN and SVM specifically have been shown to have low mean absolute error and consistent performance in similar honeybee models that were aimed at classification rather than regression (Robles-Guerrero et al., 2023). Gradient Boosting has also been shown to produce successful regression models with similar air-pollution factors as predictors, although for occurrences of respiratory diseases in humans, not yet in relation to honeybee population modeling (Ku et al., 2022). Modeling techniques like neural networks (NN) and elastic net regression were considered as well, however due to the smaller size of the data and lack of high dimensionality in the dataset these models were not chosen.

Model Selection

Linear Regression

Linear regression models are often used in environmental science due to their reliable and relatively simple nature. Linear regression analysis is a supervised machine learning model that uses known values of input and output variables to build a model that predict the value of one variable based on the value of another (Bigelow and Lawton, 2024). A linear regression model was trained using a randomly selected 80% of the dataset and considered CO, NO_2, ozone, $PM_{2.5}$, PM_{10}, year, and high, medium, or low population density as predictor variables with only ozone showing significance as a predictor variable with a p-value of 0.0051. This model was then used to predict bee populations on a test set consisting of the remaining 20% of the data. The model achieved a low R^2 value of 0.316 indicating that this model only explains ~32% of the variance in the bee population and is not a great fit to the data.

Gradient Boosting

Gradient boosting models are based on a type of supervised machine learning model called decision trees. Boosting is an ensemble method that combines multiple weak decision trees to create a more robust final model (Chen and Guestrin, 2016). For this project, the XGBoost package in R was used to train and test a boosted machine learning model with the final number of decision trees included set at 100. The final XGBoost model resulted in a significantly lower R^2 value than the previous linear regression model at 0.037.

Support Vector Machine

A support vector machine was chosen to investigate whether the link between air quality predictors and honeybee population was more complicated than a linear relationship. Support vector regression (SVR) is a type of SVM that is applied to more complicated relationships but are less prone to overfitting like decision trees or neural networks ("What Are SVMs?," 2023). For this model, the radial kernel was chosen and the hyperparameters gamma and cost were tuned to 0.1 and 1, respectively. This resulted in a model with an R^2 of 0.444.

k-Nearest Neighbors

K-NN is a simple, nonparametric supervised machine learning method. K-nearest neighbors is an instance-based model that groups data points based on the proximity of a data point to its k neighbors, with neighbors determined by a Euclidean distance metric (Shi et al., 2022). The k value represents the number of closest neighbors and a value of 5 was chosen for this model by using cross-validation to evaluate the performance of various k values. K-NN resulted in the most successful model on the basis of R^2 with a value of 0.533.

Conclusion

The objective of this chapter was to discuss the data collection and preparation process as well as the modeling methodology used to predict honeybee populations. Four supervised machine learning regression models were assessed on their ability to model bee populations using various air quality metrics, including linear regression, k-NN, XGBoost, and SVM. The models were trained and then tested on their predictive abilities and compared using a variety of metrics. Chapter 4 will discuss this comparison further using metrics like R^2, root mean squared error (RMSE), and mean absolute error (MAE) and determine which models are most successful in predicting bee population as well as which variables are most important for these predictions.

Chapter 4: Results

Chapter 3 covered the four machine learning models that were trained and tested to predict honeybee population in California using presence of air pollutants as variables. The models included linear regression, k-NN, XGBoost, and SVM. Chapter 4 will discuss the results of these four models and include a selection of the best model based on metrics like root mean squared error (RMSE), mean absolute error (MAE), and R^2. Feature importance and feature selection will also be discussed for the best model and will be used to determine the important drivers of honeybee population in California. This chapter will also include an evaluation of how well the findings of this chapter have contributed to meeting the overall objectives of this project.

Results Evaluation

Selection of Evaluation Metrics

To compare the results of the four models produced in Chapter 3, three metrics were chosen to measure models' performance: RMSE, MAE, R^2. As a visual aid, residual plots were created for each of the four models as well. Mean square error (MSE) is a measure of the difference between the values predicted by a model and the actual values borne out by the data. RMSE is then the square root value of the error and was chosen as a metric to evaluate model performance because taking the square root allows the resulting RMSE to be in the original units of the target value which is total honeybees. Similarly, MAE was chosen as it also allows the units for the error value to match the units of our target variable. While RMSE is a metric that penalizes larger errors, changes in error in terms of MAE are linear and aid in easier interpretation (Schneider and Xhafa, 2022).

R^2 is a measurement of how well a model's predictions approximate the actual data points. An R^2 value of 1 essentially says that 100% of the variance in the data is explained by the model, meaning higher values of R^2 typically represent a more successful model. R^2 was chosen over adjusted R^2, another common metric, because all four models developed in chapter 3 had the same number of predictor variables, therefore adjustment was needed for comparison.

Model Comparison

Figure 3 displays the residual plots for each of the 4 models trained in the methodology section of this project. Assessing the residual plots for patterns, outliers, or clustered data points helped determine what adjustments could be made to the models and helped guide what type of model might be the most successful for this dataset. Residuals represent the difference between the honeybee population value predicted by the training model on the segmented testing data and the actual values from the dataset. An exceptionally accurate model would have a residual plot where the points were clustered around zero, symmetrically distributed, and show no clear patterns.

