Data Science Project: Final Report

Application of Decision Theory and Machine Learning to Wild Pollinators in Apple Orchards

Author

Stephen Brownsey

Supervisor

Dr Julia Brettschneider
Department of Statistics

May 9, 2020

Abstract

Keywords:
Decision Theory, **Machine Learning**, **Clustering**, **Pesticides**, **Wild Pollinators**, **Bees**

Wild bee populations are in decline globally. In the original research (Park et al., 2015), the effects of surrounding natural landscape on wild pollinators was investigated. In this project the data were analysed from a different perspective, with the aim of forming a basis for better collection methods and statistical methodologies for the future. The effectiveness of machine learning algorithms in reducing the need for domain knowledge when deciding upon protocol clustering was evaluated and the ecologist/economist trade off was analysed from a theoretical decision theory perspective.

Results: Machine learning accurately clustered orchards using pesticide protocols to demonstrate the harmful effects of pesticides on wild bee abundance and richness. Using raw pesticide values resulted in better clustering outcomes than using standardised data. Standard machine learning approaches were shown to be superior over decision theory approaches, with analyses highlighting the need to adjust standard algorithms to cope with low numbers of observations and combat sparsity.

The theoretical decision theory applications demonstrated that in future, with better collection methods, an ecologist vs economist perspective could be investigated to determine optimal trade-off strategies for various decision rules.

Limitations included the lack of data points in the original report and the absence of specific date information.

Contents

1 Acknowledgements

The project used data collected in a previous study by Mia Park et al (Park et al., 2015), without this original research the analysis covered in this dissertation could not have been undertaken.

I would like to thank my supervisor, Dr Julia Brettschneider, for her invaluable help throughout the project. Her teaching and passion for her research inspired me to pursue this project in the first instance and changed my thoughts for career progression.

I would to thank the botanist, Dr Maria Christodoulou, for her insight into the agricultural industry, in particular the methods of data collection in apple orchards and general domain knowledge which she was able to share.

I would also like to thank my parents for the support they have always shown me.

This dissertation is dedicated to Ringo, Maya and Toto, without whose constant friendship, it may still be on the drawing board!

2 Introduction

2.1 Project Background

In this project, the data presented by Park et al. (2015), where the impacts percentage of natural area and conventional pesticide on wild pollinators was investigated, have been re-analysed from a different perspective.

On the whole, general linear models were used by the authors to display their findings, giving valuable insight into the benefits of natural areas for wild pollinators, and highlighting the importance of considering the landscape context when weighing up the costs of pest management on crop pollination services. In this project the effects of pesticides on wild pollinators were investigated, with the goal of analysing the relationship between them, using decision theory modelling. Adaptations were made to the original goals as more information became apparent, with the focus moving towards machine learning modelling with the actual data, whilst the decision theory applications were undertaken from a theoretical perspective (as summarised in Section 3).

Due to agricultural intensification over the past century, the conventional approach, by growers of the 19 conventional apple orchards included in the research article, was to apply multiple classes of compounds before, during, and after the orchard blooming season. Fungicides are usually applied just before or during bloom, when the rainfall is highest, to prevent the spread of fungal pathogens. Insecticides are applied after the flowering window to minimise the effects on pollinators. Thirdly, thinners are commonly applied after the 'June drop', when the trees naturally shed some fruitlets, to thin the crop, thereby avoiding branch damage and increasing fruit size. In the conventional setting investigated, pollination was by strategic placement of honeybee hives throughout the orchard. When this is the case, insecticides and thinners are only applied after the hives have been removed from the orchard. However, as the trees are rarely perfectly synchronised, timing of this application can only be approximate, affecting wild pollinators and their services. Understanding exactly how this affects the wild pollinators could lead to having a better grasp on how to maximise their use and longevity in the farming industry.

In the work presented in this project, machine learning algorithms were used to explore different clustering scenarios both from classical and decision theory perspectives. How these clusters affected the wild pollinators was analysed. The outcomes were developed further, enabling models to be fitted to the data which took into account confounding variables. These results were quantified. Each cluster represented a different orchard pesticide protocol, with the aim of finding those least harmful to wild pollinators. The knowledge gained was then used to build a prototype Application using RShiny (RStudio, Inc, 2013), which could be used by orchard owners to make more informed decisions when deciding upon their pesticide protocol. The quantitative results of the clustering were transformed into qualitative options for a user to select, with graphical outputs from these choices demonstrating the benefits to pollinators by choosing a less harmful pesticide protocol.

2.2 Why Bees?

The value of bees can be seen on a global scale. You may ask 'Why is this?' and certainly, relevant questions to ask include:

Why are bees so important?

And why is it worthwhile to conduct the research on these pollinators?

The benefits of bees on agriculture are well-established. They pollinate around 80% of flowering crops which, in turn, are responsible for about 33% of the world's food supply ((Speake, 2019)). From an economic perspective the pollination services of bees are valued at over £100 billion each year (Reynolds, 2019).

There are already initiatives in place by organisations such as the White House, to try and help the honey bees. In particular, the effects on honey bees can be limited by both the location placement of the hives and application times of the pesticides((USDA, 2019), (Time, 2015)). Ensuring that the hives have been removed before the pesticides are applied can help limit the damage to their populations from these chemicals.

The effects of these pesticides on wild bees however, cannot be controlled in this fashion. Wild bees are in a precarious position with more than 700 bee species in North America headed towards extinction (Time, 2017). With the loss of these vital pollinators, the threat to global food supplies is very real. Anything we can do to demonstrate the negative effects of human intervention and thereby slow and eventually stop further damage, thus protecting our food supplies, has to be worthwhile.

The study is based in North America, however the decline to bees is seen globally (Reynolds, 2019).

3 Objectives

Two separate objectives were considered in the analyses presented in this project.

3.1 Objective One

The mathematical and statistical goal, as defined in the project specification, was to analyse the data and use the findings to build decision models in an agri-environmental context, as follows:

a) Organise the variables and conduct comprehensive exploratory data analysis,

b) Develop decision models to analyse how the choice of pesticides, and the order in which they are deployed, can impact on the richness and abundance of wild bees,

c) Fit the models using the available data,

d) Compare the results obtained by alternative decision rules.

It was initially anticipated that the original data provided would include apple yield information. Since this was not the case, it was necessary to change the statistical goals of the project. In the new approach, a slightly different slant on the data was considered. Whilst utilising similar statistical methodology, the aim of the new approach was to establish if machine learning could be used to reduce the requirement on domain knowledge when undertaking statistical analyses in the ecology and agriculture industries.

Currently, experts' views are elicited, when determining the critical values to be used in the agri-environmental sector. Demonstrating how machine learning can be used in this field would open up more areas for statistical investigation. This change was discussed and agreed upon with the project supervisor, Dr Brettschneider, and hence, the original objectives using decision theory were considered purely from a theoretical perspective (Section 14). In order to establish the most useful details for inclusion into the investigations, conversations were conducted with both the crop scientist and project supervisor. It was agreed that combining new data with existing data would confer little benefit, due to the lack of industry standards for data collection. The value of any additional data points would thus be counteracted by the variation in collection methodology. The decision was also made not to simulate data as this would be unlikely to provide a realistic view on the actual effects on the bee populations. The changes to Objective One are summarised below:

a) Unchanged,

b) Develop classical and decision theory machine learning models for the data, using the protocol variables,

c) Unchanged,

d) Compare the results obtained by alternative clustering methods.

3.2 Objective Two

The ecological goal of the project was to establish a relationship between the ecologists' priorities of high bee survival rates and associated high values of bee abundance and richness vs a land owner's priority of generating as much profit as possible from their orchard. As per above, the quality of the data meant this was not feasible. The decision was again reached to discuss the original objectives from a theoretical perspective (14) and the changes to Objective 2 can be seen below:

Original Objective 2:

a) Quantify the implications of different priorities such as high yield (land manager's perspective) vs biodiversity (ecologist's perspective), and suggest trade-off strategies. In an ideal world the strategies would lead to both a high yield of apples and minimal harm to wild bees,

b) Develop a user-friendly prototype for a web application that a grower could use involving the decision process, with variables that may be altered to allow visualisation of how varying decision criteria would affect both bees and crop yield.

Updated Objective 2:

a) Quantify the implications of different pesticide protocols versus biodiversity,

b) Develop a user-friendly prototype for a web application that a grower could use involving the decision process, with variables that may be altered to allow visualisation of how varying pesticide protocol would affect wild pollinators based on the clusters generated using machine learning as well as demonstrate key findings from the analysis.

4 Research

4.1 Introduction

Two main areas of research were undertaken to support this project. The first was background information on decision theory and machine learning algorithms, leading to an understanding of what approaches to take in an idealistic scenario (see Subsection 4.2). The second was understanding more about the original research, including how the data were collected and the methods implemented by the original researchers. This understanding was gained by interaction with the research botanist (see Subsection 4.3).

4.2 Background

Knowledge of machine learning was gained from year three modules, ST344; Professional Practice of Data Analysis and CS342; Machine Learning, as well as published books (Ah-Pine and Wang (2016), James et al. (2013)), which provided a supplemental understanding of the most appropriate methods to implement.

4.3 Original Methods

Key points from the original research (Park et al., 2015) are shown below. These conclusions were not checked as part of my exploratory data analysis since the point of this project was not to validate previously published work:

- In general honey bees were put into the orchards in rented honey bee hives. Honey bee abundance was driven solely by temperature, supporting the notion that the effects of pesticides on honey bees can be limited by careful timing and positioning of their hives,
- Wild bee communities were driven by year, temperature, and characteristics of the landscape and orchard management,
- Wild bee abundance and species richness increased with an increase in the percentage of natural area in the surrounding landscape,
- Wild bee abundance and species richness decreased linearly with increasing pesticide use 1 year after application,
- Note: Only female bee numbers were used in the analyses. This was due to the male bees making up only 10% of the overall wild bees sampled, and, when the original research was re-run with the male bee data included, the conclusions were unchanged.

Main learnings from interactions with the botanist:

- Fungicides, insecticides and thinners are used throughout the year to try and increase the apple yield of orchards. It is important to understand when and why these are

used. Fungicides are typically used before or during bloom, in particular before so they can be combined with copper products, thus enabling them also to act as bactericides. Insecticides are applied outside the flowering window so as not to affect the pollinators. Thinners are applied after the natural load drop, often referred to as the 'June Drop', with the aim of further reducing a tree's load (of apples), so as to minimise the chance of branches breaking and allowing the remaining fruit to increase in size. The honey bees can be somewhat sheltered from these effects, as the hives are usually removed before the application of thinners and further insecticides. Wild bees however cannot be protected from these applications. Understanding these various effects helped to provide ideas on what to look at in the exploratory data analysis (see Section 6).

- The number of bees were calculated using a transect of the orchard. The researcher walked down the transect with a large net, moving the net in a figure of eight motion, catching and containing the bees. Once the transect had been walked, the number of each type of bee was calculated. Note on methodology: since the flight speed of bees is affected by temperature, this also has an effect on the number of bees collected.
- In an ideal scenario the data would be split into two observation points: Before Bloom and After Bloom, where Blooming is defined as 50% of the petals in the orchard being in bloom. For each orchard, there would be one observation taken in the Before Bloom period and one taken in the After Bloom period (see Figure 1).
- There are two different ways of calculating bee count. Abundance, the quantifiable number of bees present, and Richness, given by the variety of bees observed, in number of different species observed.
- The air temperature affects a bee's flight speed.
- Other literature (Thomson, 2019) demonstrates the effects of long-term variation in pollinator abundance and the diversity in reproduction of a generalised plant.

5 Data and Measurements

5.1 Measurements

For the exploratory data analysis to be started, it is first important to understand what the variables in the data refer to and how they were calculated. The following bullet points summarise the methods used in calculations:

- The Bee Impact Quotient (BIQ), is the product of a pesticide's scaled toxicity and its plant surface half-life,
- Application Rate is denoted by the quantity (of pesticide) used per acre,
- Pesticide Use Index (PUI), is a modified version of Bee Impact Quotient summed over all pesticides used in the orchard, and can be calculated by:

$$\sum_{i=1}^{n} BIQ \cdot \%active\,ingredient_i \cdot application\,rate_i$$

- PUI was broken down into Insecticide Use Index (IUI) and Fungicide Use Index (FUI). The same formula is followed but instead of summing over all pesticides, they sum over all insecticides and fungicides respectively.

5.2 Variables

This section describes how the variable names and meanings are related.

- When referring to wild bees, this always relates to wild female bees,
- The only variable related to honey bees is 'apisAb', which is honey bee abundance,
- Regions are classified as Lake Ontario (LO), Geneva area (GV) and Southern Cayuga Lake (S)
- 'AbF' in a variable name means Abundance of wild bees,
- 'RichF' in a variable name means Richness of wild bees,
- In bee variables the name before the three bullet points above implies which type of bee is being referenced. For example, 'socialAbF ' is the abundance of wild female social bees,
- The .np on a variable name means it has been calculated with Phosmet removed (an insecticide which could contribute a disproportionate amount to IUI),
- The .pre, .blm and .pos mean the values were calculated at time points 'Before Bloom', 'During Bloom' and 'After Bloom' respectively,
- All variables containing 'eiqB11 ' relate to the PUI, as defined above. The letter after the 11 relates to whether it was insecticide or fungicide and no letter implies it was

the total pesticide level. For example, 'eiqB11.np ' is the Total PUI without Phosmet and 'eiqB11I.pos ' is the Insecticide PUI After Bloom.

There are other variables used in the dataset but these are either self-explanatory, not used in the analysis, or will be explained in detail as they are used.

6 Exploratory Data Analysis

6.1 Primary Objectives

The primary objectives of the exploratory data analysis were:

- To highlight the limitations of the data,
- To identify areas in which to focus the future plans for the project.

The approach for the decision models was decided after undertaking the initial data analysis. Of the exploratory data analysis undertaken, only a small portion is reported as the majority of the time spent involved understanding the variables and the impacts they would have on the statistical goals of the project. The additional analyses can be found on the GitHub page (Brownsey, 2019b).

6.2 Orchards

The orchards were situated in North America. However, although the longitude and latitude of them was included in the published dataset, feedback from the botanist made it was clear that this should not be included in the report, as it contained personally identifiable information. These data were presented as part of my original analysis, but any inferences made have not be be included in this final report.

It can be noted that 16 orchards were visited in the first year and 19 in the second. This was due to a protocol design change in the original research. In the first year, not all the orchards were visited on two visit occasions. In the second year, the requirement for more data resulted in more orchards being visited at both time points. The visits are summarised in the following table:

Table 1: Number of Visits by Day and Year

Year and Day	Total Orchard Visits
Year 1, Day 1	16
Year 1, Day 2	7
Year 2, Day 1	19
Year 2, Day 2	19

From conversations with the botanist, it was known that blooming percentage of the orchard plays a large factor in the number of bees visiting the flowers and therefore the number caught during walks of the transect. It was therefore important to understand how these varied from Day 1 to Day 2 visits, and by different regions:

6.3 Exploratory Data Analysis

The Violin Plot of Bloom Index (see below) demonstrates how the distributions of the blooming values vary greatly in each region and on each visit day. In order to reduce misleading artifacts in the data, it would have been more ideal to have all Day 1 visits at a time with <50% blooms and all Day 2 visits at a time with >50% blooms (see Section **??**). Unfortunately there was little evidence that the Day 1 and Day 2 visits took place in relation to the bloom values as would have been hoped.

