CS 760: Machine Learning Fall 2020

Machine Learning to Predict Honey Production from Air Quality

AUTHORS: Carter Sifferman and Keaton Leppanen

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

Bees are essential to the health of our environment and agricultural industry, however, in recent decades, bee hives' ability to produce honey has drastically declined. Many of the causes for this drop in productivity are still unknown, but one possibility is that air pollution is contributing to this trend. In this project, we attempt to use machine learning to predict honey production rates from air quality. More specifically, we design three predictive models (a linear regression model, a KNN regression model, and a Geo-KNN regression model) which use historical air quality and honey production data to predict future trends. Unfortunately, the accuracy of these models turned out to be quite poor and our ultimate results inconclusive. However, through these experiments we were still able to gain valuable insight into the problem domain and provide discussion about the challenges and potential solutions work in this area presents, as well as create a novel approach to the problem that we call "Geo-KNN".

Introduction

Honey production is of keen interest to entomologists, environmentalists, and the agriculture industry. While the number of bee colonies has remained relatively stable in the past decades, honey production has been on the decline. In 2019, the United States produced 152 million pounds of honey, compared to 230 million pounds in 1987 [9]. In turn, honey prices have increased from less than 50 cents per pound to over two dollars per pound in the last thirty years [9]. Being able to predict these changes in price is helpful for honey producers and sellers. Aside from economic usefulness, bees are an indicator species, meaning that their health is a good indicator of the health of the surrounding ecosystem [6]. Being able to make accurate predictions about bee health could give accurate predictions about the health of the entire ecosystem. Additionally, accurate prediction would allow researchers to plug in different projections for air quality to see how honey production may fare in different futures with different levels of action to address climate change.

While approaches with similar goals exist, they're primarily used for crops rather than honey or have been applied only to honey production in Spain or Australia. Our approach is unique in that it combines data on air quality with data on honey production for the United States.

Related Work

Being able to predict crop yields is of obvious value to the agricultural industry and is crucial for preventing famine. As such, humans have been doing it for thousands of years [4]. Recently, however, we have gotten much better at it. Most modern approaches use some real-time sensor data and/or historical records fed into a machine learning model. Johnson (2014) uses real-time sensor data to forecast corn and soybean

yields in the U.S. [5]. To generate a prediction for an area, Johnson calculates the mean of the features over the area. Ultimately Johnson uses a regression model to generate a prediction., Bolton and Friedl (2013) use a similar approach to Johnson [2]. A more recent approach uses more sophisticated AI techniques to extract features from remote sensing data. In [10], the sustainability and artificial intelligence lab at Stanford use publicly available remote sensing data to predict crop yields two months in advance. They use a deep Gaussian process model and crucially do not assume that crop yields are mutually independent between counties, in contrast to previous approaches. Additionally they use AI to discover relevant features rather than hand-crafting them, as previous techniques such as Johnson and Bolton and Friedl did. Rather than a linear regression model, they use a CNN and LSTM Network. Ultimately this leads to better predictive accuracy than any previous approach.

The sub-area of predicting honey production is less well researched, and there is reason to believe that general crop yield prediction techniques will not transfer perfectly to honey yields, which are affected by entirely different factors. To the best of our knowledge, the only papers which attempt to predict honey yields from statistical models are Rocha and Dias (2017) [7] and Campbell et al. (2020) [3]. Rocha and Dias (2017) use "an automated forward-backward variable screening procedure" to find weather variables relevant to honey production, and use those variables with fitted radial basis functions to predict honey production in Spain. They find that spring temperatures and September sunshine are the best predictive features for honey production, but also that the factors which predict honey production vary very widely between regions in Spain. Campbell et al. (2020) use a regression trees with gradient boosted regression to predict honey production in Southwestern Australia by first predicting the growth of marri trees, a common nectar source for the bees. Campbell et al. use 18 weather and ecological features for their model, but do not include air quality data. They achieve reasonable predictive accuracy and find that November temperatures are the most important variable for honey production in Australia - this does not contradict the result of Johnson (2014) as November is Australia's spring. To our knowledge, no previous research has used air quality data to predict honey production, or used statistical techniques to predict honey production in the United States.

Datasets

The novelty of our approach comes from the combination of two datasets: Historical US Air Quality from the US EPA [1] and Historical Honey Production from the US National Agricultural Statistics Service [8]. The EPA dataset is gathered partially from the same sensor network as used by [5], although we use different features. We choose to combine these two datasets because air quality affects bees.

The EPA dataset is very large and detailed, containing over 2.4 billion rows of air quality measurements in the US from 1980-2020. The honey dataset, on the other hand, contains 627 rows of state-level honey production data from 1998-2012. This limits our training samples to the state level and to 1998-2012. Because the EPA dataset is so much more detailed than the honey dataset, we need to reduce it down to a manageable number of features for each state-year data point. First, we select eight air quality features to use in our feature vectors. Units and sample frequency were chosen to be as consistent as possible between features within the constraints of measurements available in the EPA dataset.