The residual plot from the linear regression model in Figure 3 showed extreme positive outliers which indicated that a linear model may not be the best fit for this data. The XGBoost and SVM model's residual plots showed increasing variance as predicted honeybee population values increased, indicating that these models also may not be the best fit for the data. Lastly, the residual plot for the k-NN model in Figure 3 displayed a fairly random distribution of points yet still showed signs that the model may not be very strong in predicting honeybee population.

Figure 3

Comparison of Residual Plots for Applied Machine Learning Methods

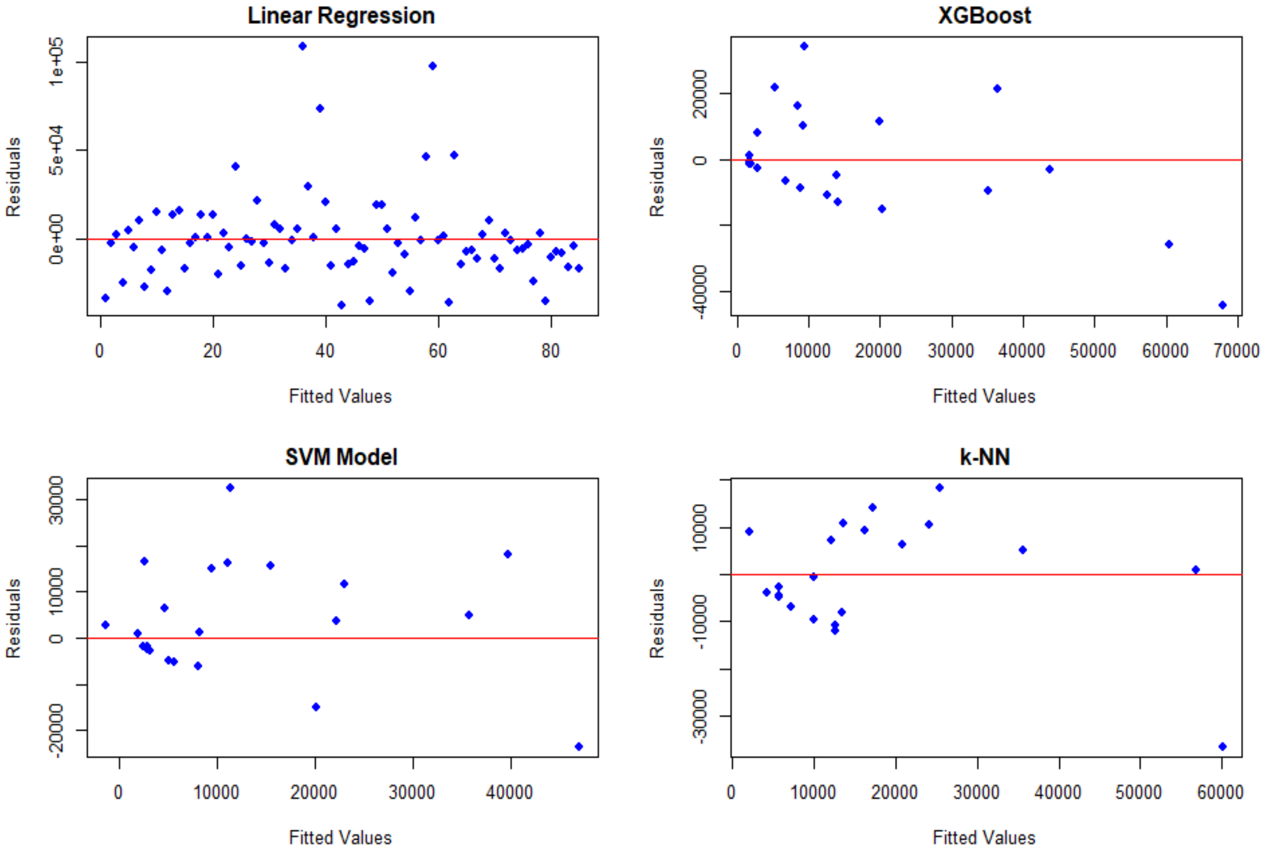


Table 1 displays the model performance metrics used to compare the four machine learning models applied to the California honeybee population data. MAE, RMSE, and R^2 values were calculated after the training phase was completed for each model. In assessing the accuracy and predictive power of each of these models, each of these metrics were considered in conjunction with one another to ensure a comprehensive analysis of each model’s performance.