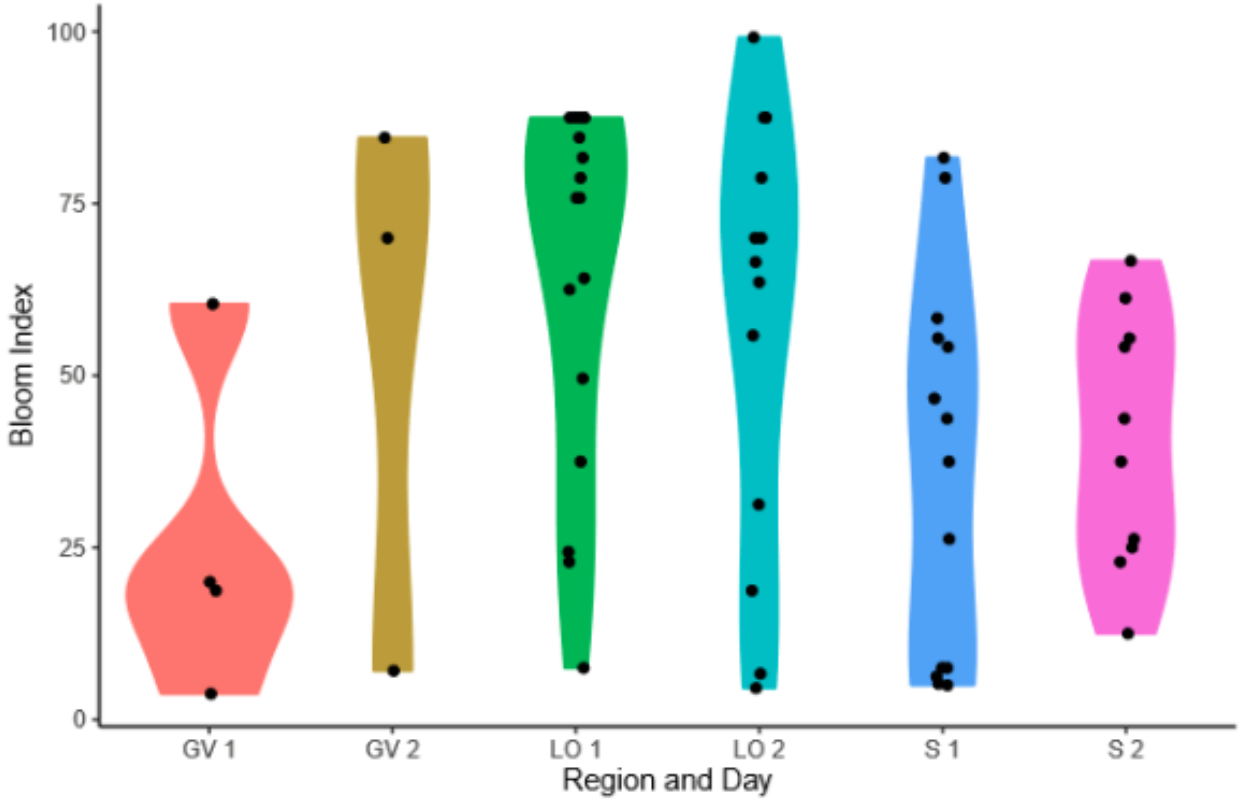


Figure 1: Violin Plot of Bloom Index by Region and Day

The distributions of all variables were investigated in order to understand the data and the effects the variables played on one another. The GGpairs function in R was utilised to summarise the interactions and distributions of the data as a whole. Violin plots, scatter plots and distribution plots, amongst others, were used to visualise the data. A specialised function was written using GGplot2 in R and then all the combinations of plot sought were passed in as inputs with the use of the lapply function. This ensured the code was reusable and clean. All the code used during EDA and the rest on the project can be found on GitHub Brownsey (2019b). Some example plots can be seen below:

Plot 1: A scatter plot of honey bee and wild bee abundance rating vs temperature, demonstrating a large variance in both honey bee and wild bee abundance.

Plot 2: A scatter plot of honey bee abundance (defined as the log of honey bee abundance + 1) vs temperature. The blue line is a Loess regression line with the 95% confidence interval shaded. The red line denotes a simple linear regression line. The plot demonstrates a non-linear association between honey bee abundance and temperature.

Plot 3: A density plot of the IUI Post Bloom.

Plot 4: A histogram plot of the IUI During Bloom coloured by region (denoted in the legend).

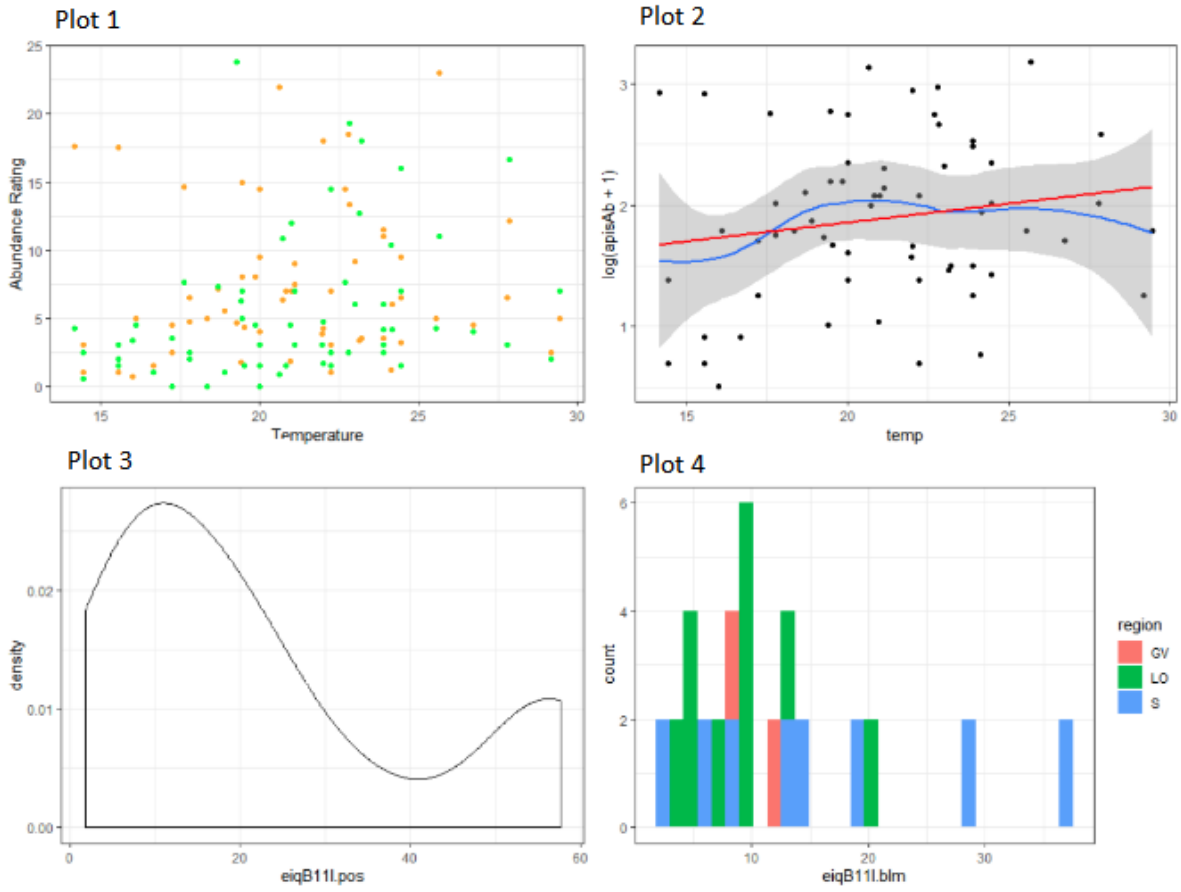


Figure 2: Example of Some EDA Plots Generated

Correlation between variables was also evaluated to establish if, and how, variables were related to one another. For example, below is the correlation table associated with all the bee variables, as defined in Section 5.2. In the variable column, Solitary is referred to as Sol and Social referred to as Soc.

Table 2: Correlation Table of Bee Variables

Variable	HoneyAb	WildAb	WildRich	SolAb	SolRich	SocAb	SocRich
HoneyAb	1.00	0.29	0.19	0.25	0.13	0.18	0.22
WildAb		1.00	0.91	0.96	0.92	0.29	0.31
WildRich			1.00	0.81	0.92	0.50	0.53
SolAb				1.00	0.92	0.02	0.05
SolRich					1.00	0.13	0.16
SocAb						1.00	0.99
SocRich							1.00

6.4 Discussion

The correlation table can be used to provide insight into the bee variables. Firstly, it can be noticed that the honey bee abundance is very lowly correlated to the other variables. This demonstrates that they are unlikely to be affected by the same variables, and backs up the previous research (Section 4.3) that honey bee abundance was determined solely by temperature. It is also heavily linked to 'hive.acr' , the density of honey bee hives per acre. The hives can also be removed to limit the effects of pesticides on their populations. For these reasons it was decided to not include honey bee abundance in our analysis. On the contrary, the exposure of wild bees to the six measured variables cannot be controlled by simply removing the hives. For the purpose of determining cluster performance, it was considered optimal to reduce the number of variables investigated. In order to do this, bee variables with a correlation greater than 0.9 were considered equivalent and only one variable was included. This subsetting method resulted in wild bee abundance and social bee richness being chosen as the two variables to be evaluated.

6.5 Main Outcomes

The main outcomes of the exploratory data analysis were:

1. Since the Year 2 data were more complete, it made sense to discard the Year 1 data for the models and analysis. In particular, as pesticide use last year affects bee count this year, using the incomplete data from the first year was likely to bring more inconsistencies to the analysis than benefits. As there seemed to be an underlying protocol factor in the orchards, it was considered that an attempt to model the orchards with respect to this could be the best way forward,

2. An orchard protocol can be summarised as the pesticides used at each time point and the environmental features of that particular orchard. The decision model process in the main analysis was iterative and the models improved over time. The first round of models considered contained pesticide data only,

3. An orchard protocol is unique for each orchard. At each visit occasion for each orchard, all the orchard protocol variables remained unchanged,
4. For the main analysis, only the wild bees were considered as honey bee exposure was found to be somewhat controlled by the timing of when the hives are placed in the orchards,
5. The Blooming Indexes were not consistent with what would have been hoped for. The Indexes for both days had a large variation and were not Day 1: $<$50% bloom and Day 2: $>$50% bloom, which would have reduced the variability and been more ideal,
6. The bee variables were highly correlated. From this it could be noted that by just looking at WildAb and SocRich, all other wild bee variables were highly correlated with a correlation score of greater than 0.9 to one of these two.

7 Machine Learning Models

7.1 Introduction

This section describes the work undertaken on Objectives 1b) - 1d), including the iterative process involved with deciding on the optimum approach to data clustering. Firstly, from the exploratory data analysis undertaken (Section 6), three time points were identified: 'before bloom', 'during bloom' and 'after bloom'.

As per the specification change 3, since no yield data were present, the shift was made to a machine learning implementation.

Further to interactions with the botanist, the decision was made to take a data driven approach when determining the clusters as she was unaware of any existing industry standards on the evaluation of pesticide levels.

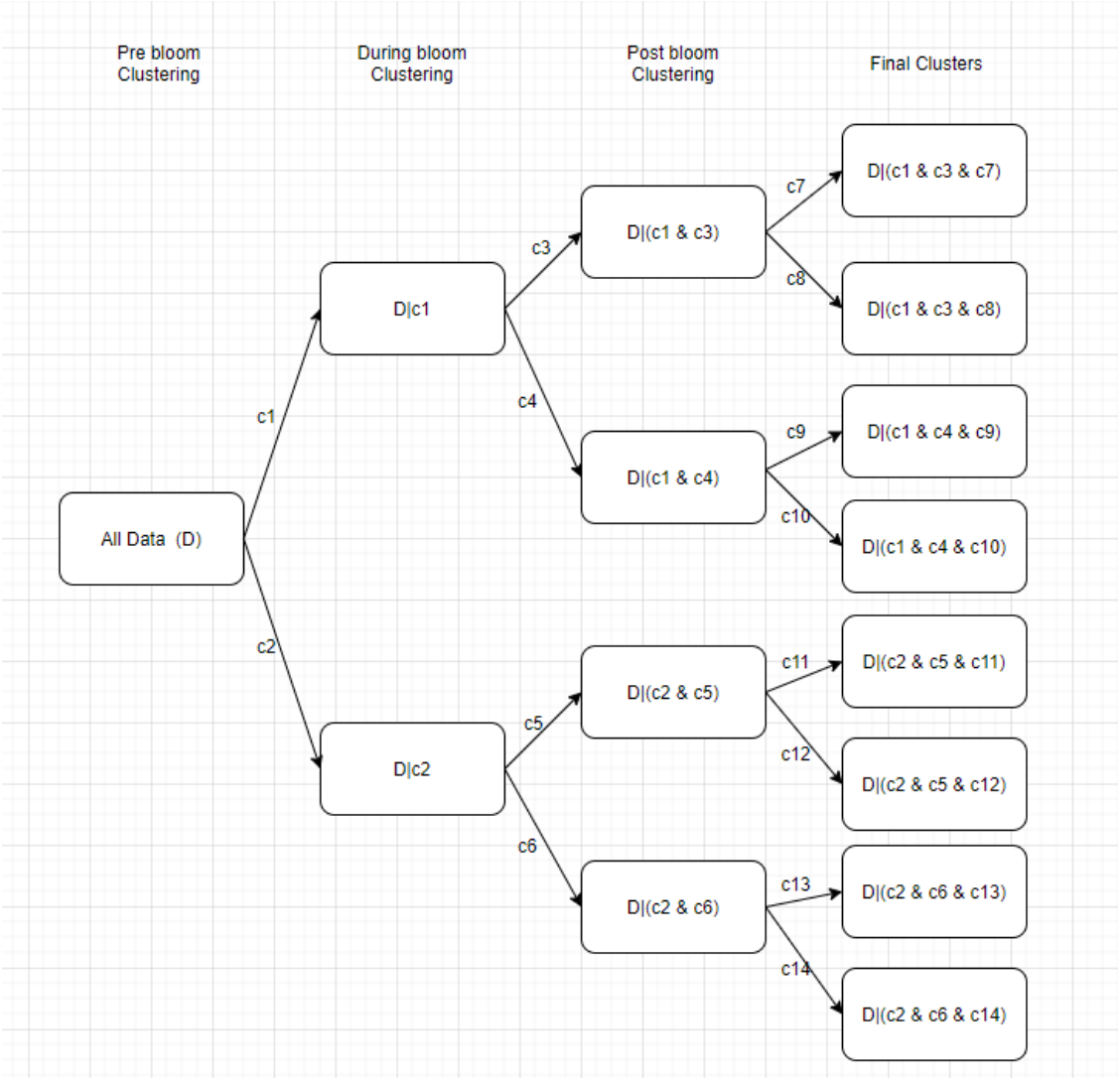
7.2 Decision Theory versus Classical Approach

The decision theory models were clustered at each time point based on the subset of the pesticide protocol relevant for that particular time point. For example in the 'before bloom' time point, the only two variables considered were before bloom fungicide and insecticide. For the 'post bloom' time point, the post bloom values for fungicide, insecticide and thinner were used to cluster the orchards. The groups are iteratively re-clustered at each time point to give the final clusters after the post-bloom clustering has been run.

Figure 3 below, depicts the decision theory approach. The dependence of each time point on the previous time point can be clearly seen. The data is subsetted at each step as it is propagated through the decision tree structure.

Conversely, the classical models consider all time point levels as one and apply the clustering to all the protocol information at once. The number of clusters were chosen using the nb_clust function in R (Documentation, 2019), which runs a maximum of 26 optimum clustering metrics and displays the summarised output from these methods. Using this method is similar to using ensemble methods in machine learning. In principle by taking an aggregation, in this scenario looking at the mode number of clusters, there is less chance of choosing a unsuitable cluster number to model the data, than if just one metric was considered.

Figure 3: Decision Theory Applications to Clustering



7.3 Why Machine Learning?

In general, analyses undertaken on similar agricultural topics have so far used manual decision-making to select the cut-off points, with low/high thresholds set for each variable and then data in each branch subsetted based on these values. Use of machine learning on such a small dataset is generally not recommended, nor indeed a good implementation of the concepts. However, the aims of this project are to see if it can be used to provide a good clustering of the data and in turn be used to find optimum splitting criteria if the datasets were to be expanded and analysis taken further. The primary aim therefore was to find an optimum clustering method, so that going forwards with additional data, the methodology would already have been tested.

One particular downside to the approach was that there were so few data points, it was not possible to split the dataset into a training dataset and test dataset. The final outputs were particularly sparse, with some clusters containing only one element, meaning that the accuracy would not be good if some data points were kept back for testing purposes. If this analysis was repeated with a larger dataset, then using both training and test datasets would allow for the optimum method to be classified more easily.

7.4 Naive Kmeans Approach

The first approach was to simply cluster the data by each time point according to the protocol, as defined by the pesticide values at that time point. This was achieved by the passing the protocol parameters into the Kmeans clustering algorithm. Once clustered, the orchards could be put into their respective end clusters based on the combination of how each time point was clustered. The reason for doing this was to very quickly get a visual aid as to how the number of nodes expanding from each time point influenced the cluster choices and sparseness of empty clusters. See Section 2 (Objective One) for the reasoning involved with this decision.

The Kmeans clustering method can be summarised as follows:

1. The number of clusters, K, to be used is chosen.
2. Centroids are initialised by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. The algorithm then iterates through until there is no change to the centroids.
4. Compute the sum of the squared distance between data points and all centroids. Denoted by:
$$F = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||^2$$
Assign each data point to the closest cluster (centroid).
5. Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.

This output can be taken and put into a tabular format, which shows the decision tree branches based on the degree of pesticide use. So far, three scenarios have been modelled, each of which has 19 data points to share amongst all the branches. The main aims of this naive approach was to quantify how many branches should be used at each step and provide insight to the later more in depth analysis methods.

Scenario One: This uses a decision tree with two branches at each step. Of the eight possible end points, five contain orchard data. This is shown by the following table (Table 3) and decision diagram (Figure 4). The same information can be extracted from the tables in this and following scenarios. The decision tree image has only been included for the first example for the purpose of visualisation as for the others, the diagrams are very sparse and can be more easily understood from a table.

