Final Report
Bee Colony Collapse - Capstone 2

Original Hypothesis

The original hypothesis was that the cause of bee colony failure in states in the upper midwest region would differ significantly from other areas in the United States, largely due to the effects of the cold climate and how that affects susceptibility to varroa mites. If that proved to be true, it would mean that disease prevention strategies and spending could vary by region. However, for the most part, this did not prove to be the case.

Data Wrangling

Data was obtained through the USDA, as collected by Cornell University. The broad description from the USDA stated that data was collected for operations with five or more colonies on a quarterly basis and on an annual basis for operations with fewer than five colonies. However, upon examining the data, only information from operations with five or more colonies was collected consistently.

The data was presented in a form that required a significant amount of rework to be usable. Rather than being presented in a continuous series of rows, the first seven variables in the data for each state were stored in the upper half of each .csv file, while the second set of variables were stored in the lower half of each .csv file. Each chunk of data was separated by several lines of text describing the data to follow. Thus, the data required a large amount of processing to present all the data for each state in one row and remove the extraneous text.

It was found that one entire season of data was not collected at all. Per notes in the source data, data collection did not occur for the 2nd quarter of 2019. Arguably, the spring quarter could be seen as the most important in terms of analysis. Spring is when hives are opened to determine losses in the colonies over the winter and when some causes of colony loss can be more clearly determined.

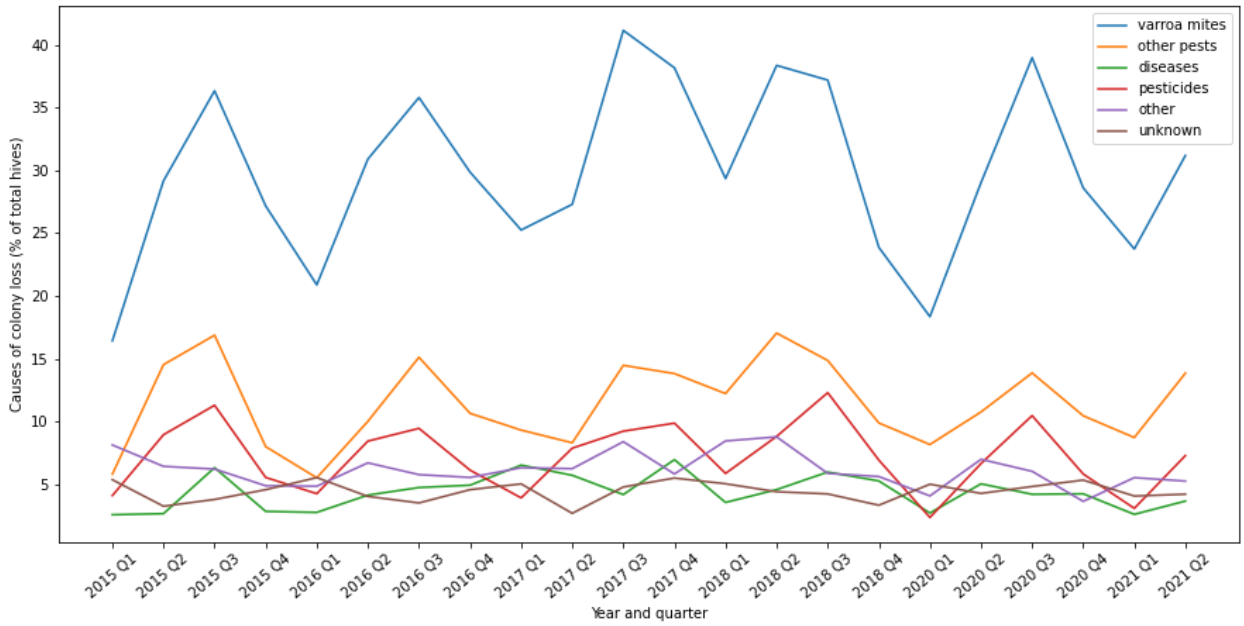
Data for colony loss was collected for varroa mites, other pests (e.g. tracheal mites, wax moths, hive beetles), diseases (e.g. foulbrood, chalkbrood, tobacco ringspot virus, sacbrood virus), pesticides, other (e.g. weather, starvation, insufficient forage, queen failure, hive damage due to storms), and unknown.