Table 1

Comparing Model Performance Metrics for Applied Machine Learning Methods

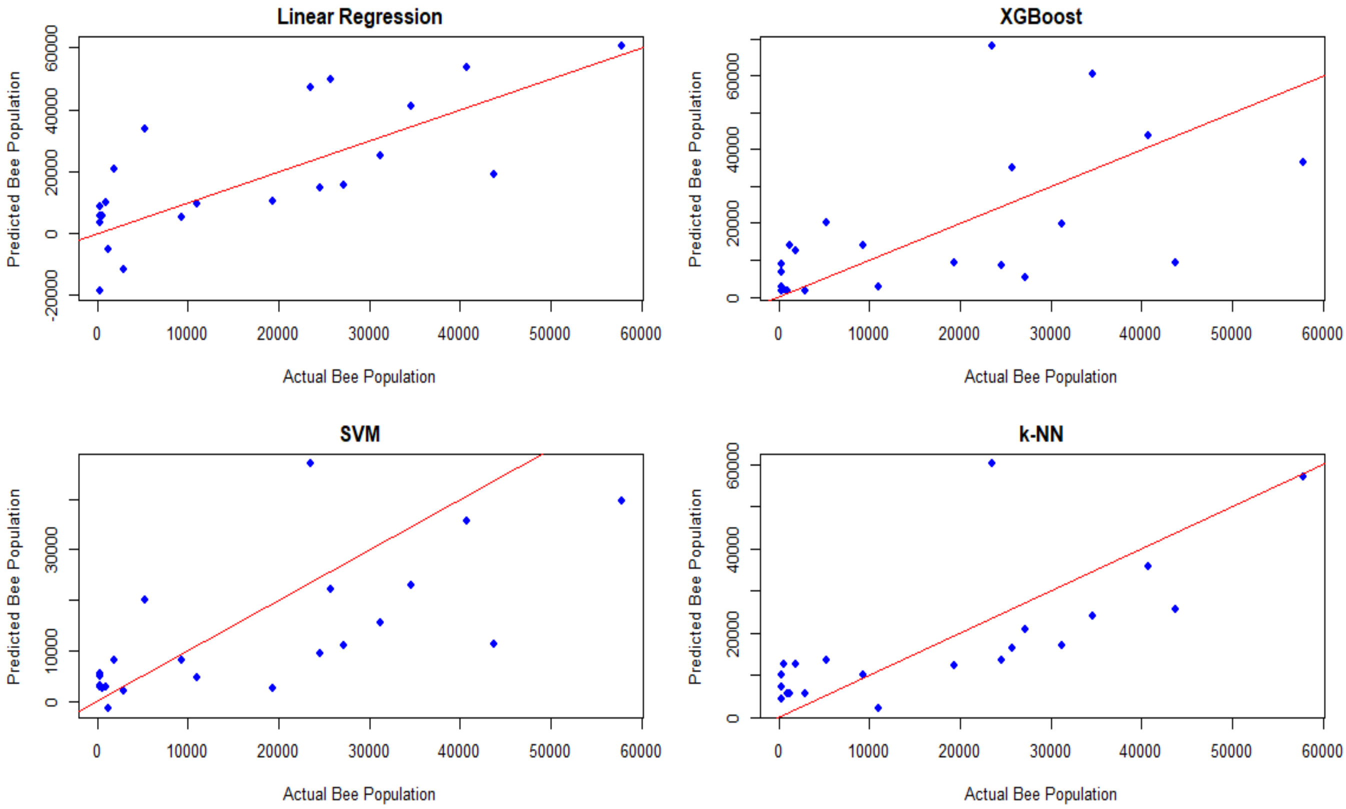
Model	MAE	RMSE	R^2
Linear Regression	11575.37	13995.66	0.316
XGBoost	12364.25	16597.33	0.037
SVM	9518.45	12611.79	0.444
k-NN	8939.32	11561.38	0.533

Alongside the metrics in Table 1 and residual plots in Figure 3, a plot of observed vs. predicted values was constructed for each machine learning model trained and tested in chapter 3. Observed vs. predicted values plots are another tool for evaluating the performance of a regression model by plotting observed values against predicted values alongside a diagonal line that represents a perfect model fit. Similar to residual plots, patterns or asymmetrical clusters of data about the 45-degree line across the plot allude to bias or lack of accuracy in the model.

Figure 4 presents the observed vs. predicted values plot for each applied machine learning method and gives a visual representation of the accuracy of predictions made by each model.

Figure 4

Comparison of Predicted Values vs. Actual Values Plots for Applied Machine Learning Methods



Best Model Selection

Based on all three model comparison metrics, k-NN was the best performing model in predicting honeybee populations. With an R^2 of 0.533, k-NN performed only marginally better than SVM and significantly better than linear regression and XGBoost. While k-NN is the best performing model of the four, an R^2 of 0.533 means that the k-NN model only accounts for around 53% of the variability displayed in the honeybee population data using all five air quality metrics and population density group as predictors in the model.

K-NN had MAE and RMSE values of 8939.318 and 11561.38, respectively. Again, while these are the lowest error scores of all four models, they are still relatively high and show that this model may not be making the most accurate predictions. From the predicted values vs. actual values plot in Figure 4, the plot for k-NN shows that the model overestimates honeybee population in areas with lower honeybee populations and underestimates population in areas with a larger honeybee population.