Table 3: Scenario 1: Clustering by Time Period

Before Bloom Cluster	During Bloom Cluster	After Bloom Cluster	Number of Orchards
1	1	1	1
1	1	2	2
1	2	1	9
1	2	2	3
2	2	1	4

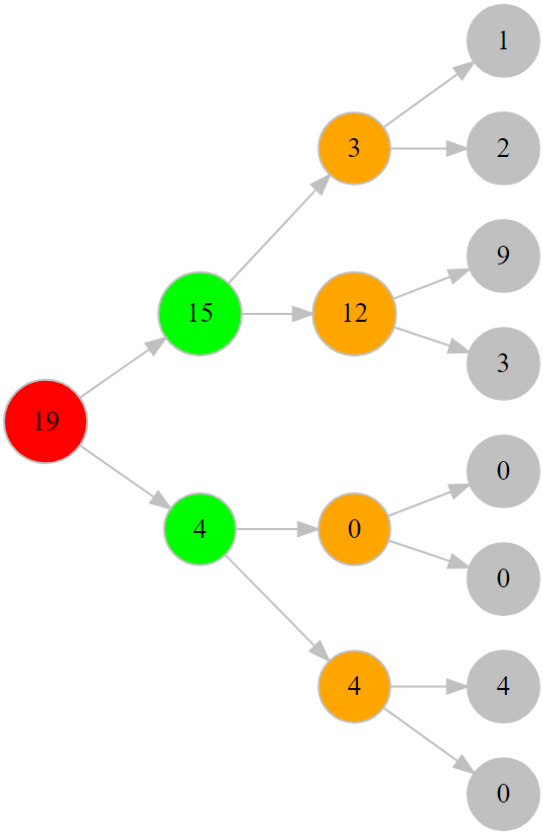


Figure 4: Splitting with 2 Nodes at each Time Point

In the diagram, the red circle denotes the number of orchards before the decision process is started, the green circles denote the number of locations of orchards at the 'Before Bloom' stage, and the yellow and grey circles represent the number of orchards at each of the 'During

Bloom' and 'After Bloom' clusters respectively.

This figure demonstrates how cluster numbering will be referred to for the decision theory approach throughout this dissertation. For example, cluster 1 would refer to the grey circle at the top with 1 observation, whilst cluster 3 would refer to the grey circle with 9 observations.

Table 4 presents a summary of the mean pesticide values in each cluster. Looking into the fungicide and insecticide values of each cluster, the differences between them can be observed. Each row of the table denotes a different cluster.

Table 4: Scenario 1: Cluster Pesticide Protocol Summaries

fung_pre	insect_pre	fung_blm	insect_blm	fung_pos	insect_pos
111.67	0.00	127.30	8.87	53.11	5.25
138.62	1.16	107.91	8.12	78.41	31.89
127.38	0.15	49.62	11.80	22.58	15.06
161.57	0.00	53.16	9.36	31.37	57.66
22.61	1.68	42.50	17.30	21.20	14.04

Scenario Two: This uses a decision tree with four branches at the first step and two branches at each other step. Of the 19 possible end points, eight contain orchard data (see Table 5).

Table 5: Scenario 2: Clustering by Time Period

Before Bloom Cluster	During Bloom Cluster	After Bloom Cluster	Number of Orchards
1	2	1	4
2	1	1	1
2	1	2	1
2	2	1	6
3	1	2	1
3	2	1	1
3	2	2	3
4	2	1	2

Table 6 contains the mean values of each fungicide and insecticide variable for each cluster, for Scenario 2.

Table 6: Scenario 2: Cluster Pesticide Protocol Summaries

fung_pre	insect_pre	fung_blm	insect_blm	fung_pos	insect_pos
22.61	1.68	42.50	17.30	21.20	14.04
111.67	0.00	127.30	8.87	53.11	5.25
113.28	0.00	102.91	3.61	68.14	11.92
102.18	0.16	43.57	12.87	24.77	11.42
163.95	2.31	11.90	12.62	88.69	51.87
143.09	0.00	58.10	9.01	14.48	24.31
161.57	0.00	53.16	9.36	31.37	57.66
195.11	0.19	63.51	4.94	20.06	21.36

Scenario three: This uses a decision tree with three branches at each step. Of the 27 possible end points, only ten contain orchard data (Table 7). Fungicide and insecticide variables are presented in Table 8.

Table 7: Scenario 3: Clustering by Time Period

Before Bloom Cluster	During Bloom Cluster	After Bloom Cluster	Number of Orchards
1	2	1	2
1	2	3	1
1	3	1	1
2	1	2	1
2	2	1	3
2	2	3	3
3	1	1	1
3	1	2	1
3	2	1	5
3	3	1	1

Table 8: Scenario 3: Cluster Pesticide Protocol Summaries

fung_pre	insect_pre	fung_blm	insect_blm	fung_pos	insect_pos
29.94	0.00	57.38	13.68	26.41	5.96
9.38	0.00	37.44	36.36	24.39	34.75
21.38	6.74	17.79	5.49	7.57	9.48
163.95	2.31	112.90	12.62	88.69	51.87
177.77	0.13	61.72	9.64	18.20	22.34
161.57	0.00	53.16	9.36	31.37	57.66
111.67	0.00	127.30	8.87	53.11	5.25
113.28	0.00	102.91	3.61	68.14	11.92
101.07	0.19	52.18	15.09	21.35	13.34
107.72	0.00	0.56	1.80	41.85	1.80

The analysis undertaken here was very useful as it clearly demonstrated that using more than two branches at each time point (Scenarios 2 and 3) resulted in sparse cluster representations. It also demonstrated one of the main shortcomings of the Kmeans algorithm, which is that the clustering outcome is not the same each time the code is run. The reason for this is that the Kmeans algorithm selects the initial cluster centroid positions at random. This means that the cluster number is somewhat arbitrary and if more data were added and the clusters re-run, the output clusters would be different.

The simulations run so far seemed to show that there were a few 'common' cluster outputs which occurred with increased frequency. The outcome from this analysis therefore suggested either the use of a different algorithm or using cross-validation to select an optimal Kmeans clustering output by running thousands of simulations. The methodology for this is covered in Subsection 7.6.

7.5 Future Methodology

In order to take the approach forwards from the naive approach, the following changes were applied to analysis sections 7.6, 7.7. Firstly, instead of just clustering over the whole dataset and combining the outcomes, the data were re-clustered at each step in the decision process to ensure the clusters were chosen as accurately as possible.

In addition, when clustering data, it is common practice to standardise the predictor variables, in this case pesticide values, before clustering takes place. This allows for each pesticide index to have an equal affect on the clustering of the data. When looking at the clusters, both standardised and non-standardised approaches were considered. A full explanation can be found in Section 9.1.

Cross-validated Kmeans were investigated to see if this approach could provide better information regarding the clustering of the data over the naive Kmeans approach. One further method evaluated was Agglomerative Hierarchical Clustering. From the literature read (Ah-Pine and Wang (2016), James et al. (2013)) this seemed like an appropriate fit.

Due to the sparseness of clusters, the final method considered was a look at the protocol as a whole, ignoring the time point decision tree approach. The aim of this approach was to reduce the number of potential clusters and thus lead to a different view. In particular, optimum cluster detection methods could be used for this, such as the elbow method Wikipedia (Wikipedia) and the NbClust function in R (Documentation, 2019).

7.6 Cross Validated Kmeans

Previous analysis using the Kmeans algorithm demonstrated that it returned different clustered scenarios each time it was run (Section 6.4). This was expected as the algorithm selects the initial cluster centroid positions at random. When the previous analysis was run how-

ever, it was clear there were a few common clustering combinations that were outputted more frequently. In this section, a description is given of how cross validation was used for the identification of the most commonly-occurring cluster labels across all the algorithm runs, in the hope that this could allow the changes between adding in new orchards to be seen clearly.

This process takes the Kmeans methodology (Section 7.4), runs it a thousand times and stores the outputs in a dataframe. The row-wise modal cluster is then calculated as the final cluster output.

Even when running a thousand simulations and selecting the most common output, the clusters were still different when the simulations were re-run. This algorithm was therefore not performing in a way that would allow orchard data to be added and clusters re-run to only show the difference in output from the new orchard data. Clearly this feature is a requirement for orchard protocol clustering if accurate decisions are to be drawn.

7.7 Agglomerative Hierarchical Clustering

Previous research undertaken, (James et al. (2013)), demonstrated that hierarchical clustering is often an improvement over Kmeans, as this algorithm does not require a pre-specified number of clusters and can produce a tree-based approach which is more in keeping with the decision theory aspect of the project. It also has the benefit of being visualised as a dendogram. Agglomerative Hierarchical Clustering (AHC), also known as bottom-up clustering, was chosen as it was a good fit for the data. It was selected over Divisive Hierarchical Clustering, in particular due to its relative computational efficiency (Ah-Pine and Wang (2016)) as well as the assumed shape of the data, since there are smaller sub-categories of pesticide contained within each pesticide application period. AHC works in a 'bottom-up' manner, so each object is initially considered as a single-element cluster, then at each step of the algorithm, the two most similar clusters are combined until there is only one cluster.

When using AHC it is necessary to choose both a distance metric and linkage function which is denoted mathematically as $D(X, Y) = min_{x \in X, y \in Y} d(x, y)$. The distance metrics considered were:

1. Average Linkage: The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.
2. Single Linkage: The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. In this case the merge criterion is local and the more distant parts of the cluster are not taken into account.
3. Complete Linkage: The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. Use of this metric tends to produce more compact clusters.

4. Ward's Method: This method minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

In the table below the higher the score the better the combination of these fits the cluster structure of the data:

Table 9: The Agglomerative Coefficients Calculated for some Linkage and Distance Methods for Non-Standardised Data

	Average Linkage	Single Linkage	Complete Linkage	Ward's Method
Euclidean	0.7206	0.5296	0.8173	0.8706
Maximum	0.7742	0.5496	0.8724	**0.8982**
Manhattan	0.6749	0.5486	0.8181	0.8358
Canberra	0.4298	0.3713	0.5215	0.5792

After the data has been standardised, the table becomes:

Table 10: The Agglomerative Coefficients Calculated for some Linkage and Distance Methods for Standardised Data

	Average Linkage	Single Linkage	Complete Linkage	Ward's Method
Euclidean	0.7295	0.7379	0.7522	0.7669
Maximum	0.7197	0.7398	0.7226	**0.8002**
Manhattan	0.7087	0.6991	0.7497	0.7557
Canberra	0.4845	0.4077	0.5920	0.6446

Ward's Method was chosen as the linkage method, due to its high score across every metric, for both the standardised and non-standardised pesticide protocol values. The best performing distance metric was the Maximum Distance metric, which provided the best fit to the data in the decision theory machine learning methods, and was therefore then used in the AHC function calls.

7.8 Classical Machine Learning Approach

The entire data set has only 19 data points. When clustering using the above methods, the end clusters contain very few data points and as such the values of the end clusters are heavily reliant on these few observations. This means that there is potential for a large bias to be seen. In this section, the classical approach to machine learning is presented for the dataset, with the protocols clustered as a whole, to highlight the main differences between this and the decision theory approach (Sections 6.6 and 6.7).

The methodology is similar to that described in Section 7.7, although this time both Maximum and Euclidean linkage functions were used with the the Ward Method distance metric. The reason for including both was that the scores were very similar and as the pesticide

variables take values in $x \in R_+$ it could be argued that a Euclidean linkage function would be a good fit for the data.

In this classical approach however, instead of clustering at each time point, the optimum number of clusters are calculated using the fviz_nbclust and NbClust functions in R. These calculate the optimum number of clusters for the data using a maximum of 26 optimisation methods. Then to select the number of clusters for each method, either the modal cluster number was taken or, if the top two varied by only one, the higher cluster number was chosen. The two images below (Figure 5 and Figure 6) demonstrate how the number of clusters was chosen, with the highlighted bar demonstrating the cluster number chosen. In both sets of graphics the Maximum linkage method is located on the left and Euclidean on the right:

Figure 5: Standardised Clustering Numbers

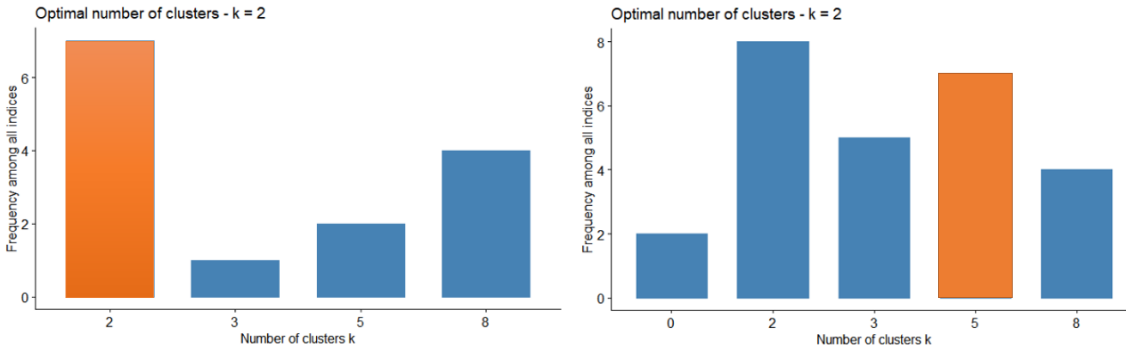
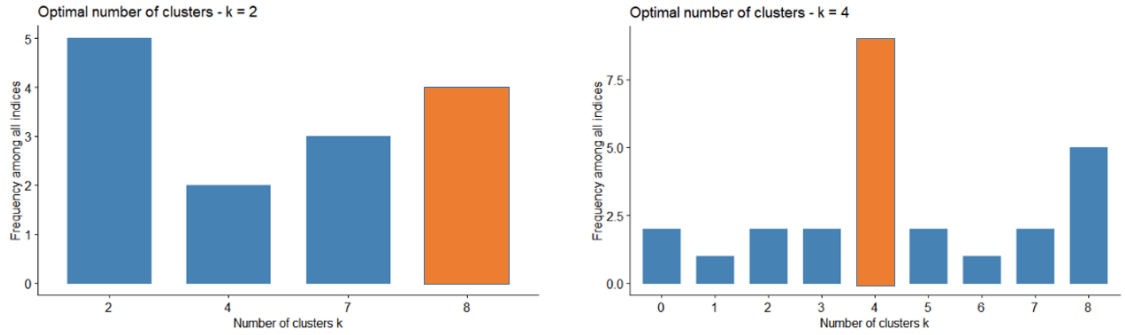


Figure 6: Raw Clustering Numbers



The outcomes from these figures are summarised in Table 11, below.

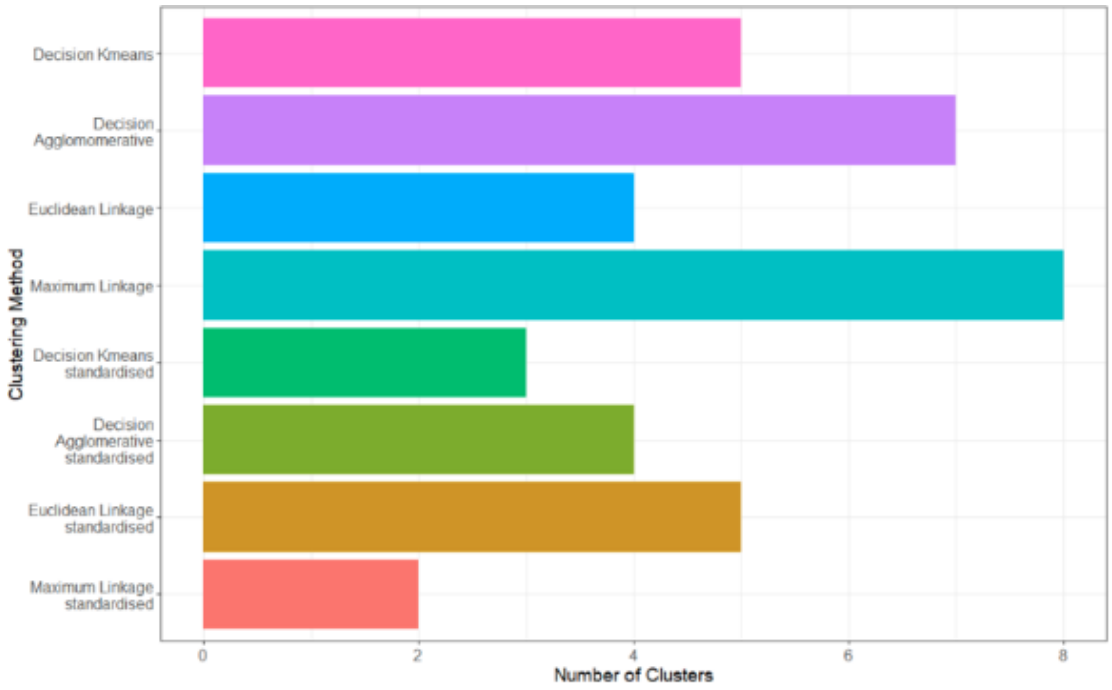
Table 11: Classical Machine Learning Cluster Numbers

Data Standardised	Linkage Function	Number of Clusters
No	Euclidean	4
Yes	Euclidean	5
No	Maximum	8
Yes	Maximum	2

7.9 Comparison of Clustering Numbers for Different Methods

To allow for easy comparison between different clustering methods, the following image summarises each method with the corresponding number of clusters chosen for analysis. When looking at the decision theory clusters, the number relates to the number of non-empty end nodes out of a possible 8.

Figure 7: Clustering Summary



A couple of key observations can be drawn from this summary, most notably, standardising the pesticide protocols results in fewer clusters being selected, in comparison to when the raw protocol values are used.

It is also apparent that using Kmeans clustering resulted in fewer clusters than agglomerative clustering for both raw and standardised data for the decision theory approach.