Caveat

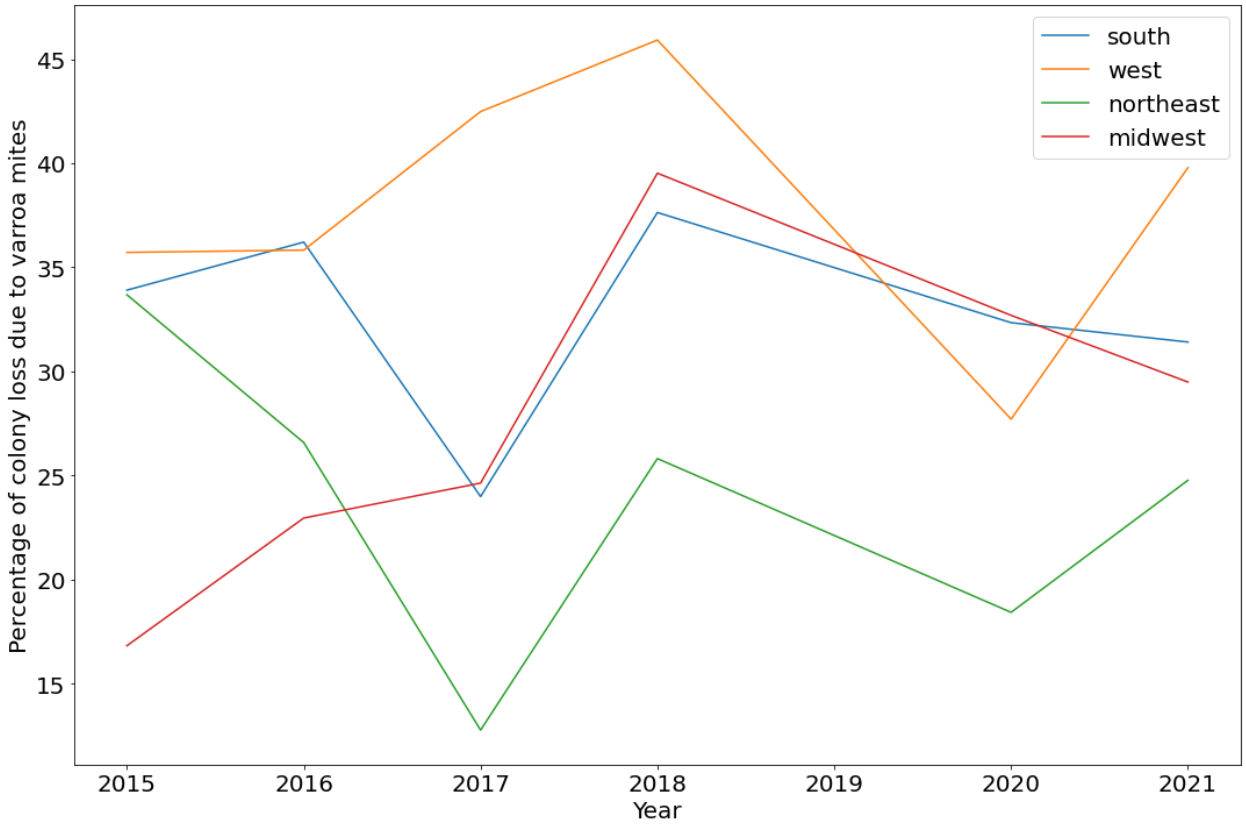
There were two data columns that were not used in the modeling. The data included the number of colonies added and the number of colonies renovated (adding a new queen to an existing colony when the previous queen had died or had reduced egg yield). More often than not, the math did not work when comparing the total of starting colonies, added colonies, and renovated colonies, less the colonies lost, to yield the maximum colonies in a season. Therefore, it was decided not to include the added colony numbers and renovated colony numbers, since any colonies added or renovated in one season would be reflected in the number of starting colonies for the following season.