Feature Selection

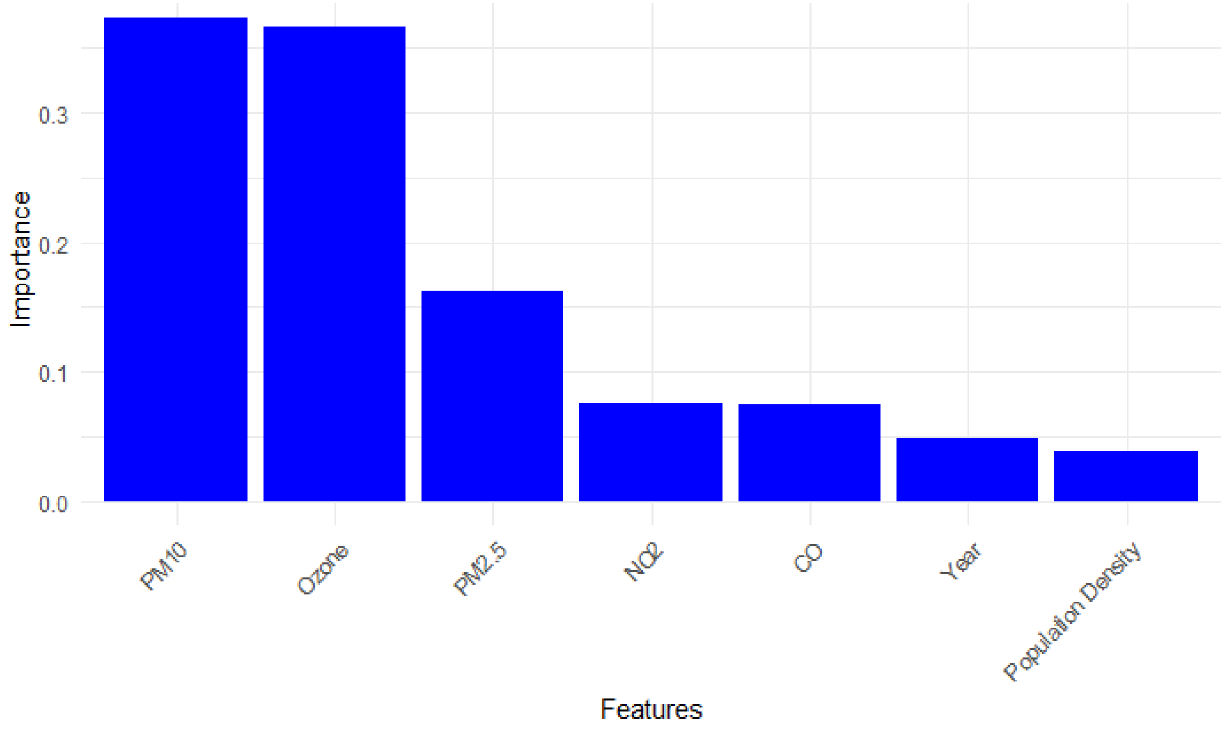
Aside from the Pb and SO_2 columns that were removed for lack of complete data, each variable included in the dataset was used to train the final k-NN model. There were no large effects to the success of the model when attempts were made to prune the number of variables used to train the model.

Feature Importance

Feature importance was determined using permutation performance for the k-NN model. Permutation importance is a method in which the values of each feature are permuted, or randomly rearranged, and performance of the model is recorded. If the performance of a model suffers when a feature is permuted, that feature is important or has more influence on the model (Abdelaziz et al., 2024). Figure 5 displays the feature importance scores for each variable used in training the final k-NN model.

Figure 5

Feature Importance Scores for Selected k-NN Model



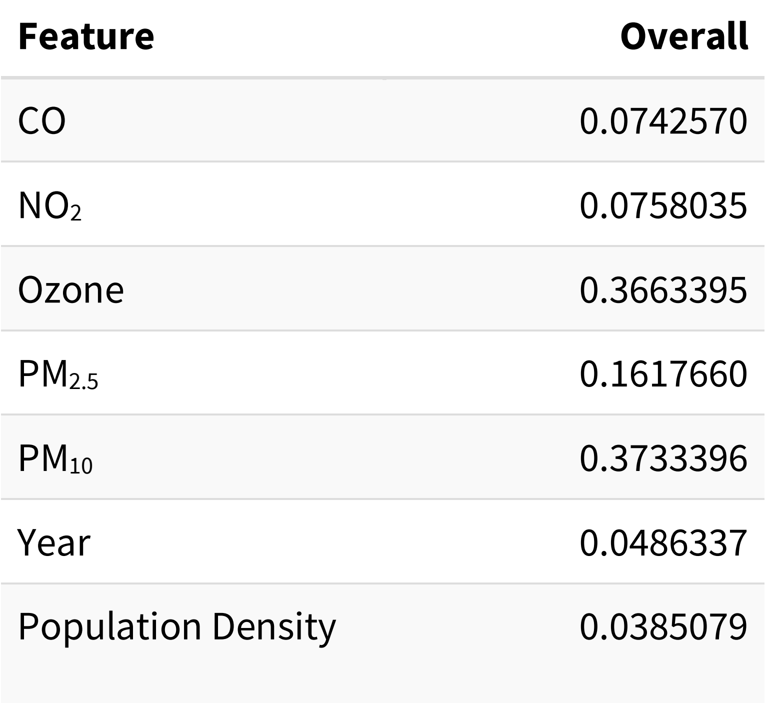
Important Drivers of Honeybee Population Decline

As revealed by the feature importance scores for the k-NN model, and shown in Figure 5, the most important drivers of honeybee population fluctuation in California were PM_{10} and ozone concentrations in the air, with similar scores to one another. After PM_{10} and ozone, $PM_{2.5}$ was the third most influential factor in predicting honeybee population in this model. This is consistent with the previous research on environmental factors contributing to honeybee population decline. This finding is also consistent with the current understanding of how climate change has affected air quality in the past 20 years, specifically by increasing particulate matter and ozone concentrations in the atmosphere.