The Kmeans clustering method was deemed sub-optimal for the purpose of clustering orchard protocols, for the following reasons:

- The random centroid initialisation of clusters,
- Even when the final clusters were chosen using cross validation and a thousand Kmeans iterations at each stage, the end cluster results varied each time the code was run,
- It also resulted in fewer clusters being detected and as such provided less insight into the different effects of pesticides on bee populations,
- If the analysis were expanded to include more orchards, then more clusters would appear, it is likely that they would be of differing sizes, for which Kmeans does not perform very well.

For these reasons, Kmeans clustering was not considered in the model comparisons (Section 9) for the best clustering method, though the outputs were calculated and have been published as part of the application (Section 10).

8 Bee Adjustment

8.1 Confounding Factors

Now the clustering combinations of the data have been defined (Section 7), in this section the effect that confounding factors may have on the output is presented. Many variables have the potential to influence the bee count observed. To achieve an accurate representation of how pesticide clusters themselves affect the bee count, these confounding variables must be adjusted for each considered bee variable (Section 8). First the bee abundance values were considered. Statistical tests were performed on all potential confounding variables to test for statistical significance. This was compared with the statistically significant variables chosen using linear models. If they matched, then for bee richness only the linear model method was used. If they did not match then both methods were considered. The reason for choosing linear models was that due to the dearth of data points it was decided these would provide a good adjustment without the risks of overfitting the models to the data.

The previous research undertaken (Park et al., 2015) concluded that both temperature and surrounding percentage of natural area available to the bees (X2000nat), did have an effect on bee count. Other potential confounding variables considered were; Local Diversity, Day, Region, and Bloom.

From the exploratory data analysis undertaken (Section 6), it is known that a $log(x) + 1$ transformation accounts for a better fit of the data. The responses can be assumed to follow a normal distribution once this transformation has been made, as tested using a Shapiro-Wilk Normality test and shown below:

Table 12: Shapiro-Wilk Normality Test Results

Data Log Transformed	Shapiro-Wilk Test p-value
No	0.000009
Yes	0.6046

This demonstrates the effect of the log transformation and how it allows the response variable to now satisfy the normal assumptions. This is a requirement for some of the tests used in the next subsections. For this reason all counts were their respective logged bee abundance and logged bee richness values.

8.2 Temperature

The first potential confounding variable considered was temperature. Figure 8 demonstrates the clear positive correlation between temperature and bee abundance, backing up the claims from the original analysis.

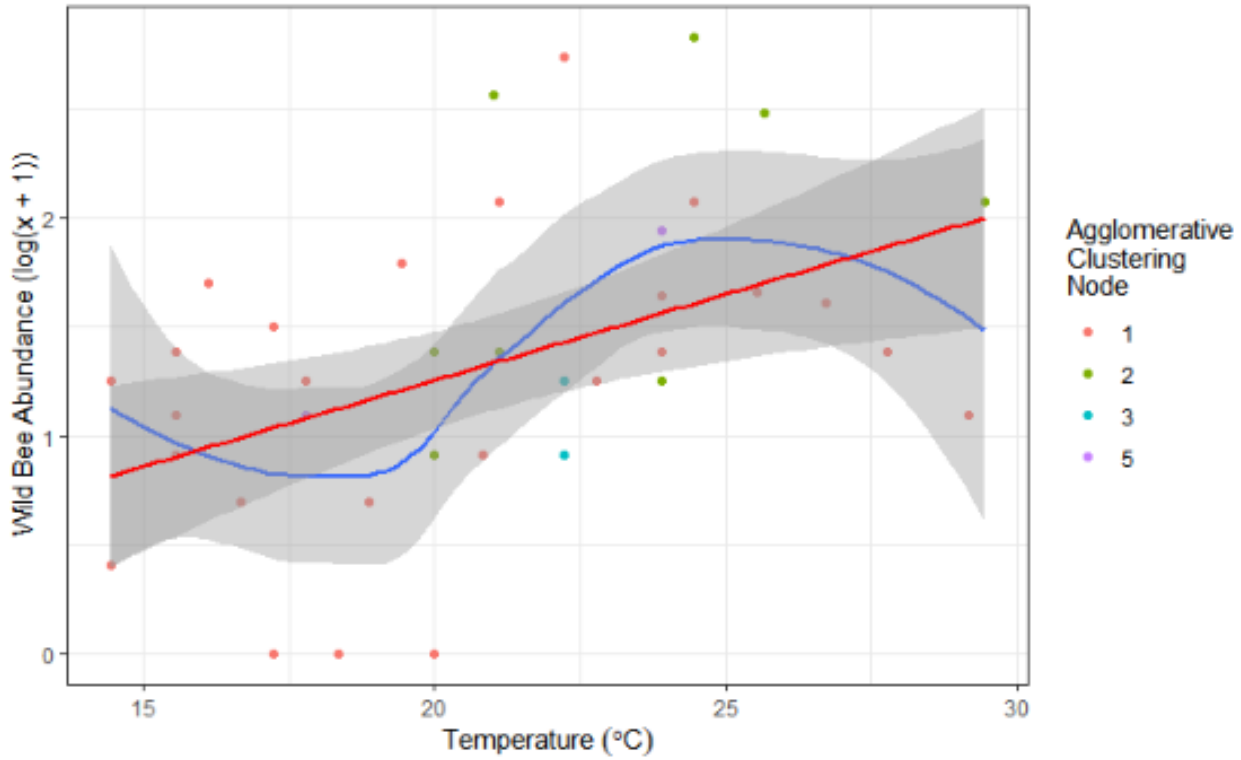


Figure 8: Temperature vs Bee Abundance

To give an idea on the effects of correcting for a confounding variable, there should be no obvious trend in the data. Instead points should vary randomly about a mean line. For the temperature variable, all abundance values were corrected as though the temperature was 20°C. This temperature was selected as it was close to the mean and median temperatures. This is denoted in the graph below:

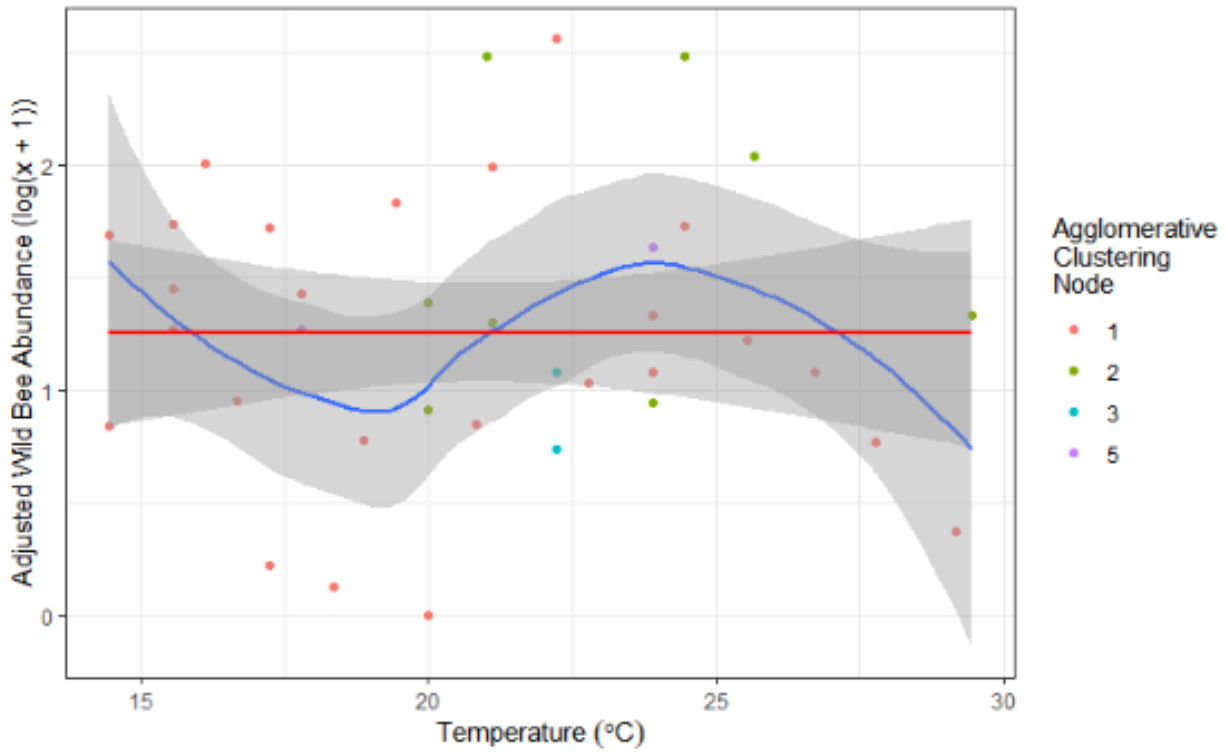


Figure 9: Adjusted Temperature vs Bee Abundance

Figure 9 clearly demonstrated that adjusting for temperature was beneficial, as it allows for the abundance values to be more related to the pesticide protocol used and independent of the temperature confounding factor.

8.3 Percentage of Surrounding Natural Area

The percentage of surrounding natural landscape available to the bees, given by the X2000nat variable, was also shown to have a positive correlation with bee abundance, as was seen for temperature (Section 7).

This showed similar trends to temperature and, after corrections were applied to the data, it was again shown that adjusting the data could provide the same benefits in removing dependence to confounding variables. One thing to note is that, if the data was iteratively adjusted in this manner, it would lead to the order of correction having an effect on the end adjustment. To ensure this is not the case, all the statistically significant confounding variables were detected, passed into a linear model and the parameter estimates from this model used to make all the adjustments from all the variables in one go (see Section 8.8).

8.4 Local Diversity

Local diversity was classified as either simple or complex, given by numeric variables of 0 and 1 respectively in the data. First an F-test was used to assess whether the variances were equal among the two diversity groups. The output from this was a test statistic of 1.17 and corresponding p-value of 0.725, showing that there was insufficient evidence to reject the Null hypothesis claim that the variances came from equal populations. A T-test was then applied to the data with the variance parameter set to equal. This returned a test statistic of -0.896 and corresponding p-value of 0.376 (see table X below).

Table 13: Statistical Test Outputs for Local Diversity

Test	Test-Statistic	p-value	Significant
F-test	1.17	0.725	No
ANOVA	-0.896	0.376	No

These outcomes demonstrate that there was insufficient evidence to reject the Null hypothesis, that local diversity does not have an affect on bee count.

8.5 Region

The region variable was classified as either Lake Ontario (LO), Geneva (GV) or Southern Cayuga Lake (S). Since there were three possible classifications, it was first necessary to use a Levene's Test to test the homogeneity of variance. This returned an F-value of 0.1954 and corresponding p-value of 0.8234, giving insufficient evidence to reject the Null hypothesis that the variances were equal among regions. Based on this an ANOVA test was used to assess whether there was a significant difference between the regions. The outcome of this was an F-value of 0.624 and corresponding p-value of 0.542 (see Table X, below).

Table 14: Statistical Test Outputs for Region

Test	Test-Statistic	p-value	Significant
Levene's Test	0.195	0.823	No
ANOVA	0.624	0.542	No

These outcomes demonstrate that there was insufficient evidence to reject the Null hypothesis that region has no effect on bee abundance.

8.6 Day

The data for each orchard were collected on two different days. Again both an F-test, for testing the variance assumption and a paired two sample T-test with variance equal parameter set to true, were applied. The outcomes from these tests can be seen below:

Table 15: Statistical Test Outputs for Day

Test	Test-Statistic	p-value	Significant
F-test	0.936	0.891	No
T-test	-2.02	0.0588	No

These outcomes show that there was insufficient evidence to reject the Null hypothesis that day has no effect on bee abundance.

8.7 Bloom

The first step in assessing the bloom index was to plot the graph of bloom index vs bee abundance:

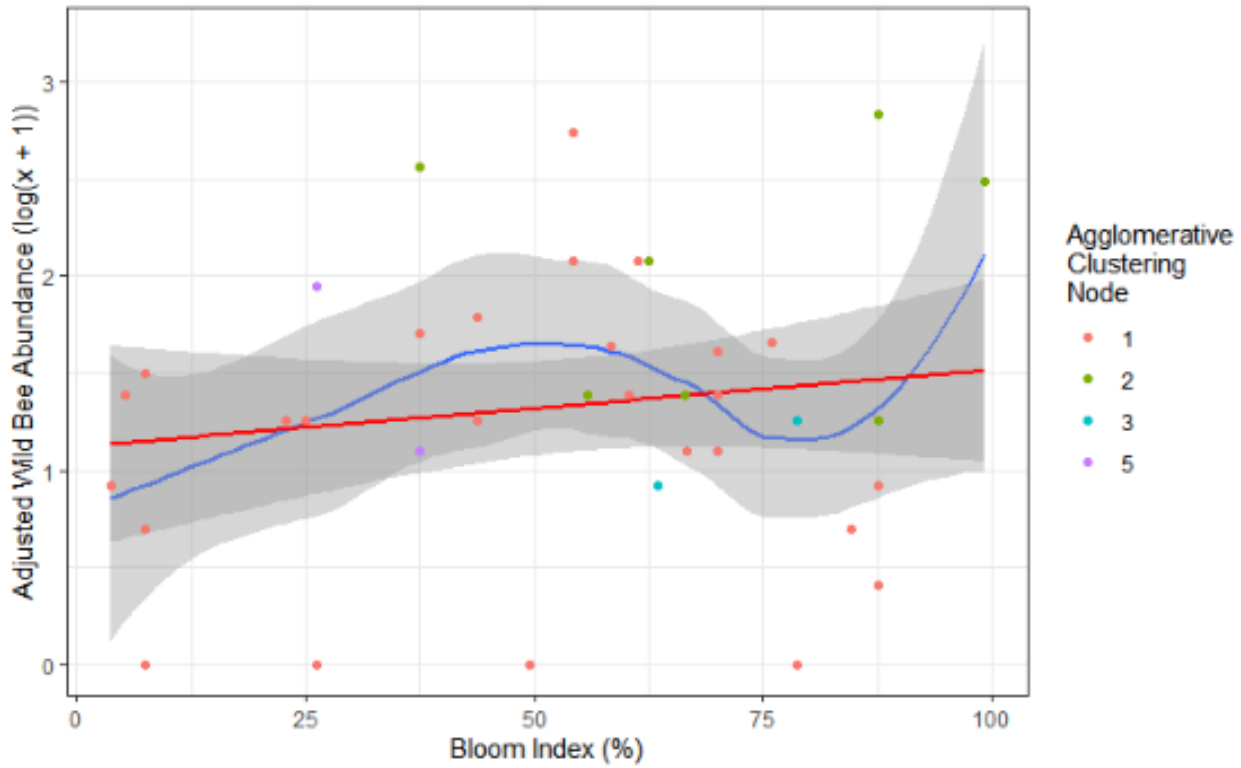


Figure 10: Bee Abundance vs Bloom Index

Figure 10 seemed to show a very slight positive trend between bloom index and bee abundance. For the purpose of testing further, the bloom values were split into three categories; 'Low', 'Medium', and 'High', denoted by less than 33% bloom, between 33% and 66% bloom and greater than 66% bloom, respectively. An ANOVA test was applied to this to see if there was any noticeable difference between the categories. The Null hypothesis for this test was 'There is no difference between the three bloom categories', whilst the alternative hypothesis was 'There is a difference between at least one of the three bloom categories'. There was

insufficient evidence to reject the Null hypothesis (test statistic of 0.456 and corresponding p-value of 0.504).

Table 16: Statistical Test Outputs for Bloom

Test	Test-Statistic	p-value	Significant
ANOVA	0.456	0.504	No

These outcomes demonstrate that bloom does not have a statistically significant effect on bee abundance.

8.8 Abundance Final Confounding Variables

Based on the analyses presented in the above sections, the two significant variables with respect to wild bee abundance are X2000nat and temperature. This was checked by fitting a full linear model to the data and confirming that the same two parameters were significant:

Table 17: Bee Abundance Full Linear Model Output

Coefficient	Test-Statistic	p-value	Significant
Intercept	-2.151	0.040	Yes
temp	2.642	0.013	Yes
X2000nat	3.346	0.010	Yes
local.diversity	0.766	0.450	No
regionLo	-1.345	0.189	No
regionS	-0.112	0.911	No
day	0.818	0.420	No
bloom.index	0.005	0.946	No

It can be seen from the table that the two confounding factors to be adjusted for are indeed X2000nat and temperature, with 'temp', and 'X2000nat' as the only relevant statistically significant values, since adjustment for the intercept variable is not required as this variable is a constant. As a result of this, another linear model was generated just using these two parameters:

$$y = t_c * (t_i) + n_c * (n_i)$$

This model returns the ordinary least squares estimator for both temperature (t_c) and X2000nat (n_c), as 0.064 and 2.138 respectively. These estimates were then used to adjust the wild bee abundance values to account for both confounding variables.

To fix the data for both variables, appropriate constant values were required. A temperature value of 20°C was selected as it was both close to the mean and median temperature value. A X2000nat value of 0.388 was selected as this was the mean value and also close to the median.