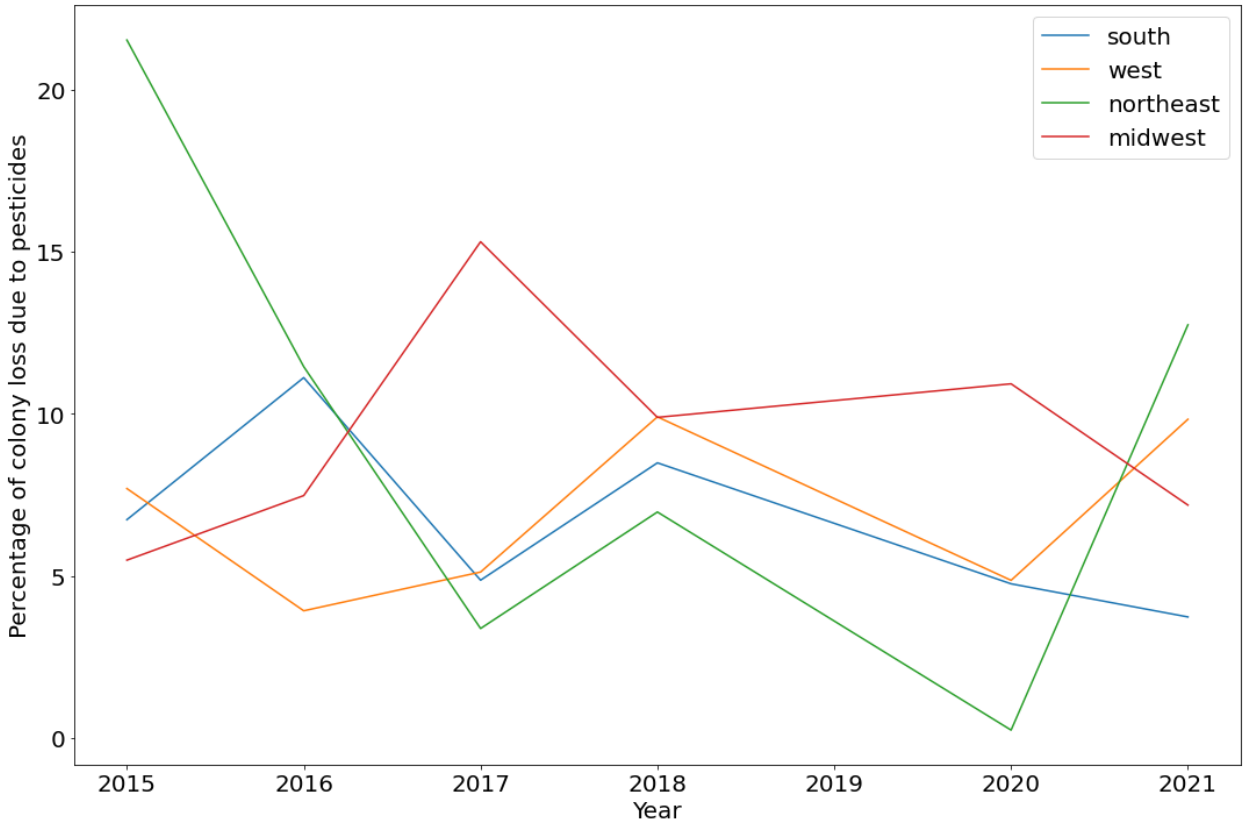
Exploratory Data Analysis

When examining the causes of colony loss, varroa mites outweighed other causes in all seasons. Of course, the data cannot reflect how hives weakened by pesticides or disease could be more susceptible to infestation of varroa mites.