The remaining factors included in the model had much smaller importance scores than particulate matter and ozone concentrations. NO_2, CO, year, and population density were the least influential factors to the final selected model, in that order. Importance scores for each model feature are listed in Table 2 below. As shown in Table 2, NO_2 and CO featured similar importance scores to one another, as did year and population density.

Table 2

Feature Importance Scores for Selected k-NN Model



Feature	Overall
CO	0.0742570
NO_2	0.0758035
Ozone	0.3663395
$PM_{2.5}$	0.1617660
PM_{10}	0.3733396
Year	0.0486337
Population Density	0.0385079

While CO and NO_2 did not rank as highly as PM_{10}, $PM_{2.5}$, and ozone as influential factors in modeling honeybee population, they may still be detrimental to honeybees in more nuanced ways. Studies show that increased NO_2 levels in the atmosphere from vehicle exhaust can chemically alter floral odors leading to decreased pollinator efficiency in flower visits and lead to direct effects on bee health (Ryalls et al., 2022).

CO may have similar downstream rather than direct effects as NO_2 as CO reacts overtime with oxygen in the atmosphere and can lead to increased levels of CO_2. Increased levels of CO_2 in the atmosphere have been shown to decrease protein concentrations in plant species that support pollinators and therefore lead to negative health outcomes for honeybees and other pollinators (Ziska et al., 2016).

Objectives Evaluation

The first of the three objectives for this project was to determine a method for segmenting California counties based on human population. With the size of the dataset as a limiting factor, separate datasets for high, medium, and low population density counties were not made and instead a new variable was introduced to the dataset. Each county was given a population density ranking of high, medium, or low, and this variable was used in model building in the methodology section of this paper.

While this first objective was met and population density was considered, based on the feature importance scores shown in Table 5, this was not a leading factor in honeybee population prediction with the k-NN model. Despite low importance scores relative to the other variables in the final model, the idea of segmenting counties could still be considered in future work where dataset size may not be a limiting factor.

The second objective of this project was to determine which specific air pollutant or class of air pollutants was the most greatly linked to bee population decline in California. This was achieved through analyzing feature importance and backed by research in the literature review section in this project. Ozone and particulate matter were found to be the most influential factors in predicting bee population decline in California using the k-NN model. Ozone and particulate matter are both air pollutants that are highly associated with vehicle or industrial emissions and are more prevalent in or near large cities. Increases in particulate matter can also occur from natural sources like wildfires or windblown dust.

Determining the machine learning model that predicted honeybee population with the highest accuracy was the third objective of this project and was achieved by developing various models and comparing the accuracy and predictive power of the models. Linear regression, SVM, XGBoost, and k-NN models were all developed, and the most successful model was determined by analyzing a variety of performance metrics including MAE, RMSE, and R^2. Of the four machine learning models that were analyzed, k-NN was determined to be the most successful model for its higher accuracy in predicting honeybee populations in California.

Conclusion

In summary, Chapter 4 conducted a results evaluation by evaluating the success of each model developed in Chapter 3 using a variety of performance metrics to compare accuracy and predictive ability of each machine learning model. The best model selected for predicting honeybee population in California was the k-NN model. For k-NN, all features were used to train the model and the variables determined to be most important in predicting honeybee population were ozone and particulate matter, specifically PM_{10}. The overall objectives of the project were also assessed and determined to have been met by the work laid out in Chapters 3 and 4.

The objective of Chapter 4 was to cover the results of the project, provide preliminary discussion on the importance of the project objectives that were met, and to outline the valuable insights that will be discussed in Chapter 5. In the final chapter of this project, Chapter 5, there will be suggestions for future research and work yet to be done in the realm of air quality research and honeybee population modeling that build upon the discussion in Chapter 4. Chapter 5 will also touch on environmental policy surrounding air pollution and honeybee population management that could be introduced in California and other states with high agricultural output. Finally, there will be a summary of results including limitations of this project and a conclusion that includes the significance of this project.

Chapter 5: Discussion

Chapter 4 evaluated the results of this project, in depth, and touched on some of the significant insights to be gained from the results of this project. Chapter 5 will summarize the findings of the project in the context of the objectives stated in Chapter 1, as well as provide recommendations for managing honeybee population decline in California and offer recommendations for future research in this area. Lastly, Chapter 5 will discuss the limitations of the project and conclude by providing insight into the significance of the study to this field of research.

Summary of Findings

This project had three main objectives with the overarching goal of examining the role of air quality in bee population decline by determining whether accurate models can be built to predict bee population using machine learning. While Chapter 4 contained an evaluation of the ways in which the objectives of this project were achieved through the work done in the methodology section of this paper, this next section will focus on the findings of this project through the lens of each of the three objectives set forth in Chapter 1.

Conclusions by Project Objectives

Objective One: Segmenting California Counties by Population Density

The intention of the first objective was to examine the relationship between population density and bee population. Studies have shown that there is a causal relationship between increased population density and increased presence of environmental pollutants in the atmosphere, specifically pollutants like particulate matter

(Borck and Schrauth, 2019). To investigate the significance of this relationship, a new variable was added to the dataset that segmented California counties into three population density groups: low, medium, and high. High population density areas were considered to be greater than 500 persons per square mile, medium was between 499 and 100 persons per square mile, and low was 99 or less persons per square mile ("California Counties by Population," 2024).