The adjustment, in terms of variables, can then be denoted as follows:

$$y_{new} = y_{old} + (t_i - 20) * t_c + (n_i - n_{mean}) * n_c$$

With the variables substituted for their respective numbers this becomes:

$$y_{new} = y_{old} + (t_i - 20) * 0.064 + (n_i - 0.388) * 2.138$$

Where y_{new} denotes the new abundance value, y_{old} denotes the abundance value observed, t_i denotes the temperature of observation i, n_i denotes the X2000nat value for observation i, t_c denotes the temperature coefficient factor, and n_c denotes the X2000nat coefficient factor.

To demonstrate the benefits of adjusting for these two variables, the linear model was re-run taking in the adjusted dataset. The outcome from this shows that both temp and X2000nat are now not statistically significant with respect to bee abundance and emphasises that now the abundance is independent of both these variables:

Table 18: Final Bee Abundance Linear Model Output

Coefficient	Test-Statistic	p-value	Significant
temp	0.025	1.000	No
X2000nat	0.091	0.902	No

8.9 Richness Final Confounding Variables

Based on the fact that both methods of determining the statistically significant confounding variables for abundance match, for bee richness, just the linear modelling method was used. The model used to detect which variables were statistically significant in the linear modelling approach was:

$$adjusted.social = temp + X2000nat + local.diversity + region + day + bloom.index$$

The output from this model can be seen as follows:

Table 19: Bee Richness Full Linear Model Output

Coefficient	Test Statistic	p-value	Significant
Intercept	-1.397	0.173	No
temp	-0.158	0.875	No
X2000nat	3.346	0.002	Yes
local.diversity	-0.917	0.366	No
regionLo	-0.633	0.532	No
regionS	0.609	0.546	No
day	1.715	0.097	No
bloom.index	0.890	0.380	No

As demonstrated above, of all the variables considered, only X2000nat was statistically significant. This makes sense as whilst it is known that temperature affects the bees flight speed (see Section 4.3) and thus the number of bees caught, it is unlikely to affect the number of species of bees observed.

The bee richness was then adjusted using the following formula:

$$y_{new} = y_{old} + (n_i - n_{mean}) * n_c$$

Where y_{new} denotes the new richness value, y_{old} denotes the abundance value observed, n_i denotes the X2000nat value for observation i, n_c denotes the X2000nat coefficient factor (1.467).

9 Model Comparisons

In this section the main results from the clustering methods used to model the orchard protocols are highlighted. There were four different methods considered, each of which was run both on standardised and raw (non-standardised) data for the pesticide levels. Kmeans clustering outputs are not covered here as (per Section 7.9) it was deemed a sub-optimal clustering method when compared with Agglomerative clustering. The outputs may be viewed in the application created should they be of interest.

9.1 Standardised versus Raw Data

In machine learning, it is general practice to standardise the data before passing it into the clustering algorithms. This is because standardising data transforms it from the raw values onto a normalised scale so each of the variables can be directly compared. This has an effect of bringing the values closer together and means that each variable will have the same weighting in the clustering algorithm used. For these reasons, it is generally good practice to always standardise the data before clustering. In the case presented here, the raw data is related to the PUI calculated based on the BIQ as denoted in the Methods Section (5.1). By standardising the variables in this way, some of this bee impact information is 'lost'. The difference between the two methods has been investigated to see how standardising the data varies the cluster outputs in this situation.

In the raw data the fungicide values are considerably higher than both insecticide and thinner values. This means that in the clustering algorithms fungicide has a bigger weighting on how the data are split into sub-clusters. As a direct effect of using the raw data, the optimum number of clusters increases as it allows for the protocols to be split more. The reason that both options were considered in this analysis was to see how additional weighting on the higher pesticide values affected the clustering and associated estimated bee count values, as in reality the actual pesticide value does have an effect on the damage caused to bees. For example, the difference to bees' health between a fungicide value of 120 and a fungicide value of 12 is likely to be a lot more than a thinner value of 0.4 being used in comparison to a thinner value of 0.04.

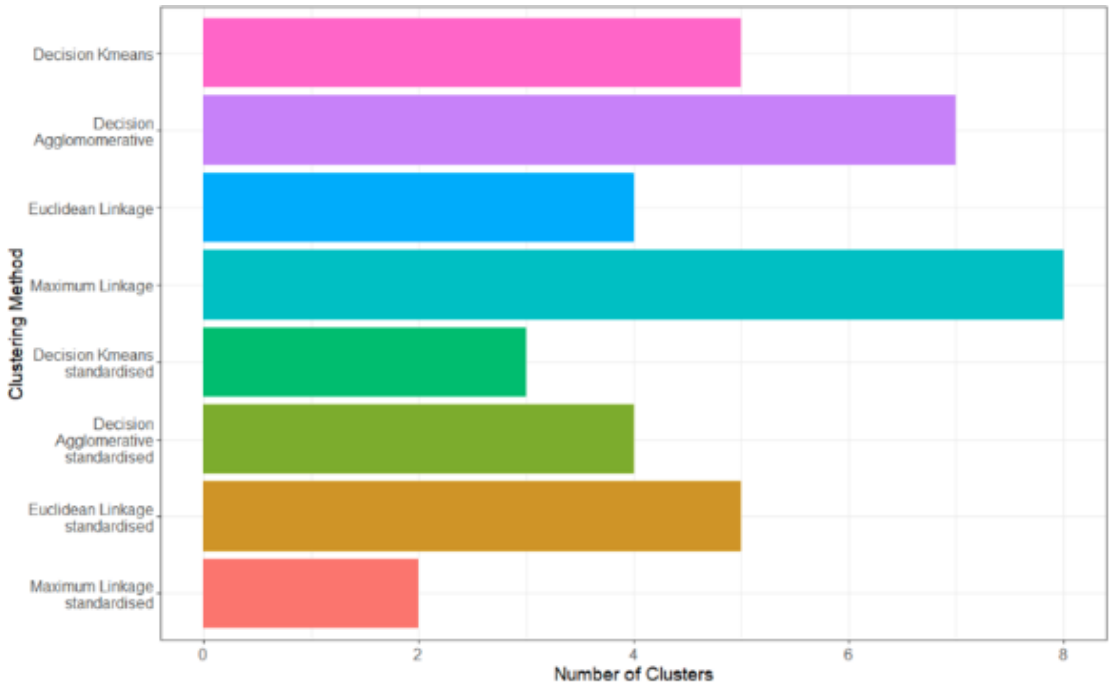


Figure 11: Summary of Machine Learning Methods and Cluster Numbers

9.2 Clustering Outputs

In this section the outputs from six of the eight clustering methods considered are presented. The bee values displayed are the logged versions. Non-logged versions are not displayed here, but can be seen using the application. In all the tables the following abbreviations are used for pesticides: Fungicide as F, Insecticide as I, Thinner as T. For time points, Pre refers to pre-bloom, Blm refers to during bloom and Post refers to post-bloom. For ease of reading the outputs, wild bee abundance will be referred to as abundance and social richness as richness

For the decision theory methods the cluster number refers to the end cluster of the decision process and not the actual number of the cluster. More information on these cluster numbers can be found in Figure 4.

9.2.1 Standardised Maximum Linkage Agglomerative Method

The standardised Maximum Linkage Agglomerative method resulted in two different clusters being generated, as shown in Table 20. This was the lowest number of any of the methods considered. One thing to note was the large disparity in number in each cluster, with 17 in the first cluster and two in the second. This is suboptimal and is discussed in detail in Section 9.3.

Table 20: Standardised Maximum Linkage Function Agglomerative Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T
1	17	1.28	0.40	0.10	-0.21	0.06	0.07	-0.22	0.07	0.08	-0.22
2	2	1.15	0.22	-0.86	1.76	-0.53	-0.63	1.88	-0.63	-0.65	1.91

9.2.2 Standardised Euclidean Linkage Agglomerative Method

The standardised Euclidean Linkage Agglomerative method resulted in five clusters being generated from the data (Table 21). The first cluster from the previous method has now been split into three clusters (1, 2 and 3), whilst the second cluster has now been split into two separate clusters (4 and 5).

Table 21: Standardised Euclidean Linkage Function Agglomerative Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T
1	9	1.44	0.49	-0.54	-0.28	-0.47	0.56	-0.19	-0.39	-0.44	-0.23
2	5	1.17	0.30	1.12	-0.29	-0.05	-0.51	-0.30	-0.22	1.04	-0.20
3	3	1.01	0.27	0.33	0.14	1.84	-0.42	-0.19	1.95	0.01	-0.24
4	1	1.57	0.30	-1.58	3.87	-1.36	-0.75	-0.31	-1.18	-0.68	-0.29
5	1	0.74	0.14	-0.15	-0.34	0.31	-0.51	4.07	-0.09	-0.61	4.11

9.2.3 Standardised Agglomerative Decision Theory Method

The standardised Agglomerative Decision Theory method resulted in four clusters being generated from the data (Table 22). Comparing this with the standardised Maximum Linkage Agglomerative method described in Section 9.2.1, it can be noticed that the first cluster has been split into two clusters (1 and 2), whilst the second cluster has now split into two. This is similar to the outcomes seen with the standardised Euclidian Linkage Agglomerative method described in Section 9.2.2.

Table 22: Standardised Agglomerative Decision Theory Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T
1	13	1.24	0.43	-0.14	-0.28	-0.01	0.16	-0.20	-0.13	-0.43	-0.21
2	4	1.44	0.28	0.90	0.02	0.30	-0.21	-0.30	0.73	1.71	-0.28
3	1	0.74	0.14	-0.15	-0.34	0.31	-0.51	4.07	-0.09	-0.61	4.11
5	1	1.57	0.30	-1.58	3.87	-1.36	-0.75	-0.31	-1.18	-0.68	-0.29

9.2.4 Maximum Linkage Agglomerative Method

The Maximum Linkage Agglomerative method resulted in eight clusters being detected in the data (Table 23). Of these clusters, the modal number of orchards in each was two, whilst the minimum and maximum were one and five respectively. A wide variation in abundance values can be noticed from the different clusters observed.

Table 23: Maximum Linkage Function Agglomerative Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T	Total
1	4	1.48	0.36	156.95	0.00	54.39	11.78	0.01	27.15	49.32	0.00	299.60
2	2	1.85	0.36	29.94	0.00	57.38	13.68	0.01	26.41	5.96	0.00	133.38
3	2	1.58	0.30	15.28	3.37	27.61	20.93	0.23	15.98	22.12	0.00	105.52
4	1	1.06	0.27	163.95	2.31	112.90	12.62	0.02	88.69	51.87	0.04	432.40
5	2	0.98	0.27	112.47	0.00	115.10	6.24	0.11	60.62	8.59	0.04	303.17
6	5	1.12	0.50	101.07	0.19	52.18	15.09	0.64	21.35	13.34	0.67	204.53
7	1	1.58	0.48	107.72	0.00	0.56	1.80	0.00	41.85	1.80	0.00	153.73
8	2	0.58	0.33	195.11	0.19	63.52	4.94	0.00	20.06	21.36	0.16	305.34

9.2.5 Euclidean Linkage Agglomerative Method

The Euclidean Linkage Agglomerative method resulted in four clusters being generated from the data, with the number of orchards in each cluster varying from a minimum of three to a maximum of six (Table 24). This clustering method had the best result in terms of the data sparsity. This is both due to the optimum cluster number of four and with the method choice. One thing to note is that using the Maximum Linkage Agglomerative method with four clusters yielded an almost identical result.

Table 24: Euclidean Linkage Function Agglomerative Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T	Total
1	6	1.48	0.36	169.67	0.06	57.44	9.50	0.01	24.79	40.00	0.05	301.52
2	4	1.85	0.34	22.61	1.68	42.50	17.30	0.12	21.20	14.04	0.00	119.45
3	3	1.58	0.30	129.63	0.77	114.37	8.37	0.08	69.98	23.01	0.04	346.25
4	6	1.06	0.27	102.18	0.16	43.57	12.87	0.53	24.77	11.42	0.56	196.06

9.2.6 Agglomerative Decision Theory Method

The Agglomerative Decision Theory method, resulted in seven clusters being detected from the data, of which three only contained one orchard observation and one contained nine observations (Table 25). This demonstrates that the decision theory clustering method is prone to bias due to the low numbers of observations in the nodes.

Table 25: Agglomerative Decision Theory Clusters

Cluster	N	Abundance	Richness	Pre F	Pre I	Bloom F	Bloom I	Bloom T	Post F	Post I	Post T
1	9	1.06	0.47	127.38	0.15	49.62	11.80	0.36	22.58	15.06	0.41
2	3	1.56	0.28	161.47	0.00	53.16	9.36	0.00	31.37	57.66	0.00
3	1	1.06	0.27	163.95	2.31	112.90	12.62	0.02	31.37	57.66	0.00
4	2	0.98	0.27	112.47	0.00	115.10	6.24	0.11	60.62	8.59	0.04
5	2	2.05	0.27	31.52	3.37	31.27	9.87	0.01	19.93	9.55	0.00
6	1	1.59	0.29	9.38	0.00	37.44	36.36	0.46	24.39	34.75	0.00
7	1	1.17	0.45	18.03	0.00	70.09	13.10	0.00	20.53	2.28	0.00

9.3 Clustering Conclusions

Since the original research yielded so few data points, the clustered nodes also contain very few data points, with some comprised of just one orchard. This meant that using Leave-One-Out cross validation (LOOCV) would not assess the accuracy of the varying methods reliably, since in this case it would tend to favour whichever clustering method returned the fewest clusters. Other methods of assessing machine learning accuracy numerically also have the same limitation. With so few data points, any training and test data sets are so small that it is very difficult to distinguish a meaningful difference. The resulting statistical power of the test is very low and the chance of making a type II error comparatively large.

As a result of the limitations described, it was decided that visual inspection would be used for analysing the clusters to determine which clustering method best demonstrates the harmful effects of the pesticides. Whilst this is by no means perfect, the damaging effects of pesticides on wild bees is well established, and the best clustering method will demonstrate most clearly, the association of the high pesticide protocols with low bee abundance and richness values, and the low pesticide protocols with high bee abundance and richness values. The downside of this method is the lack of statistical back up, so all conclusions made are subject to an individual's perspective on the clustering data and bee counts observed.

A number of conclusions could be drawn from the analyses presented in this section. The first is that raw data provides a better clustering than the standardised data. This, in general can be seen throughout, but the most clear example is the Euclidean Linkage Agglomerative method (Sections 9.2.2 and 9.2.5). The standardised clustering resulted in five clusters with a wider range of data points in each cluster, varying from one in two clusters to nine in the largest cluster. The raw data clustering however resulted in four clusters, which vary from three data points in the smallest cluster to six in the largest, demonstrating that using the raw pesticide values can help combat the sparsity problem when using small data sets in the ecology sector.

The models which best demonstrate the link between pesticides and bee count are the classical machine learning models using the raw data. This is due to a wider range of bee abundance and richness values being observed. These values can be clearly associated with the variation in pesticide protocols.

The level of fungicide at the blooming time point seems to have a large effect on the bee abundance. The standardised Agglomerative Decision Theory and Euclidean methods (Sections 9.2.2, and 9.2.3) clearly demonstrate this, with the clusters having high bee counts also having negative values for blooming fungicide. Similarly, in all the clustering methods using the raw data the clusters with highest abundances can be linked to those with low levels of fungicide during the blooming period.

Fungicide levels have the biggest effect when it comes to determining if the expected bee

abundance will be high or low, as shown by the clusters with the highest fungicide levels having, in general, the lowest bee abundance counts. This could be linked to the fungicide values being considerably higher than both insecticide and thinner, demonstrating that bees are affected by the total amount of pesticides used. Code was written to generate these summaries and the respective column added to both (Sections 9.2.2, and 9.2.3). There was a clear trend that linked total pesticide applied with bee count, especially with the Euclidean linkage function where there was a direct inverse correlation between highest pesticide protocol and lowest bee abundance. There was not a clear trend with total pesticide use and social bee richness however.

A classical machine learning approach also performed better than the decision theory approach considered. Similar to the standardised data, the decision theory approach tends to have one large cluster and then the remaining clusters contain only a few data points (Sections 9.2.3, 9.2.6), by comparison to the classical approaches which had a less varied number of orchards per cluster.

Social richness has much less variation than abundance. If the output from the Maximum and Euclidean Linkage Agglomerative methods (Sections 9.2.4, and 9.2.5) is considered, it can be noted that the variation seems less dependent on the orchard protocols, than bee abundance. Although, there could be deemed to be a correlation similar to the relationship observed between bee abundance and pesticides, there are insufficient data at the moment to conclude this with any certainty.

There is insufficient variation in the clusters generated from the data to choose between the two classical machine learning techniques considered. Both Euclidean and Maximum Linkage functions returned a good spread of results with a clear association with bee abundance and quantity of pesticides in the orchard protocol. If just one had to be chosen going forwards, it would be the Maximum Linkage function. This is because it scored slightly higher than the Euclidean Linkage function in the scoring table of distance metrics and linkage functions (See Table 9).

10 Shiny Application

10.1 Reasoning

The reason behind building a prototype web application (app) is to give the end-user, an orchard owner, a way of visualising how different orchard protocols will affect bee populations. The ultimate aim for the app would be to allow these orchard owners to make more informed decisions on the most appropriate pesticide protocol to use on their orchards, with a focus on demonstrating how the app could be used to gain value. Whilst the data set used to generate the app is very small, the principles can be extended in future and the outputs generated in the app could be changed to reflect this. The app will also summarise some of the key findings of the research in other tabs, so those who are interested in how the data are generated can see for themselves.