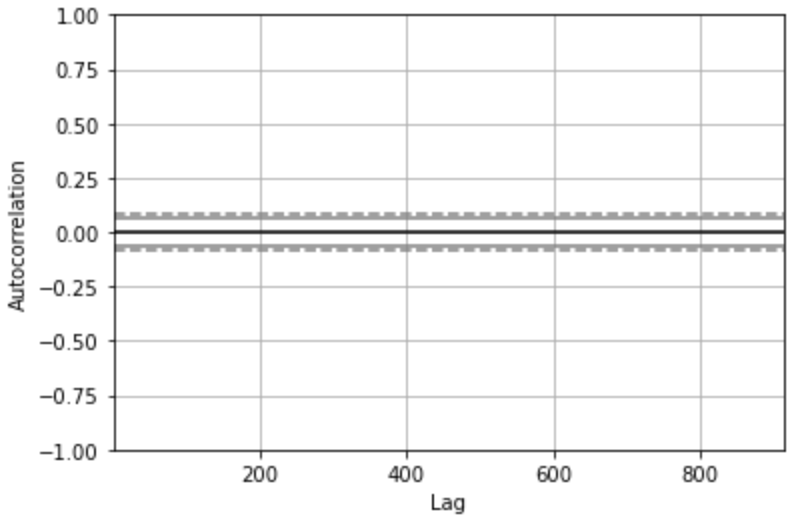
When examining specific causes of colony loss by region and quarter, no clear trends by region stand out. The first graph shows the percentage of colony loss due to varroa mites in Q2 (April to June). The second graph shows the percentage of colony loss due to pesticides in the same quarter.


south
west
northeast
midwest
Percentage of colony loss due to varroa mites
45
40
35
30
25
20
15
2015
2016
2017
2018
2019
2020
2021
Year



Model Selection

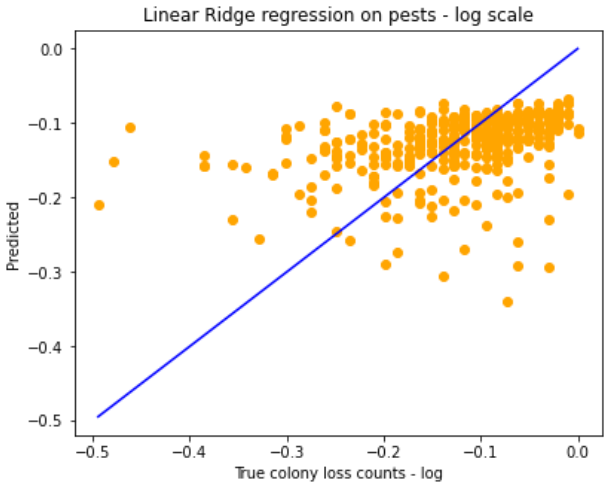
The initial plan was to do time series analysis using SARIMA, allowing a prediction of causes of colony loss in coming seasons. However, the data was too sparse for ARIMA and SARIMA models to work. On running the autocorrelation plot to determine the required number of AR terms, the autocorrelation came out as uniformly zero, meaning the starting value for the p and d variables would also be zero.