An analysis of feature importance for the machine learning model selected as best in Chapter 4 determined that population density was the least important variable of all included in training the selected k-NN model. This, however, does not mean that population density is not an important factor relating to bee population decline in areas with larger populations. It could be argued that population density is still a factor in honeybee population decline as it is directly related to the amount of air pollution in a county and could be linked to loss of habitat and declines in available food sources for bees which indirectly relate to negative bee health outcomes ("Threat to pollinators," 2024).

Objective Two: Identify Pollutants Most Greatly Linked to Bee Population Decline

The second objective of this project was to determine which specific air pollutant or class of air pollutants was the most greatly linked to bee population decline in California. The feature importance analysis conducted in Chapter 4 indicated that PM_{10} and ozone concentrations were the most significant features used by the selected k-NN model to predict honeybee population in California.

Ozone concentrations, represented in this dataset as average yearly ozone concentration in parts per million, increase due to factors such as climate change, vehicle emissions, and industrial pollution. Increased ozone concentrations have been found in the guts and organs of forager bees that were examined after death and are believed to be linked to increased susceptibility to disease and weakened immune systems (Negri et al., 2015, Feldhaar & Otti, 2020).

Particulate matter accumulated on forager bees is known to greatly impact entire honeybee colonies as they are social insects whose health as a colony is interconnected. Particulate matter settles on nectar, pollen, and water sources that forager bees ingest and collect, and later bring back to the hive, affecting the immune systems of and jeopardizing survival for the entire colony (Feldhaar & Otti, 2020).

Pollutants like CO and NO_2 achieved lower feature importance scores than particulate matter and ozone, however, it could be argued that they still have an influence on the health of honeybee populations. Increased CO and NO_2 concentrations in the air can have negative effects on the local flora that foragers depend on which can directly impact honeybee foraging efficiency and bee colony health outcomes.

Objective Three: Compare Machine Learning Models and Select Best-Performing

The third objective of this project was to construct and compare models for predicting honeybee population using four machine learning techniques. The models assessed were linear regression, SVM, XGBoost, and k-NN. The models were compared using MAE, RMSE, and R^2 as performance metrics. The k-NN model emerged as the most successful with an R^2 of 0.533, MAE of 8939.32, and an RMSE of 11561.38. While this model had a lower than optimal R^2 value, it still significantly outperformed the other models which shows promise for using this type of machine learning technique to estimate future honeybee populations using more robust datasets of air quality data.

Implications of Findings

The findings of this report validate the existing research in the field of bee population dynamics by confirming a link between decreased air quality likely due to climate change and pollution and bee population. The k-NN model showed promise in accurately predicting honeybee population in California, an important agricultural state, using air quality metrics.

Recommendations for Managing Honeybee Population

Based on the findings of this report, summarized above, there are several recommendations that can be made for managing honeybee population and monitoring honeybee population decline. There are currently several successful honeybee population models like BEEPOP, HoPoMo, and BEEHAVE, all of which are incredibly complex in the number of variables included. While some do include chemicals found in pesticides, they all overlook the importance of integrating air pollutants that are due to climate change, vehicle emissions, and industrial emissions. Integrating air quality data into existing honeybee population models will enhance the predictive ability and accuracy of these models. The limited success of the machine learning models in this project lends to the idea that while air quality is certainly a factor in honeybee population decline, it may not stand alone as a strong predictor and would provide more insight when evaluated alongside other known factors causing colony collapse disorder like pesticide use, disease, and mites.

Second, targeted policy to decrease vehicle and factory emissions that contribute to elevated levels of ozone is important to control this important factor in bee colony health outcomes. The EPA designated ozone as a pollutant whose levels need to be limited in outdoor air based on its proven negative health outcomes and it is one of the six main air pollutants identified in the Clean Air Act. The government currently requires states to develop plans that garner EPA approval to meet the air quality standards set forth in the Clean Air Act ("Ground-level Ozone Basics," 2024). This type of environmental policy is based on health criteria for humans and could in the future seek to include health criteria for other vulnerable groups, like honeybees, in drafting new regulations and determining standards.

Particulate matter was another important factor in honeybee population modeling and while increases in particulate matter can be partially attributed to pollution, there are natural causes as well. Wildfires and draught caused by climate change can lead to conditions that may potentially cause increases in airborne particulate matter. Monitoring the effects of climate change on increasing air pollutants that are known to affect the bee population can assist in creating climate adaptation strategies that can be applied alongside existing environmental policy, potentially on a county-by-county basis.

Limitations

While this study has important implications and significant findings, there were many limitations that may have been a factor in inhibiting its success. The lack of detailed honeybee population data collected on a daily, weekly, or monthly basis was a significant issue. While the EPA provides an abundance of publicly available high quality air pollutant data collected daily, this data needed to be aggregated to yearly values to be compatible with the limited publicly available bee population data. Bee population data was also only reported every five years, so while the span of the dataset was from 2002-2022, there were only five years in that time where bee population and air quality data was able to be considered. Missing data was another significant issue. Collection of honeybee population data was inconsistent in many California counties and led to the removal of valuable rows of data and resulted in a quite small dataset which may have impacted the predictive abilities and accuracy of each of the machine learning models tested.