It was decided to use RShiny to build the app. This was the most intuitive way considering that all the programming analysis for the project had been conducted in 'R'. This package sits in the R environment and allows the outputs from the analyses to be fed directly into the app.

10.2 User Experience

The main focus for creation of the app was the user interface, with the aim of making it simple, quick and intuitive to use. The app is broken down into three subsections: a 'Clustering Summary' for different orchard protocols; a 'Data Analysis' section to highlight some key areas covered by this dissertation, and an 'About' section to give credit to the original authors as well as provide links to the project GitHub page, where all the code is stored.

In Figures 12 and 13, images have been displayed to demonstrate how the application works. The first depicts how the load screen appears if the app is used on a computer. The second gives one of the examples of the key analysis points of the project. This particular example shows a summary of the distance metric and linkage functions considered in this project. For an interactive view see the actual application (Brownsey, 2019a).

Figure 12: Load Screen Example on Computer

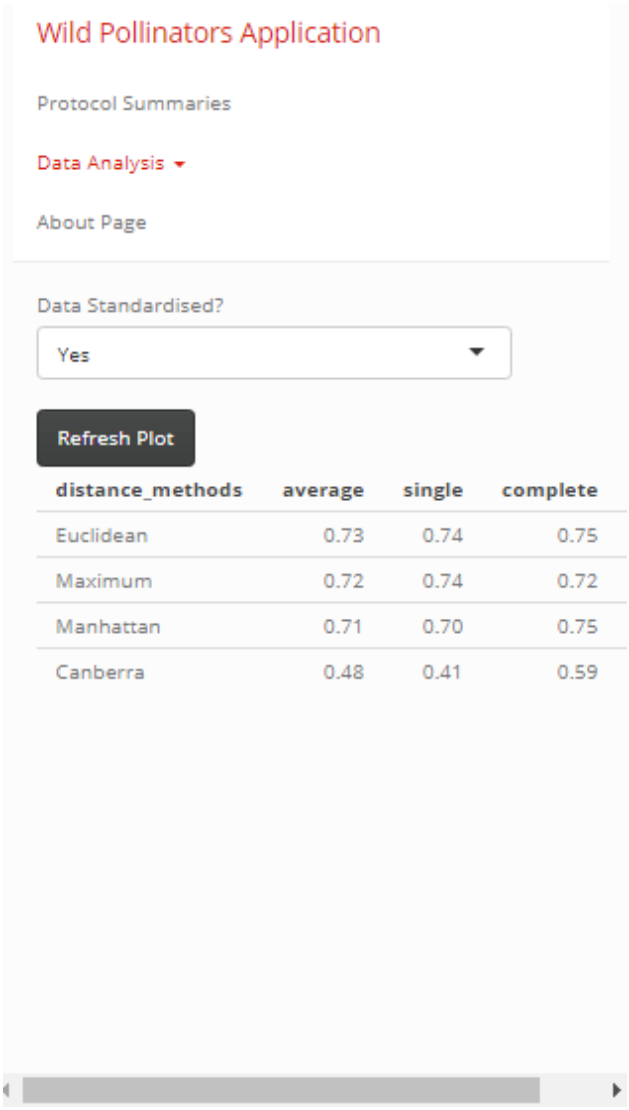


Figure 13: Clustering Page on a Mobile Display

In the interest of usability and inputs which are easy to understand, quantitative results of the clustering are transformed into qualitative options for the user to select. The presentation of graphical outputs from these choices allows the benefits to pollinators by choosing a less harmful pesticide protocol to be demonstrated. The 'Key Analysis ' tabs are self-contained to allow easy navigation on mobile phones and tablets.

10.3 Hosting

The decision was made to host the app on 'Shinyapps.io', a free web hosting platform for RShiny applications. It not only allowed for testing to be carried out by friends more easily but also provided a 'proof of concept' as to how the app could be shared with the target audience. In addition, by using this platform, users may access the app on a variety of devices, which has the added benefit of reaching a wider audience.

Shinyapps.io has the benefit of being able to deploy apps directly from the RStudio programming session and the app can be updated as many times as required by simply redeploying the changes. This was particularly useful when implementing changes to the app, with testers immediately able to see the difference and provide feedback on whether or not it was what they had meant. The only downside of the Shinyapps.io approach is the 25 hour monthly usage limit, but for the purposes of the prototype this was deemed an acceptable drawback. If a commercial version was ever required a premium Shinyapps.io subscription could be purchased to increase the usage allowance. The URL for the RShiny app can be found by clicking on: **Application**

10.4 Testing

To test the usability of the app, some family members and friends were asked to use the app, test the settings and provide feedback on where improvements may be made. This feedback is summarised in the following points:

1. It would be more visually appealing to have side by side graphs for comparison, rather than different graphs on different pages,
2. Whilst I like the setup of using different pages I think simpler is better,
3. Is it possible to have one tab for the stuff you want to show to an orchard owner and another tab with the key points?
4. Less complex graphics, just a bar-chart would be easier for me to understand,
5. If you can subsection your key analysis section into sub-parts for different things, that would be easier to use than one very complicated tab,
6. Hard to navigate on a mobile phone.

In general the feedback was that a simple to understand user interface would be an improvement, as well as self-explanatory graphics and based on this feedback, changes were made by page type as follows:

The feedback from my non-statistical test audience was really beneficial as it demonstrated that 'less is more' in many scenarios. As a result, the number of visualisations was reduced and simplified. This led to the final implementation of this page having just two visualisations which would update the graphs with the latest pesticide combinations every time the user clicked a button. These visualisations demonstrated the difference in expected wild bee abundance and expected social bee richness resulting from the chosen pesticide protocol. The tabs showcasing some of the key points of the research were also made more intuitive to use. The following bullet points highlight some of the key changes made:

1. All the data relating to the qualitative protocols were situated in the one default tab of the app which is loaded on startup,

2. The additional information in the app, relating to the research undertaken, was split into sub-pages within the 'Data Analysis ' tab to allow for easy navigation of the application, especially on mobile phones and tablets,
3. An 'About' tab was also added to act as a reference source, which provided links to both the original research and the research carried out as part of this dissertation,
4. The sizing of the tabs was improved to allow for easy navigation on mobile phones and tablets.

10.5 Future App Usage

This RShiny app is very basic but it does demonstrate how this could be both of use and of interest to those in the industry. As more data are collected, it would allow more cluster combinations, each of which could contain more data points and these clusters would then provide a better insight into the exact effects of the different pesticide protocols.

If data relating to apple yield were collected, this would allow the web application to be taken further and enable a comparison of an agricultural perspective vs economic gain to be visualised. In particular it would allow the ecologist vs economist trade-off to be visualised. A simple slider could also be implemented to give weight to each perspective. The expected monetary gain and bee counts would then also be displayed. The current layout of the application could be retained but the drop-down list expanded to contain these future possibilities as options.

11 Project Management

In general the project progressed in accordance with the updated specification, and the objectives were completed in accordance with the Gantt chart defined in the project specification. This can be seen in Figure 14:

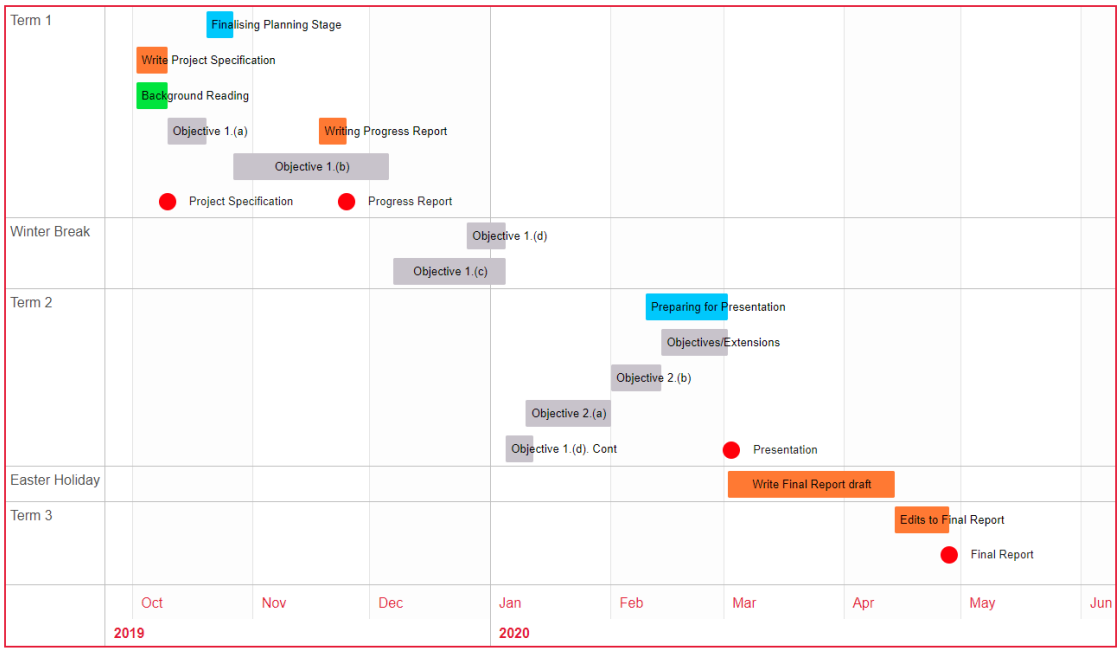


Figure 14: Project Specification Gantt Chart

Slight deviations from the plan were noted. In particular, the data were not as complete as had been anticipated and this lead to additional time taken with the exploratory data analysis to understand the limitations of the data fully, and decide upon the best ways to deal with these. The extra workload was caught up over the Christmas break, so no delay was taken into the second term. In addition, due to the change in specification, extra reading was required to understand the machine learning principles deployed more completely, and to select the appropriate algorithms to be used for comparisons. Again this was completed over the Christmas period. The time for objectives and future extensions was utilised to analyse the original objectives from a theoretical perspective (see Section 14).

An RShiny app was written to produce a fully interactive Gantt chart image. A still screenshot from this is presented in Figure 14. This allowed my progress to be tracked accurately, and enabled the impact of any delays to be seen visually. The problems could then be managed and mitigated accordingly. As the app took in a .csv file, it was very easy to recreate the Gantt image to incorporate any changes. Similar to the application built for the project work, this was hosted on Shinyapps.io to allow for access from anywhere and any device. This was particularly useful when discussing the objectives with my project supervisor.

12 Legal, Social, Ethical and Professional Issues

12.1 Legal Issues

The data used are publicly available from the Park et al. (2015) publication.

12.2 Social Issues

The project does not involve any human-related or other social issues.

12.3 Ethical Issues

The only potentially ethical issue related to the treatment of the bees once caught, and possible harm caused. However, since these data had already been collected, any possible harm caused was not as a direct or indirect result of the analyses conducted as part of this dissertation.

12.4 Professional Issues

It was noted that the dataset online contained confidential information as to the exact location of each orchard used in the study. Originally, I did use this information to generate some analysis plots. However, for the sake of retaining confidentiality with any personally identifiable information (PII), any inference and figures made from this PII data have not been reported.

13 Results and Conclusions

Based on the analyses undertaken, the conclusions drawn from the data can be summarised as follows:

1. The analyses supported previous conclusions that percentage of natural surroundings and temperature were positively correlated with bee abundance levels,
2. The outcomes demonstrated that machine learning could be used to highlight the harmful effects of pesticides on bee abundance and richness, reducing the need for domain knowledge to draw conclusions,
3. The raw pesticide protocol data provided a better clustering summaries than the standardised protocol values,
4. The classical machine learning clustering approach performed better than the decision theory time point approach,
5. When assessing the performance of clustering methods, sparsity of the nodes plays a big role in the effectiveness of an algorithm. The addition of a minimum node capacity could help adapt existing methods to perform better on small data sets,
6. Fungicide and total pesticide use was strongly linked to bee abundance.
7. The best two cluster methods were classical Agglomerative clustering using the Ward distance metric with both Euclidean and Maximum Linkage functions. There was insufficient evidence to distinguish between these two methods visually using the clusters, but the Maximum Linkage method came out on top when making numerical comparisons,
8. An RShiny application was built to showcase key results. This was web-hosted to allow access on computers, mobile phones and tablets.

14 Theoretical Decision Theory

14.1 Introduction

As covered in the analysis so far, the data from the study were not in a suitable format to cover the original decision theory concepts posed, as there were insufficient data points to densely populate the decision tree. As mentioned previously the decision was made not to simulate data as this would not provide a representative overview of actual trends in the data. Instead, the aim of this section is to cover how the original questions could be answered from a purely hypothetical and theoretical perspective.

The analysis presented considers both the concept of data collection and decision theory methodology, with the aim of demonstrating how, given a statistically planned study, decision theory could provide insight into the ecology:economy trade-off. This could eventually lead towards an optimal pesticide protocol selection strategy to both provide a profitable yield and minimal effects on the bee populations.

Since no data were used for the creation of the decision trees, no probabilities can be numerically elicited, instead they will be given in a purely theoretical statistical representation. Then, should a future study have the required data, the probabilities themselves could be calculated.

14.2 Data Collection

In this subsection, a discussion of the existing reported data and required data in presented.

In addition to the variables currently reported, a record of apple yield, the unit (kg) price of the apples and the cost of the fertiliser used would be required to produce meaningful metrics on the ecology:economy trade-off.

The day variable, currently reported simply as Day 1 or Day 2 would be more beneficial as an actual date. In particular, if there were recordings in each blooming period, this would enable specific sub-analysis of the effects at different time periods to be investigated accurately. This could potentially give insight into how the timing of pesticide application could be used to maximise protocol effectiveness within both a time period (Pre-bloom, bloom, post-bloom) and the growing season as a whole.

A variable for apple progress would be required to track the growth of the apples for both the economic and combined decision theory methods. An additional requirement here would be the requirement for human experience and time to score the apple progression, and as such even with a prespecified score scale would be prone to bias. Apple progress could be based on size and ripeness, but an expert view would be required for the most accurate specification. The bias is due to there being a human effect to the score which would be hard to control,

especially since it would be likely that different people measured it for different orchards.

The bee counts are currently collected manually with a person walking a transect of the orchard with a net, with the number of bees collected in a 15 minute time period then recorded as the observed number. This value is very prone to bias. If cameras were used instead to collect the number of bees, this would provide a much more consistent and accurate result. It is likely that some bees will be photographed more than once by the cameras, so one assumption required would be that the rate of returning bees is constant. Since the bee counts are being compared, this would not affect any conclusions made, given the assumption holds true.

Only one variable of each bee type needs to be considered due to the high level of correlation between the variables. By recording bee abundance, bee richness can be calculated easily, as the observations have already been made. As such it would be worth reporting both variables in a summary data set.

A summary of all the required variables for the decision theory applications is summarised below:

1. Orchard ID - Unique ID for privacy,
2. Date of visit - Date variable in form DD/MM/YYYY,
3. Apple yield (kg/square metre) - Yield of apples from the orchard,
4. Orchard size (square metres) - Size of orchard in square metres,
5. Apple price - Selling price of apples per kilo,
6. Apple progress - A score to demonstrate how well the apples are growing,
7. Bloom variable - Three categories: Pre-bloom, During-bloom and Post-bloom,
8. Natural surroundings - Percentage of natural surroundings,
9. Temperature - Temperature on day that counts were taken,
10. Wild abundance - Abundance count for wild bees,
11. Wild richness - Richness count for wild bees,
12. Social abundance - Abundance count for wild social bees,
13. Social richness - Richness count for wild social bees,
14. Pesticide cost - Total cost of the pesticides used for the orchard,
15. Pre-bloom insecticide - PUI value for pre-bloom insecticide,
16. Pre-bloom fungicide - PUI value for pre-bloom fungicide,
17. During-bloom insecticide - PUI value for during-bloom insecticide,

18. During-bloom fungicide - PUI value for during-bloom fungicide,
19. During-bloom thinner - PUI value for during-bloom thinner,
20. Post-bloom insecticide - PUI value for post-bloom insecticide,
21. Post-bloom fungicide - PUI value for post-bloom fungicide,
22. Post-bloom thinner - PUI value for post-bloom thinner,

14.3 Decision Theory

14.3.1 Introduction

Given that the data are in the format discussed above, and there was a significantly large enough number of observations to allow for the tree to be sufficiently dense at each stage, then the decision theory approach discussed below could be investigated, probabilities elicited and accurate conclusions made.