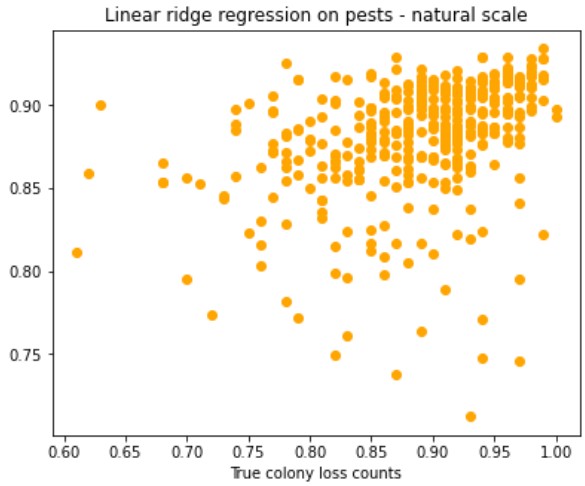


Instead, a series of regression models was used to determine what variables were the best predictors of colony loss. Multiple parameters were fitted to each of three regression models: linear ridge regression, random forest regression, and support vector regression. Multiple sets of predictors were used for each model. These included region data alone, subregion data alone, all loss reasons, varroa mite data combined with region data, and a Patsy matrix combining varroa mites, pesticides, and region data.

Linear Ridge regression

The two best linear ridge regression models that provide value in their predictive abilities are the model of all the pest data, and the model using the Patsy matrix of varroa mites, pesticides, and region data. Of the two, the model using all the colony stressor data as the predictor is the better model in terms of metric values. The difference between the train and test data indicate that the model may be slightly underfit. However, the Patsy model of varroa mite, pesticide, and region data has closer alignment between the training and testing sets, indicating that the consistency of the Patsy matrix as predictor may be higher.



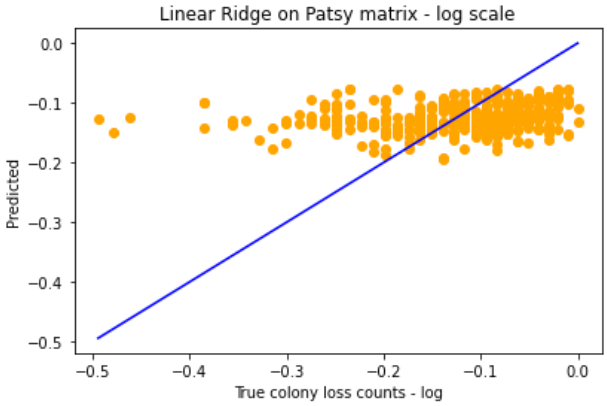


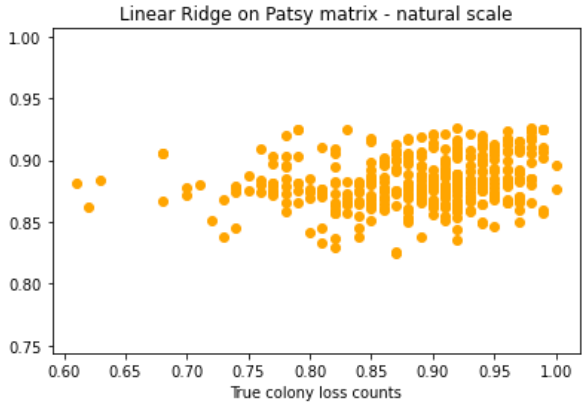
Metrics for all colony stressor data:

MAE on training data: 0.045 MAE on test data: 0.046

R-squared on training data: 0.125 R-squared on test data: 0.095

RMSE on training data: 0.095 RMSE on test data: 0.079





Metrics for Patsy matrix of varroa mites, pesticide and region data:

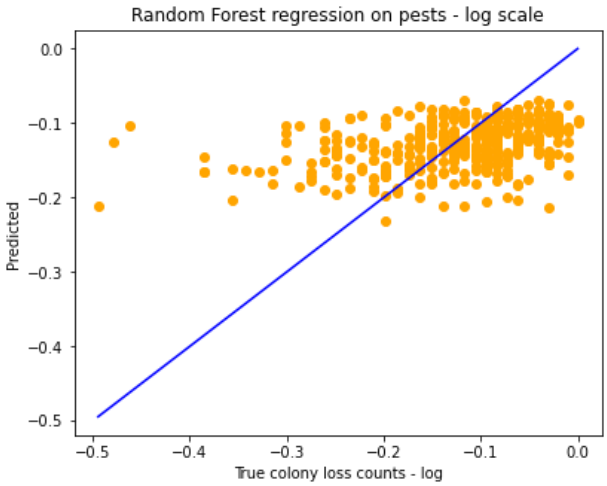
MAE on training data: 0.049 MAE on test data: 0.049