Recommendations for Future Research

To enable a more thorough understanding of the relationship between bee population and common air pollutants, future research should consider collecting firsthand data at more regular intervals to allow for larger datasets with which to build predictive models. Access to more data would allow for a closer examination of the role of population density and could allow for separate models to be constructed for high, medium, and low population density counties or even built on a county-by-county basis. This would allow for an even greater understanding of the influence of air quality on bee populations and could lead to more targeted environmental policy. Future research could also utilize other air quality factors and introduce new pollutants that were not able to be considered in this project due to lack of data, like SO_2 and Pb. Future research should also consider assessing the effects of air pollution on bee populations in other important agricultural states with different landscapes, climates, and population dynamics to begin to identify trends across the United States.

While it was important to narrow the scope of this project to only consider air quality as a factor in bee populations, future research could build on the findings discussed in this paper by modeling more complex relationships. Future research could incorporate other environmental factors like water quality, changes in weather conditions like precipitation and temperature due to climate change, or changes in habitat and local flora due to pesticide use and human settlement to create a more robust model focused solely on the relationship between bee population and environmental factors.

Another consideration for future research is the incorporation of various data science tools like AI or more complex machine learning models like artificial neural networks. This project utilized only a sampling of the modern data science tools available.

Conclusion

This case study examined the role of air quality in bee population decline and determined that accurate models can be built using machine learning to predict bee population based on presence of pollutants in the air. The findings of this project indicated that the k-NN model was the most successful model in predicting honeybee population using CO, NO_2, ozone, PM_{10}, $PM_{2.5,}$ year, and county population density as variables. Ozone and PM_{10} concentrations were found to be the most important drivers of honeybee population decline, of all variables considered. The results of this project affirmed the significance of the negative effects of climate change and air pollution on both vulnerable and critically important honeybee populations. These findings were used to inform recommendations for managing of honeybee populations in California and other important agricultural states and provide direction for future research opportunities in this field.

References

Abdelaziz, M. T., Radwan, A., Mamdouh, H., Saad, A. S., Abuzaid, A. S., AbdElhakeem, A. A., Zakzouk, S., Moussa, K., & Darweesh, M. S. (2024). Enhancing network threat detection with random forest-based NIDS and permutation feature importance. Journal of Network and Systems Management, 33(1). https://doi.org/10.1007/s10922-024-09874-0

Anwar, O., Keating, A., Cardell-Oliver, R., Datta, A., & Putrino, G. (2023). Apis-prime: A deep learning model to optimize Beehive Monitoring System for the task of daily weight estimation. *Applied Soft Computing*, *144*, 110546. https://doi.org/10.1016/j.asoc.2023.110546

Becher, M. A., Osborne, J. L., Thorbek, P., Kennedy, P. J., & Grimm, V. (2013). Review: Towards a systems approach for understanding honeybee decline: A stocktaking and synthesis of existing models. *Journal of Applied Ecology*, *50*(4), 868–880. https://doi.org/10.1111/1365-2664.12112

Bigelow, S. J., & Lawton, G. (2024, September 23). *What is Linear Regression?*. Search Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/linear-regression

Borck, R., & Schrauth, P. (2019). Population density and urban air quality. *CESifo Working Papers*, *7629*. https://doi.org/10.2139/ssrn.3387665

California counties by population (2024). World Population Review. (2024). https://worldpopulationreview.com/us-counties/california

California Department of Food and Agriculture. (2023). *California agricultural statistics review 2022-2023*. https://www.cdfa.ca.gov/Statistics/PDFs/2022-2023_california_agricultural_statistics_review.pdf

Carozzi, F., & Roth, S. (2023). Dirty density: Air quality and the density of American cities. *Journal of Environmental Economics and Management*, *118*, 102767. https://doi.org/10.1016/j.jeem.2022.102767

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

DeGrandi-Hoffman, G., Roth, S. A., Loper, G. L., & Erickson, E. H. (1989). BEEPOP: A Honeybee population dynamics simulation model. *Ecological Modelling*, *45*(2), 133–150. https://doi.org/10.1016/0304-3800(89)90088-4

Environmental Protection Agency. (2024, May 14). *Ground-level Ozone Basics*. EPA. https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

Feldhaar, H., & Otti, O. (2020a). Pollutants and their interaction with diseases of Social Hymenoptera. *Insects*, *11*(3), 153. https://doi.org/10.3390/insects11030153

IBM. (2023, December 23). What Are SVMs?. IBM. https://www.ibm.com/topics/support-vector-machine

Jerrett, M., Burnett, R. T., Beckerman, B. S., Turner, M. C., Krewski, D., Thurston, G., Martin, R. V., van Donkelaar, A., Hughes, E., Shi, Y., Gapstur, S. M., Thun, M. J., & Pope, C. A. (2013). Spatial Analysis of Air Pollution and Mortality in California. *American Journal of Respiratory and Critical Care Medicine*, *188*(5), 593–599. https://doi.org/https://doi.org/10.1164/rccm.201303-0609OC

Khoury, D. S., Myerscough, M. R., & Barron, A. B. (2011). A Quantitative Model of Honey Bee Colony Population Dynamics. *PLoS ONE*, *6*(4). https://doi.org/10.1371/journal.pone.0018491

Ku, Y., Kwon, S. B., Yoon, J.-H., Mun, S.-K., & Chang, M. (2022). Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. *Clinical and Experimental Otorhinolaryngology*, *15*(2), 168–176. https://doi.org/10.21053/ceo.2021.01536