There are endless possibilities when it comes to designing a decision tree. In this discussion, three possible scenarios are investigated to see how the concepts of decision theory are applicable to both bees and the agricultural industry in general. One area in particular which was considered is how they vary from both a stochastic process, where the decision is dependant on all previous steps, and a standard Markov stochastic process, where each step is only dependent on the previous step. The decision tree and indeed the Markov chains considered are discrete-time Markov processes, this is because the 'steps' occur at known bloom points. A formal definition of these is as follows (Ross, 2019):

A stochastic process $\{X_n, \quad n = 0, 1, 2, \dots\}$, takes a finite or countable number of possible values, with n taking the set of non-negative integers. If $X_n = i$ then the process is said to be in state i at time point n. Then, given the process is in state i, there is a fixed probability p_{ij} that the next state will be j. This can be denoted mathematically as:

$$p_{ij} = P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\}$$

A matrix can be used to summarise all the one-step transition probabilities (p_{ij}). Three definitions that will be used later regarding this are:

$$p_{ij} \geq 0, \qquad i, j \geq 0, \qquad \sum_{j=0}^{\infty} = 1$$

Let S be a measurable state space as in the decision models discussed here then, under the Markov assumption, the stochastic process can be formulated as:

$$P(X_n = x_n | X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}) \qquad [1]$$

Taking this one step further, higher order Markov stochastic processes, whereby the sequence is dependent on the last m steps are also considered. For the discrete set considered here, a formal definition is as follows:

$$P(X_n = x_n|X_{n-1}, ..., X_0 = x_0) = P(X_n = x_n|X_{n-1} = x_{n-1}, ..., x_{n-m}) \qquad [2]$$

A key difference to note is between the bee outcomes being dependent on previous nodes, and the decisions themselves being dependent on previous nodes. The bee numbers are dependent on all the prior decisions whereas the decisions are only dependent on the previous timepoint for both bee and/or apple values. In other words, the expected outcome of a node depends on all possible routes between that node and the end outcome nodes and their associated probabilities. By contrast, the decision as to which pesticide protocol to use is only dependent on the previous one or two nodes. The number of nodes at the timepoint is dependent on the utility strategy considered. In general, the decisions follow either a stochastic Markov process or a higher order Markov process with $m = 2$. However, the values follow a standard stochastic process whereby the outcome depends on all previous decisions and associated probabilities.

Three types of models were considered in the research presented here. The respective decision theory approaches considered were:

1. An ecologist's perspective,
2. An economist's perspective,
3. The ecologist vs economist trade-off.

Utility functions are used in decision theory to evaluate the value of each branch, with the value of a node referring to its expected utility. This expected utility is dependent on the utility function defined and will be used to determine the optimum path of decisions based on the data collected, and could be used to determine the optimum strategies going forwards. More formally a utility function (u) in a domain space S maps the raw values into a meaningful but subjective value, which can represent the differing views of each perspective:

$$\text{For any A, B} \in S : u(A) \leq u(B)$$

It is through utility functions that the same raw values can be used to represent each of the different perspectives. An important observation about utility theory is it allows the specific views to be seen. For example, a bee abundance of 600 might not be twice as 'valuable' as a bee abundance of 300. Through utility theory these preferences can be accurately elicited to determine the optimal strategy under each decision maker's priorities.

Utility functions will be referred to in two ways in this section: U refers to the utility as

whole and will be used when referencing the expected utility of a given node $\mathbb{E}[U]$, whilst U_j denotes the utility of the realised path j in the decision tree.

When defining these models, two branches are shown at each stage, both for the sake of simplicity and to allow the models to be drawn and displayed in a more understandable manner. In reality, the number of branches would be data-driven using the Agglomerative clustering algorithm with Ward's distance function and Maximum Linkage function. A data-driven approach would also be implemented to decide upon the limit of the minimum number of data points in each cluster so as to avoid sparsity and bias. When the models refer to n, they are referring to the optimal value calculated from the specified machine learning clustering method. This value of n would have the ability to vary at each stage of the decision tree.

Based on the conclusion that total pesticide use is significant, as well as individual pesticide use (Section 13), the pesticide protocols are grouped by the blooming time points, rather than by each individual pesticide application. For all decision trees drawn, a square denotes that a decision was made, a diamond signifies that an observation was made and a circle denotes an outcome node. In this example, only the bee abundances were considered in the utility, and hence just the b_{i1} and b_{i2} appear. The difference can be visualised in Figure 15 below, for one time point of the whole tree, with the protocol method on the left and time point on the right.

The Fungicide time points are denoted by $F11$, where 11 signifies it is the first time point and the first cluster. $I11$ denotes the same for Insecticide. To denote the pesticide protocol used for time point 1, $V11$ was used. To signify whether the bee count was high or low at the first time point B_{11} and B_{12} were used, respectively.

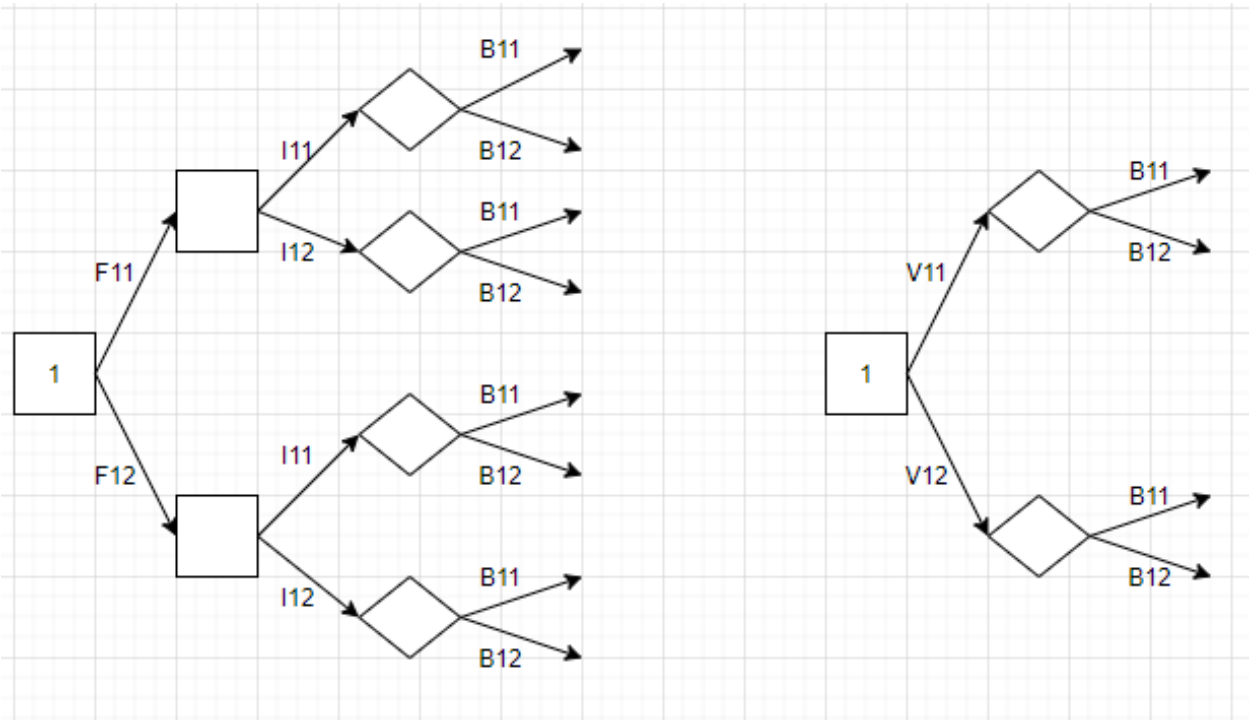


Figure 15: Difference between Clustering by Time Point vs Pesticide Application

In terms of the statistical difference with dependence in relation to the bee count variables, the protocol method can be denoted as following the higher order Markov process [2] with $m = 2$ whereas the time point methodology follows the standard Markov process [1].

In the analysis considered here, the decisions and observations can be summarised in the following way, where i denotes the time point which is being considered $i \in (1, 2, 3)$ and, note in the definitions below both $\mathbb{Z}_+$ and $\mathbb{R}_+$ include 0. For simplicity, throughout the letter P will be reserved for referencing probabilities. For example, $P(x)$ references the probability of event x occurring:

1. The protocols values lie in $\text{Pesticide}_i \in \mathbb{R}_+$, the protocols clusters are $V \in (v_{11}, v_{12}, \ldots, v_{32})$ where x_i denotes the cut-off pesticide value for each cluster.

$$V = \begin{cases} v_{i1}, & 0 \leq \text{Pesticide}_i \leq x_i \\ v_{i2}, & \text{Pesticide}_i > x_i \end{cases}$$

2. The bee observations take a value in $\text{Bee}_i \in \mathbb{Z}_+$, the abundance clusters are $B \in (b_{11}, b_{12}, \ldots, b_{32})$. For the purpose of decision making, these observations are grouped together as either high or low counts at each stage based on a threshold value y_i:

$$B = \begin{cases} b_{i2}, & 0 \leq \text{Bee}_i \leq y_i \\ b_{i1}, & \text{Bee}_i > y_i \end{cases}$$

3. The apple observations take a value in $\text{Apple}_i \in (0, 1, 2, \ldots, 100)$, then the apple clusters are $A \in (A_{11}, A_{12}, \ldots, A_{32})$. For the purpose of decision making, these observations are grouped together as either high or low apple scores at each stage depending on whether they are above or below a threshold value z_i:

$$A = \begin{cases} A_{i2}, & 0 \leq \text{Apple}_i \leq z_i \\ A_{i1}, & \text{Apple}_i > z_i \end{cases}$$

Taking bee count as an example, each bee decision, in this scenario, low or high can be summarised as the mean of all the bee observations:

$$U_{low} = \frac{1}{n}\sum_{i=1}^{n} b_{2i}, \quad U_{high} = \frac{1}{n}\sum_{i=1}^{n} b_1 i$$

Then, the expected utility of a bee node can be described as the expected number of bees to be observed from that specific node, using the lower and upper branches and the respective probabilities of each branch being selected:

$$\mathbb{E}[U] = p_{low} * U_{low} + p_{high} * U_{high}$$

Generalising both of these to account for the future possibilities, in which there are potentially greater than two options at each stage:

1. The total pesticide levels of each protocol take a value in $\text{Pesticide}_i \in \mathbb{R}_+$. This value gets assigned to a given protocol cluster: $V \in (v_{i1}, \ldots, v_{in})$ with x_{ix} denoting the cut-off points for each cluster.

$$V = \begin{cases} v_{i1}, & 0 \leq \text{Pesticide}_i \leq x_{i1} \\ v_{i2}, & x_{i1} < \text{Pesticide}_i \leq x_{i2} \\ \ldots & \\ v_{in-1}, & x_{in-2} < \text{Pesticide}_i \leq x_{in-1} \\ v_{in}, & \text{Pesticide}_i > x_{in-1} \end{cases}$$

2. The bee observations take a value in $\text{Bee}_i \in \mathbb{Z}_+$. Since the model being considered is discrete, these values will get mapped into one of n clusters $B \in (b_{i1}, \ldots, b_{in})$, with y_{iX} denoting the cut-off points for each cluster:

$$B = \begin{cases} b_{i1}, & 0 \leq \text{Bee}_i \leq y_{i1} \\ b_{i2}, & y_{i1} < \text{Bee}_i \leq y_{i2} \\ \ldots & \\ b_{in}, & \text{Bee}_i > y_{in-1} \end{cases}$$

3. The apple observations take a value in $\text{Apple}_i \in (0, 1, 2, \ldots, 100)$, the apple clusters are $A \in (A_{i1}, A_{i2}, \ldots, A_{in})$. The number of discrete groups would again be data driven, with z_{iX} denoting the bounds of each group:

$$A = \begin{cases} A_{i1}, & 0 \leq \text{Apple}_i \leq z_{i1} \\ A_{i2}, & z_{i1} < \text{Apple}_i \leq z_{i2} \\ \ldots & \\ A_{in}, & \text{Apple}_i > z_{in-1} \end{cases}$$

In principle, the utility theory is the same when there are n clusters as when there are only two, the difference being that instead of summing over the two possible scenarios, the summation of probabilities runs over the n scenarios. The expected utility output of a given node is denoted by U_i and the associated probability of this occurring is p_i. With these variables, the expected utility can be formulated as:

$$\mathbb{E}[U] = \sum_{i=1}^{n} P(U_i) * U_i$$

The decision trees can get very large, but they are symmetrical at each stage. For this reason only one branch is fully shown, with the rest cut off early. This should enable the points to made more clearly and reduce confusion from an overly large decision tree.

An assumption that will be made throughout is that the apple progress scores and bee abundance levels depend solely on the previous time points pesticide and the decision as to which pesticide protocol to take will depend solely on the previous time points results for apple progress and bee abundance.

14.3.2 Ecologist's Perspective

The analysis undertaken using machine learning demonstrated that bee abundance was linked to pesticide protocol, with a higher quantity of pesticides resulting in fewer bees observed. The assumption is made that there are no situations in which pesticide application would boost bee populations. Only bee abundance was used in the utility function of this section. There was insufficient evidence to demonstrate the same correlation applied for bee richness. However, with more data it is likely that a connection could be made here, and if it were the case that bee richness was demonstrated to be affected by protocol, then it could be added with ease into the utility function. This would be achieved by added extra variables in [3], below, for each time point to correspond to the richness values.

Ecologists would naturally be risk averse and their perspective would clearly prioritise the abundance of bees. As such the pesticide protocol decisions are driven solely by the bee abundances. This means that the model is somewhat trivial as the utility function would be based purely on the bee abundance counts. An example of this, with each time point

weighted equally, can be seen below:

$$U = \frac{1}{3}(bee_1 + bee_2 + bee_3) \qquad [3]$$

This demonstrates that whilst the state space contains pesticide values and costs, the only variable which is of relevance to the utility function is bee abundance.

Since pesticides are known to affect bee populations, the way to maximise the expected utility would always be to apply no pesticide. This is shown in the highlighted route in the decision tree below (Figure 16):

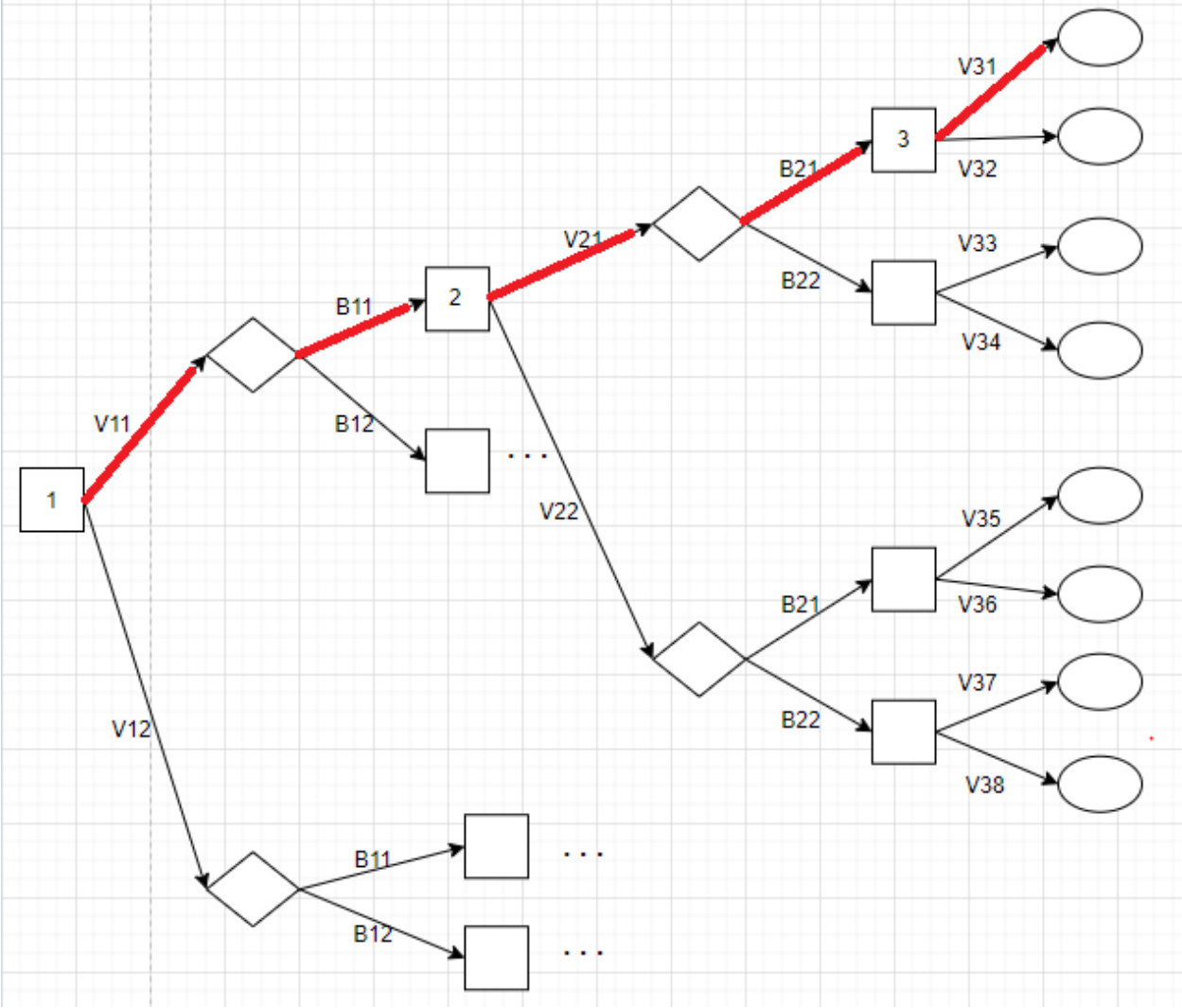


Figure 16: Ecologist's Perspective

There is a chance that the 'route' could deviate down a b_{i2} path, but in each scenario the

expected pesticide decision would always be to use no pesticide. The ecologist's perspective would therefore be 'Never apply pesticides under any conditions'. Then based on the one-step transition probability matrix definitions, it is known the rows sum to 1. In this scenario, the one-step transition matrix will have a '1' at each position, which denotes the decision to apply no pesticide and a '0' to all other decision routes:

$$V = \begin{cases} V_{ij} = 0, & j \neq \text{No pesticide} \\ V_{ij} = 1, & j = \text{No pesticide} \end{cases}$$

In this way, the outcome decision of every node is known with certainty, and hence the traversal of a decision tree is rather trivial, as the majority of nodes will have a probability of 0 of ever being traversed.