Once we have selected our eight features, we still have thousands of air quality data points for each honey data point, as the EPA dataset is at the county and daily level while honey is at the state and annual level. To aggregate the air quality data, we average each feature over each county in each state and over each measurement in each year. This averaging method has proven effective before in [5]. In our final feature vector, we also include a dummy variable for the year which the measurements come from - this helps us to control for the overall reduction in honey production and change in air quality (increase in greenhouse gasses and decrease in particulate matter) year over year. We cover how this year feature affects our model's predictive ability in the approach section. Finally, we include the latitude and longitude of the center of the

state which the sample refers to, for use in a geography-based KNN implementation detailed in the approach section. Table 1 shows each feature, each of which may or may not be used in a given implementation:

Table 1: Final Extracted Features and Response Variable

Measure	Unit	Sample Frequency
o3	parts per million	8-hour average
co	parts per million	8-hour average
so2	parts per billion	3-hour average
no2	parts per billion	1 hour
particulate matter	micrograms per cubic meter (25 C)	24 hour
barometric pressure	millibars	1 hour
temperature	degrees Fahrenheit	1 hour
wind speed	knots	1 hour
state	position in alphabetical order	n/a
latitude of state center	decimal degrees	n/a
longitude of state center	decimal degrees	n/a
honey yield per colony	pounds per year	n/a

We use honey yield per colony as our predictive variable y. We do so rather than total yield or number of colonies for two reasons. Firstly, in the last 30 years, the number of bee colonies in the US has remained relatively stable, while the amount of honey produced has drastically decreased. This means that the limiting factor for honey production is not the number of colonies, but the amount of honey produced by each colony. To get an estimate for total honey production, one can multiply last year's number of hives by this the next year's predicted honey production, as the number of hives is unlikely to change drastically. Second, yield per hive does not depend on the size of the area being predicted in the same way as total honey production. By using yield per hive we are controlling for the number of hives, and our models will be able to be generalized to any area, from a single farm, to a county, or to a country.

In total we have N = 305 feature vectors x of size D = 9 and their corresponding responses y.

***Note*: Regarding the Size of N**

The most unfortunate aspect of our data is the relatively low number of samples N = 305. This issue is primarily due to gaps in the seemingly very thorough EPA Air Quality dataset. Despite having billions of records over a wide span of years, state monitoring practices and policies are far from uniform. This means that many states lack appropriate data for certain measurements in certain years, resulting in incomplete or 'holey' feature vectors (e.g. Wyoming did not monitor SO2 levels in 1998). We dispose of many of our initially assembled feature vectors because they have one or more missing attribute value.

Before removing incomplete vectors we have 1000 feature vectors, but afterwards, we are left with our current N value. We attempted to remedy this deficiency by exploring the process of imputation - which would have helped to fill in some of the gaps and increase the number of feature vectors drastically. However, given the nature of the gaps, we did not feel comfortable employing any of the common imputation methods such as 'carry forward' or 'linear imputation' as we believe it would have resulted in data not necessarily characteristic of reality. Another attempted solution we explored was using a different Honey Production dataset released by the USDA which focused on the county level. The theory being that even with the same proportion of gaps, the increased number of counties when compared to states would help increase our N. However, we met with similar issues as we had with the state level data. In addition, due to decrease in overlap between the specific counties monitored in the EPA dataset and those in the Honey Production dataset, the total N value was actually less than that of the state level data.

Approach

We implement three different approaches to predict honey yield per colony, all of which perform linear regression on some subset of our dataset.

Linear Regression

Our linear regression approach is a simple linear regression by minimizing the MSE.

For this approach, $\mathbf{X}$ begins as a 305 x 9 feature vector containing the features: o3, co, so2, no2, particulate matter, pressure, temperature, and wind speed. First we check that all the features we have are relevant by computing the covariance matrix:

$$\hat{\mathbf{\Sigma}} = \sigma^2(\mathbf{X}^\top \mathbf{X})$$

We can then compute the significance of each feature with the following, where $\hat{\theta}_j$ is the jth entry in θ and ν_j is the j,jth entry in $\mathbf{\Sigma}$

$$\frac{\hat{\theta}_j}{\nu_j}$$

We then compare the significance value to $\tau = \mathbf{\Phi}_\chi^{-1}(0.05) = 0.2357$ to determine whether a feature is significant and get the following results:

Table 2: o3 is the only feature deemed insignificant

Feature	Significance	significant?
o3	0.0001	no
co	0.403	yes
so2	37.92	yes
no2	17.11	yes
particulate matter	26.07	yes
pressure	11945.69	yes
temperature	271.48	yes
wind speed	3762.53	yes
year	291.24	yes

We find that o3 is the only insignificant feature and do not include it in our feature vector for any of our three approaches. We also augment our feature vectors to include a bias term at θ_0. This makes the $\mathbf{X}$ we plug use to minimize the MSE is a 305 x 9 feature vector.