Mayack, C., Cook, S. E., Niño, B. D., Rivera, L., Niño, E. L., & Seshadri, A. (2023). Poor Air Quality Is Linked to Stress in Honeybees and Can Be Compounded by the Presence of Disease. *Insects*, *14*(8), 689. https://doi.org/10.3390/insects14080689

Negri, I., Mavris, C., Di Prisco, G., Caprio, E., & Pellecchia, M. (2015). Honey bees (apis mellifera, L.) as active samplers of Airborne Particulate Matter. *PLOS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0132491

Olivares-Pinto, U., Alaux, C., Le Conte, Y., Crauser, D., & Prado, A. (2024). Using honey bee flight activity data and a deep learning model as a toxicovigilance tool. *Ecological Informatics*, *81*, 102653. https://doi.org/10.1016/j.ecoinf.2024.102653

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution, 25*(6), 345–353. https://doi.org/10.1016/j.tree.2010.01.007

Reilly, J. R., Artz, D. R., Biddinger, D., Bobiwash, K., Boyle, N. K., Brittain, C., Brokaw, J., Campbell, J. W., Daniels, J., Elle, E., Ellis, J. D., Fleischer, S. J., Gibbs, J., Gillespie, R. L., Gundersen, K. B., Gut, L., Hoffman, G., Joshi, N., Lundin, O., … Winfree, R. (2020). Crop production in the USA is frequently limited by a lack of pollinators. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1931), 20200922. https://doi.org/10.1098/rspb.2020.0922

Robles-Guerrero, A., Saucedo-Anaya, T., Guerrero-Mendez, C. A., Gómez-Jiménez, S., & Navarro-Solís, D. J. (2023). Comparative study of machine learning models for

Bee Colony Acoustic Pattern Classification on low computational resources. *Sensors*, *23*(1), 460. https://doi.org/10.3390/s23010460

Russell, S., Barron, A. B., & Harris, D. (2013). Dynamic modelling of honey bee (apis mellifera) colony growth and failure. *Ecological Modelling*, *265*, 158–169. https://doi.org/10.1016/j.ecolmodel.2013.06.005

Ryalls, J. M. W., Langford, B., Mullinger, N. J., Bromfield, L. M., Nemitz, E., Pfrang, C., & Girling, R. D. (2022). Anthropogenic air pollutants reduce insect-mediated pollination services. *Environmental Pollution*, *297*, 118847. https://doi.org/10.1016/j.envpol.2022.118847

Schmickl, T., & Crailsheim, K. (2007). Hopomo: A model of Honeybee intracolonial population dynamics and resource management. *Ecological Modelling*, *204*(1–2), 219–245. https://doi.org/10.1016/j.ecolmodel.2007.01.001

Schneider, P., & Xhafa, F. (2002b). Chapter 3 - Anomaly detection: Concepts and methods. In *Anomaly detection and complex event processing over IOT data streams: With application to eHealth and patient data monitoring* (pp. 49–66). essay, Academic Press. Retrieved November 22, 2024, from https://doi.org/10.1016/B978-0-12-823818-9.00013-4.

Shi, Y., Yang, K., Yang, Z., & Zhou, Y. (2022). Chapter Two - Primer on Artificial Intelligence. In *Mobile Edge Artificial Intelligence: Opportunities and challenges* (pp. 7–36). essay, Academic Press, an imprint of Elsevier. Retrieved November 20, 2024, from https://doi.org/10.1016/B978-0-12-823817-2.00011-5.

Threats to pollinators: U.S. Fish & Wildlife Service. FWS.gov. (2024). https://www.fws.gov/initiative/pollinators/threats

Torres DJ, Ricoy UM, Roybal S (2015) Modeling Honey Bee Populations. *PLoS ONE 10*(7): e0130966. https://doi.org/10.1371/journal.pone.0130966

United States Department of Agriculture, National Agricultural Statistics Service (NASS). (1998). Quick Stats Database. Retrieved from https://quickstats.nass.usda.gov/#EE3FA4B0-CF5C-3721-92DB-4B59E85BA7BC

United States Environmental Protection Agency. (2011, February). Air Quality Guide for Nitrogen Dioxide. https://document.airnow.gov/air-quality-guide-for-nitrogen-dioxide.pdf

United States Environmental Protection Agency. (1980). Outdoor Air Quality Database. Retrieved from https://www.epa.gov/outdoor-air-quality-data/download-daily-data

Watanabe, M. E. (2008). Colony collapse disorder: Many suspects, no smoking gun. *BioScience*, *58*(5), 384–388. https://doi.org/10.1641/b580503

Zhu, S., Horne, J. R., Mac Kinnon, M., Samuelsen, G. S., & Dabdub, D. (2019). Comprehensively assessing the drivers of future air quality in California. *Environment International*, *125*, 386–398. https://doi.org/10.1016/j.envint.2019.02.007

Ziska, L. H., Pettis, J. S., Edwards, J., Hancock, J. E., Tomecek, M. B., Clark, A., Dukes, J. S., Loladze, I., & Polley, H. W. (2016). Rising atmospheric CO2 is reducing the protein concentration of a floral pollen source essential for North American bees. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1828), 20160414. https://doi.org/10.1098/rspb.2016.0414

Appendix

Link to code: https://github.com/j-roskam/DS785-Capstone