ALCM Report

Rodelyn Jaksons

07 January, 2020

## Data pre-processing

The trap counts for each orchard were aggregated to weekly counts based on the calendar week in which it was collected. As the number of ALCM is proportional to the effort, the number of traps laid in that week were also accounted for. Therefore the capture rate for each week was also analysed. The capture rate is defined as:

Where is the capture rate for the week, is the number of captures for the week and are the nuber of traps laid.

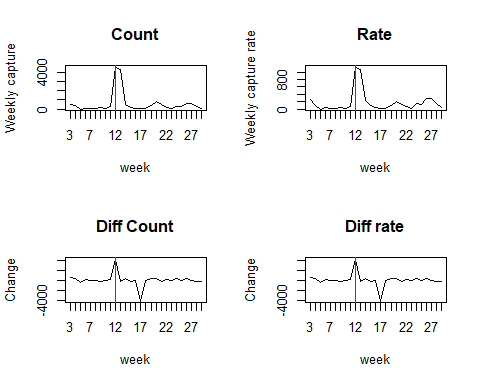
A start date of the 1st of July was used to calculate growing degree days with a base temperature of 7C. The models also used meterological predictors such as average weekly temperature and total weekly rainfall. The meterological data was obtained from HortPlus.

As the data is varying temporally, an important predictor is what occured previously and is known as a lag predictor. However, the use of a lag predictor means that one is only able to forecast one week ahead. Therefore to cicumvent this we difference the series to obtain stationarity. Stationarity is a data transormation technique so that the observations become independent of time. By doing so, we are modthe changes in the following week from the previous week. Therefore, large positive changes will show the peak timings of emergence.

The differenced series for the trap captures is defined by

and for the capture rate:

Th differencing is demonstarted for the Palmer orchard in the 2017-18 growing season. In these plots, the first column is of the total weekly trap capture, while the second column is the weekly rate capture. The first row is of the raw data, while the second row depicts the differenced series. It can be seen that the peak occurence is unaffected by the differencing.



The difference series for each orchard was normalised using each seasons sample standard deviation. Normalisation makes the trap captures between the different orchards easier to compare. The differenced series for the trap captures is defined as

and is defined as the following for the capture rate:

## Methods

We investigated a number of machine learning algorithms to see if a one algorithm would be best to for prediction, or if an ensemble of models would be better for forecasting. The models comprised of regression trees, random forests, support vector machine (SVM) regression, and gradient boosting machine (GBM).

### Regression trees

Regression trees recurively partition the variable space, thereby creating a set of rectangles. In each partition a linear regression model is fitted in each one. The resulting tree is based on the different partition combination which best describes the observed process. Regression trees have been successfully used across many disciplines for modelling and prediction. However, regression trees are known to suffer from instability, as small changes in the data can yield different trees.

### Random forest

Random forest, are extensions of regression trees and addresses the instability that a single tree faces. In a random forest model, the same tree is fitted to bootstrapped sample versions of the training data. The predictions are based on a average of many trees.

### Support vector machines

SVMs uses hyperplanes for prediction. In an SVM, the hyperplane acts as a classifier which best describes the data, thus separating the observations into their respective groups.

### Gradient boosting machine

GBM’s are similar to random forests. However, unlike a random forest where a many trees are bulit independent of each other, GBM’s build each tree sequentially. In a GBM the aim of each new tree is to improve the previous tree’s weaknesses.

## Model Evaluation

To evaluate model performance we use cross validation. In cross validation, the data set is partitioned into a traning set and testing set. The training set is used to build the model, where the model is then used to obtain the prediction error in the testing set.

The cross validation error rate for validation is given by

where is the test fold and is the number of folds.