We optimize the linear regression using the matrix solution to minimize the mean-squared error of theta:

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

By minimizing the MSE over the entire dataset we get the following weights in $\hat{\theta}$:

Feature	Weight
intercept	84.86
co	-20.08
so2	1.01
no2	0.15
particulate matter	-1.19
pressure	0.039
temperature	0.278
wind	-0.45
year	-2.07

To get an estimate $\hat{\mathbf{y}}$ for a query feature vector $\mathbf{x}$, we calculate:

$$\hat{\mathbf{y}} = \mathbf{x}^\top \hat{\theta}$$

K-Nearest Neighbors Regression

Our K-nearest neighbors regression approach finds a prediction point's k nearest neighbors and performs a linear regression on only those neighbors, and uses that linear regression to predict the response $\mathbf{y}$. We use the same $\mathbf{X}$ as used in our linear regression approach.

First we find the k nearest neighbors to a query feature vector $\mathbf{x}$. Due to the size of our dataset we are able to do this with a brute-force approach (calculate all the distances and sort). Distance is defined as the following where $\mathbf{X}_i$ is the ith column of $\mathbf{X}$:

$$\text{distance} = \|\mathbf{x} - \mathbf{X_i}\|$$

We then minimize the MSE as in the linear regression approach, but using an $\mathbf{X}$ which is a k x 9 feature vector containing only the k nearest neighbors to $\mathbf{x}$. And similarly for $\mathbf{y}$.

$$\hat{\theta}_{\text{neighbors}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

We then use $\hat{\theta}_{\text{neighbors}}$ to get an estimate $\hat{\mathbf{y}}$ for a query feature vector $\mathbf{x}$:

$$\hat{\mathbf{y}} = \mathbf{x}^\top \hat{\theta}_{\text{neighbors}}$$

We tune the k parameter manually and find that k=20 gives the best results. This is a reasonable number, as it still restricts the regression to neighbors but provides sufficient datapoints to be robust to noise in the data.

Geo-K-Nearest Neighbors Regression

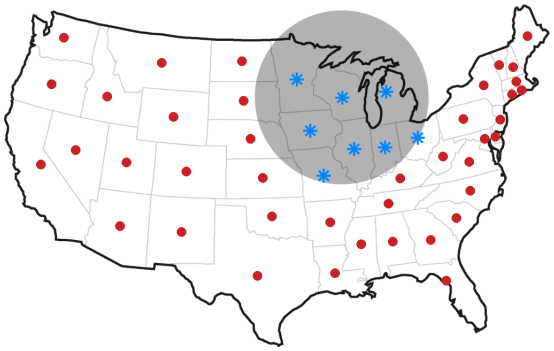
We introduce a somewhat novel spin on k-nearest neighbors which is especially suited to our dataset. We call this approach geo-k-nearest neighbors as the nearest neighbors are found based on geography (location) rather than the norm of the entire feature vector. The motivation for this approach is the work of Rocha and Dias (2017) in which the authors find that the factors that predict honey production vary widely between regions within Spain [7]. If this is also true in the United States, it suggests that performing regression on measurements from a geographical neighborhood could give better predictive accuracy than using the entire United States dataset that we have used so far.

This approach is identical to K-nearest neighbors regression described above except we use a different notion of distance, which is based only on the latitude and longitude of the geographical center of the state which each feature vector comes from

$$\text{distance} = \|\mathbf{x}_{\text{latitude,longitude}} - \mathbf{X_i}_{\text{ latitude,longitude}}\|$$

This has a nice visual representation, which is a circle (because we are using l2-norm) growing out from the query point until it finds k nearby neighbors.

Figure 1: Example of a Geo-KNN search for nearest neighbors for a prediction in Wisconsin



Once we have found neighbors we preform regression just as in k-nearest neighbors regression, using the same features as in k-nearest neighbors regression. Latitude and longitude are used only to find nearest neighbors, not for regression.

This approach allows for arbitrary query points - we can query any latitude and longitude (within the United States, reasonably) and find neighbors, even if the query point is not the center of a state. Additionally, this can easily be expanded to counties rather than states, or even use the specific latitude and longitude of measurement stations or honey farms, if the information were available.

Results

To evaluate the effectiveness of our three approaches, we performed a 10 fold cross-validation. We randomly partitioned our 305 feature vectors into 10 approximately even groups. We then used each of these partitions as the test set to measure the accuracy of our predictive models when trained on the other 9 partitions. The aggregate results in the form of the mean squared error are recorded in Table 3.

