ALCM Report

Rodelyn Jaksons

09 January, 2020

## Data pre-processing

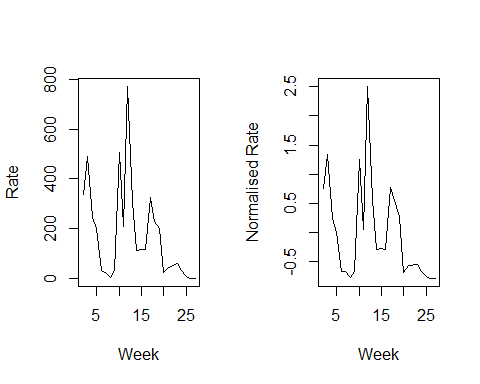
### Trap captures

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

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

To make the capture rates comparible between orchards and growing seasons, the capture rate was normalised. The normalised capture rate is given by:

The normalisation is demonstrated for the Nelson 2018-19 growing season. The first row, shows the raw capture rate, while the second row is of the normalised capture rate. It can be seen that normalisation leaves the peaks unaffected.



### Predictors

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 varies temporally, an important predictor is what occured previously, which is known as the 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 use historical data to find the average of capture rate of the week question, and use it as a predictor in the model.

## 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 recursively 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. In our case the folds were defined by season. The model performances were also inspected visually, to see if they predicted dynamics could reasonably explain the data.

## Results

The table below shows the cross validation error rate for the different models by region, where the lower error is preferred. For Central Otago and Hawke’s Bay, the random forest and GBM performed the best. For the Nelson region the SVM and GBM models were superior.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Region | Regression Tree | Random Forest | SVM | GBM |
| Central Otago | 45.4 | 38.2 | 45.5 | 38.1 |
| Hawke’s Bay | 34.4 | 26.8 | 28.3 | 26.2 |
| Nelson | 38.6 | 37.1 | 34.5 | 33.3 |
|  |  |  |  |  |

The plots below show the performance of each model. The black lines are the observed captures, while the red are the model predictions. In general the models can adequately describe the emergence dynamics.