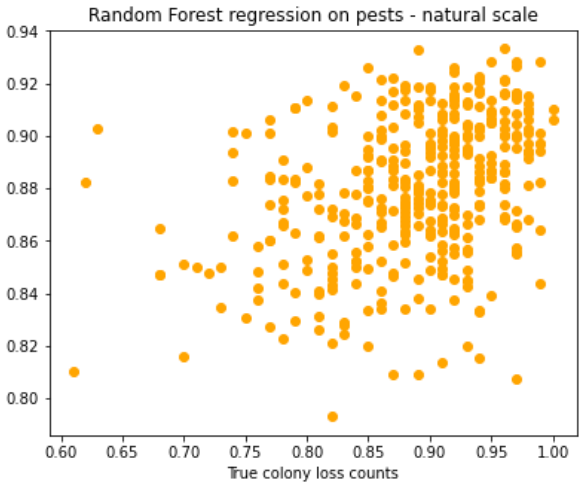
R-squared on training data: 0.058 R-squared on test data: 0.042

RMSE on training data: 0.099 RMSE on test data: 0.081

Random Forest regression

The best random forest regression model that had good predictive abilities was, like linear ridge regression, the one built using all the colony stressor data. The RMSE comparison between the train and test data sets showed slight underfitting, while the comparison on MAE showed slight overfitting.





Metrics for all colony stressor data:

MAE on training data: 0.045	MAE on test data: 0.047
R-squared on training data: 0.150	R-squared on test data: 0.132
RMSE on training data: 0.094	RMSE on test data: 0.077

Findings

Across the two models that performed well, linear ridge regression and random forest regression, using all the colony stressors (varroa mites, other pests, diseases, pesticides, other, and unknown) as the independent variables provided the best predictive qualities of the dependent variable of colony loss. In one case, using a Patsy matrix of varroa mites, pesticides, and region data proved to also have good predictive qualities, but not quite as good as the colony stressor data alone.

Further, when examining the feature coefficients yielded by running OLS with the statsmodels package, the state data fluctuates wildly across states. Neighboring states that would be expected to have similar feature coefficients, like Minnesota and Wisconsin, instead have very different coefficients: Wisconsin with 52.3267 and Minnesota with 6.0684.

OLS Regression Results

Dep. Variable:	percent_lost	R-squared:	0.332
Model:	OLS	Adj. R-squared:	0.294
Method:	Least Squares	F-statistic:	8.602
Date:	Sat, 12 Mar 2022	Prob (F-statistic):	2.33e-46
Time:	18:56:29	Log-Likelihood:	-2807.2
No. Observations:	878	AIC:	5712.
Df Residuals:	829	BIC:	5946.
Df Model:	48		
Covariance Type:	nonrobust		

Recommendations and Future Research

Even though colony collapse disorder has been recognized as an issue using that title since 2006, when attrition climbed dramatically, the USDA has only been keeping data on bees since 2015. The data is poorly collected, with missing data for colony stressors and inconsistent math. The strongest recommendation would be to collect data for more years and maintain a more robust data collection strategy.

Future research suggestions:

With more robust data collection, a time series analysis could be conducted, allowing beekeepers to focus their loss prevention efforts on most likely causes of colony loss.

Combining this data with temperature and rainfall data would be illuminating. A very wet spring, an extended winter cold snap, or a mid-summer drought can all have considerable effects on colony health and ability to withstand disease and predation.

Since colony stressors can combine to cause the death of a colony, listing a secondary cause of loss, when present, would be useful. For instance, even if the final cause of colony loss is varroa mites, the hive may have been weakened earlier in the season from pesticide use.