Table 3: Average MSE of Different Approaches Across 10 folds of Cross-Validation

Implementation	MSE
Linear Regression	201.9
KNN Regression	166.2
Geo-KNN Regression	176.6

Our most exciting result is that our Geo-KNN approach is more accurate than linear regression on the entire dataset. This gives further evidence for the conclusion of Rocha and Dias (2017) that the factors

which influence honey production vary significantly by region [7]. We believe that applying Geo-KNN or an approach like it to any crop yield prediction problem may improve accuracy, and would love to see future work expriment with an algorithm like Geo-KNN in different domains.

Unfortunately, the MSE for all of our models is relatively high, indicating that none of our models are particularly effective at predicting honey production based off of air quality data. We believe the inaccuracy can be traced to two primary sources.

First, our small N value. In order for predictive models to be effective they need a large quantity of data to base their predictions on - our 305 feature vectors fell short of this mark. Despite initially seeming quite large, when processed it was revealed that due to limited year ranges and frequent missing data points the amount of usable data was actually quite limited. The EPA Air Quality dataset in particular has quite a few large gaps for many of its measurements due to states having differing monitoring policies and those policies changing over time. These holes necessitated us removing the majority of potential feature vectors bringing us from around 1000 potential feature vectors to the 305 we use in our experiments. We attempted to rectify this through various means including exploring using counties instead of states, however none were met with success.

Second, our base premise could potentially be flawed. It could be that these particular air quality features are not a reliable indicator of honey production either because they have no impact on bees' or that those effects are too small and varied to be picked up on at a state level. Similarly, it could be because there exist much greater factors impacting production, such as pesticide usage or predation, which effectively add noise to our data.

Conclusion and Future Work

We set out working on this project because we were interested to see the effects air pollution had on bees and despite our results being inclusive in that respect, we still gained valuable insight into the available data and the mechanisms which could be used to answer that question. The next step in pursuing this line of inquiry is to address the limited data problem. Ideally, years from now there will be sufficiently consistent and numerous data points which would allow for a larger N to be substituted into our experiments. Failing that, it would interesting to see the results of finding a way to deal with the incomplete feature vectors. Exploration of predictive models which can deal with holes in the feature vectors, such as some variants of KNN, could be used to circumvent the problem. Another option could be using some form of data imputation to fill in the gaps - we experimented around with this as before dismissing many of the common data imputation methods to not be sufficient to provide us with accurate results, but with a more sophisticated system of imputation one could take advantage of the fact that many of the pollutants have a general exponential trend to fill in the gaps. Even though we didn't attain the answers we wanted, our experiments can serve as a basis for further exploration into this interesting area.

Resources

Source Code - GitHub Repository
Honey Production Dataset on Kaggle
EPA Air Quality Dataset on Kaggle

References

[1] US Environmental Protection Agency. *Historical Air Quality*. URL: https://www.kaggle.com/epa/epa-historical-air-quality. (accessed: 12.04.2020).

[2] Douglas K. Bolton and Mark A. Friedl. "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics". In: *Agricultural and Forest Meteorology* (2013). URL: https://www.sciencedirect.com/science/article/abs/pii/S0168192313000129.

[3] T. Campbell et al. "Machine Learning Regression Model for Predicting Honey Harvests". In: *Agriculture* (2020). URL: https://www.mdpi.com/2077-0472/10/4/118.

[4] Helaine Elsin. "Encyclopaedia of the History of Science, Technology, and Medicine". In: 2008, p. 1753.

[5] David M. Johnson. "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States". In: *Remote Sensing of Environment* (2014). URL: https://www.nass.usda.gov/Research_and_Science/Cropland/docs/JohnsonRSE14_Yield.pdf.

[6] Heather Pilatic. *Honey Bees – An Indicator Species in Decline*. URL: http://www.panna.org/sites/default/files/PAN%5C%20UK%5C%20News_March2011_CCD.pdf. (accessed: 12.03.2020).

[7] Humberto Rocha and Joana Dias. "Honey Yield Forecast Using Radial Basis Functions". In: *International Workshop on Machine Learning, Optimization, and Big Data* (2017). URL: https://estudogeral.sib.uc.pt/bitstream/10316/45897/1/honey_yield.pdf.

[8] National Agricultural Statistics Service. *Honey Production (1998 - 2012)*. URL: https://www.kaggle.com/jessicali9530/honey-production. (accessed: 12.04.2020).

[9] USDA. *Honey Bees Statistical Summary*. URL: https://www.nass.usda.gov/Publications/Highlights/2019/2019_Honey_Bees_StatisticalSummary.pdf. (accessed: 12.03.2020).

[10] J. You et al. "Combining Remote Sensing Data and Machine Learning to Predict Crop Yield". In: *AAAI Conference on Artificial Intelligence* (2017). URL: https://cs.stanford.edu/~ermon/papers/cropyield_AAAI17.pdf.