14.3.3 Economist's Perspective

From an economic perspective, the deciding factors will be apple progress, apple yield, and profit. Expected Monetary Value (EMV) theory is well suited to this problem, as the utility function can be clearly quantified in terms of monetary gain. This can be denoted as:

expected apple yield - cost of pesticides per square unit measurement (e.g., metres)

to allow consistency between orchards. Each pesticide protocol V_{ij} has a cost C_{ij} associated with it. In Figure 17, this can be seen by each V_{ij}. The apple progress variable $\text{Apple}_i \in [0, 100]$, in the decision tree will be split into two clusters for each pesticide choice, and can be denoted as:

A_{ij}, with A_{1j} and A_{2j}

denoting high and low apple scores respectively. This score is based on how well the apples are progressing and as such is linked to the associated expected apple yield, which is calculated from the end node. The formula for calculating expected value of a node can be elicited by:

the expected yield - the expected cost of pesticides

The state space for this includes the variables: apple yield, pesticide use and associated cost. More formally all possible combinations of these variables can be shown by:

$\{A_{ij}, i \in \{1, 2\}, j \in \{1, 2\}, V_{ij}, i \in \{1, 2, 3\}, C_{ij}, i \in \{1, 2, 3\}, j \in \{1, 2\}\}$

To demonstrate this difference from the ecologist's perspective, a different utility is defined. For an economist, profit is the main goal, as such the utility function is:

apple yield (Y_j) of a given end pesticide protocol outcome - the cost of pesticides (C_{ij}),

where j is the indexing of that particular route. Then the utility of an end node j can be summarised by:

$$U_j = Y_j - \sum_{i=1}^{3} C_{ij}$$

This utility function demonstrates that although the state space contains pesticide protocol values, the utility function is independent of the details of the pesticides and just dependent on their economic cost values as represented:

$$U(V, A, C) = U(A, C)$$

The assumptions being made are that the decisions in the tree follow the standard Markov property [1], whereby the apple cluster is dependent solely on the pesticide levels at the previous time point, and the pesticide protocol levels are dependent on the apple progress score, observed at the previous node in the decision tree.

The decision tree is similar to that of the ecologist's approach, however the bee abundance observations (b_{i2} or b_{i1}) are changed to represent the apple progress (A_{ij}) with the final circle node representing the actual yield rather than progress:

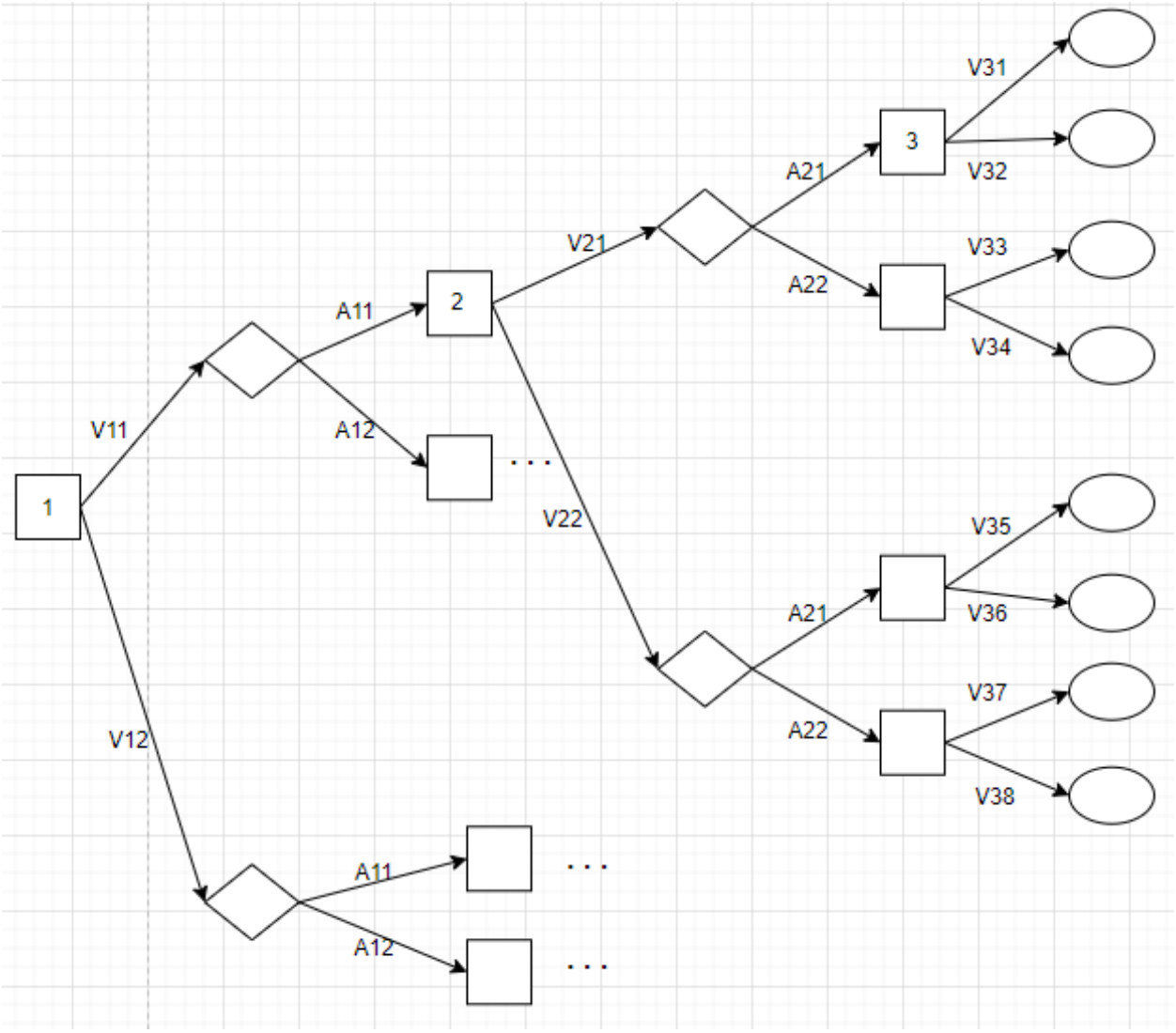


Figure 17: Economist's Perspective

In the simplest scenario, the expectation for node 3 from Figure 17 is summarised as the expected yield of the two outcomes multiplied by the probability of each one occurring minus the expected pesticide costs involved with reaching this node (C_{11} and C_{21}), this can be summarised as a function of the utilities:

$$\begin{aligned}\mathbb{E}[U] &= P(V_{31} * Y_1 + P(V_{32}) * Y_2) - P(V_{31}) * C_{31} + P(V_{32}) * C_{32} - C_{11} - C_{21} \\ &= \sum_{j=1}^{2} [P(V_{3j}) * (Y_J - C_{3j})] - C_{11} - C_{21} \\ &\text{Using the utility function } U_j \\ &= \sum_{j=1}^{2} P(Y_{3j}) * U_j\end{aligned}$$

This expected utility for this node can be generalised for n possible clusters, by simply adjusting the exponent of the sum:

$$\mathbb{E}[U] = \sum_{j=1}^{n} P(V_{3j}) * U_j$$

Taking this one step further to look at the next round of nodes, node 2 from Figure 17, will have the probability of both the P21 and P22 to consider as well as the probabilities that these return a positive or negative apple progress. V_i relates to the pesticide protocol cluster and A_i relates to the apple score cluster at time point i. The variable A_2j will be used to represent the apple decision at time point 2 relating to the end outcome of Y_j. In this scenario the expected utility can be formulated as:

$$\begin{aligned}\mathbb{E}[U] &= P(V_{21}) * P(A_{21}) * P(V_{31}) * U_1 + P(V_{21}) * P(A_{21}) * P(V_{32}) * Y_2 + \dots \\ &\quad + P(V_{22}) * P(A_{22}) * P(V_{38}) * U_8 \\ &= \sum_{j=1}^{8} P(V_3 = V_{3j}, A_2 = A_{2j} | V_2 = V_{21}) * U_j \\ &= \sum_{j=1}^{8} P(A_{2j} | V_2 = V_{21}) * P(V_3 = V_{3j} | A_2 = A_{2j}) * U_j\end{aligned}$$

One important point to note, when referencing these formulae, is that only one route to each end node has a positive probability $p \in (0, 1]$ of occurring, therefore inside the sum when referring to the route with respect to each U_j.

Expanding on this theory to consider node 1 from Figure 17, the theory will be generalised to hold for the first node and then indeed any node in the tree. In each case, the expected yield for each route will be all possible utility outcomes multiplied by their associated probability

of occurring.

Applying this concept to the first node, there are 32 possible end utilities that can be reached from this node, each of which has only one route which returns a positive probability. Then the expected utility is given by:

$$\begin{aligned}\mathbb{E}[U] &= \sum_{j=1}^{32} P(V_3 = V_{3j}, A_2 = A_{2j}, V_2 = V_{2j}, A_1 = A_{1j}, V_1 = V_{1j}) * U_j \\ &= \sum_{j=1}^{32} P(V_3 = V_{3j}|A_2 = A_{2j}) * P(A_2 = A_{2j}|V_2 = V_{2j}) * P(V_2 = V_{2j}|A_1 = A_{1j}) \\ &* P(A_1 = A_{1j}|V_1 = V_{1j}) * P(V_{1j}) * U_j\end{aligned}$$

Therefore, for a given node n at time point $t \in \{1, 2\}$, the formula for expected utility can be generalised as:

$$\mathbb{E}[U] = \sum_{j=1}^{2^{(6-t)}} P(V_3 = V_{3j}|A_2 = A_{2j}) * P(A_2 = A_{2j}|V_2 = V_{2j}) * \cdots * P(V_t = V_{tj}) * U_j$$

Should the number of time points considered increase, the formula could be expanded to accommodate this although, as above, the final node, in this case at time point 3, could not be incorporated due to the ... used in the formula. Then, for a decision process with n time points, time points in $t \in \{1, 2, \ldots, n-1\}$, the expected utility can be summarised as:

$$\begin{aligned}\mathbb{E}[U] = \sum_{j=1}^{2^{(n+1-t)}} & P(V_n = V_{nj}|A_{n-1} = A_{(n-1)j}) * P(A_{n-1} \quad = A_{2(n-1)j}|V_2 = V_{(n-1)j}) * \\ & \cdots * P(V_t = V_{tj}) * U_j\end{aligned}$$

If the data were to be collected, the expected output would be that the expected utility would increase with respect to pesticide used up until a certain threshold. At this point, the increased cost of pesticide would not return sufficient increase in apple yield to generate a profit to the orchard owner. The expected yield of each pesticide protocol could be determined for each time point, which would allow for the optimised protocols to be determined solely from an economic perspective. If future data included rainfall, frost or other natural events, an optimised strategy could be determined for differing weather scenarios. It would also allow for risk seeking or adverse strategies to be investigated.

An example of this could be rainfall which, if it occurs shortly after pesticide has been applied causes the pesticide to be diluted, and can therefore provide less benefit to apple yield. To overcome this, pesticide could be applied more than once in a time period. Whilst this would increase expected apple yield it would also increase the costs incurred.

14.3.4 Ecologist vs Economist Trade-off

An ecologist vs economist perspective takes both the apple progress score and bee abundance into consideration when deciding which pesticide protocol to use.

The main differences in the decision tree is the pesticide decision being dependent on both the apple yield (a) and bee abundance (b). Here, higher order Markov theory [2] can be observed rather than the standard Markov theory used in the previous section when deciding which pesticide protocol cluster to follow. In this case, the pesticide decision depends on both the bee abundance and apple score from the previous time point. In this case D_{i1} andD_{i2} are used to demonstrate whether the combined decision outcome from both perspectives was for positive progress or negative progress. For a given time point i, A_i denotes the apple event for that time point and a_i denotes the score obtained. Likewise with bee abundance B_i denotes the bee event and b_i the abundance observed. Then as an example the probability that at time point $t = 2$ there is a positive outcome for progress, given the starting node was node 1, can be denoted by:

$$\begin{aligned} P(D_2 = D_{21}|N = 1) &= P(D_2 = D_{21}|a_2 = a_2, B_2 = b_2, V_2 = V_{21}, D_1 = D11, \\ &\quad A_1 = a_1, B_1 = b_1, V_1 = V_{11}) \\ &= P(D_2 = D_{21}|A_2 = a_2, B_2 = b_2) \end{aligned}$$

This demonstrates the higher order Markov property with $m = 2$ and shows how the dependence of the outcome for the pesticide decision relies on just the apple progress score and bee abundance.

Then based on this decision (D_{21}), the next step was to apply either a high or low pesticide protocol at the next decision step. The values for both the bee abundance and apple progress score will follow the standard Markov process as in the previous scenarios. This is demonstrated in Figure 18, below:

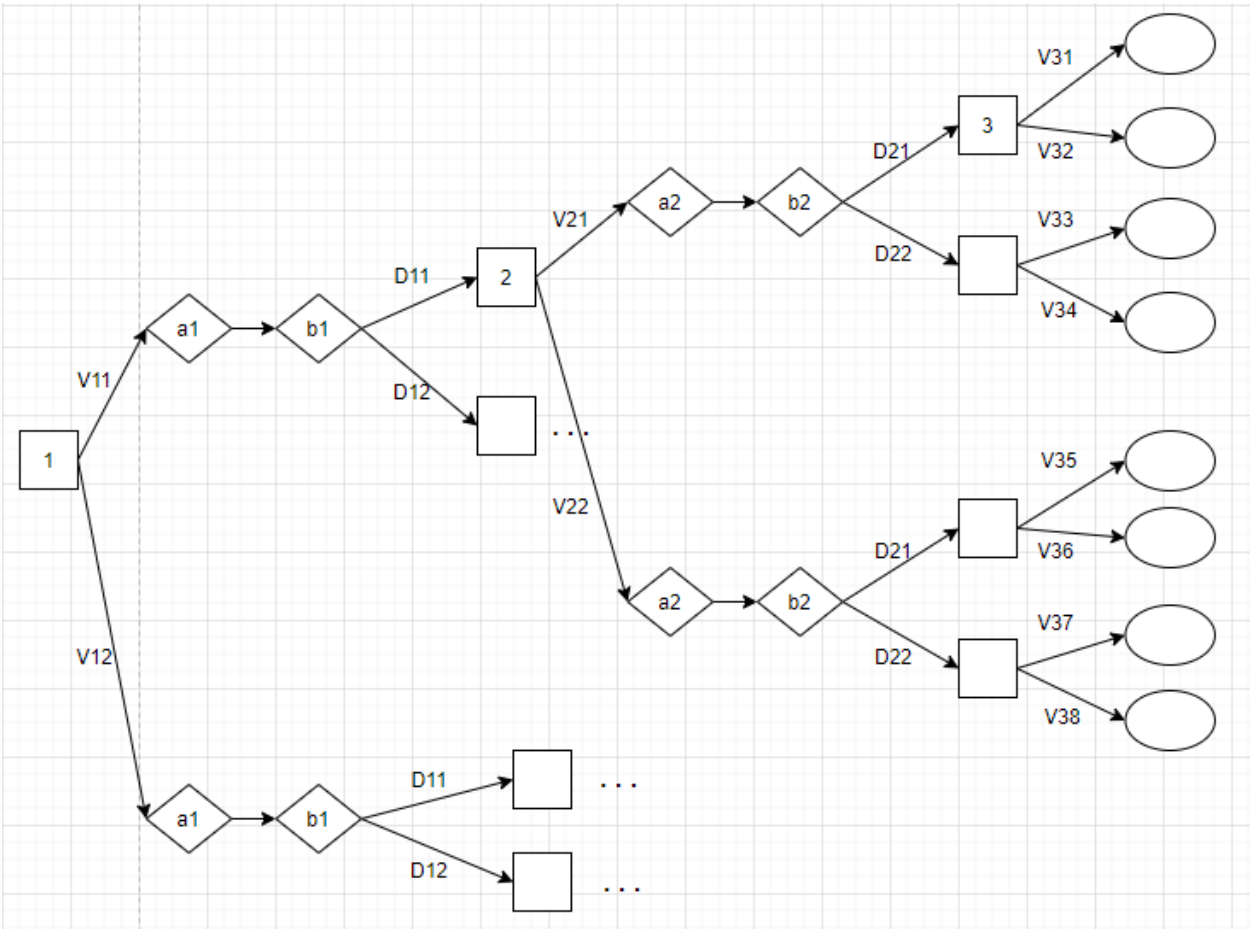


Figure 18: Ecologist vs Economist Trade-off

Previously EMV theory was used when eliciting the utility functions, where the best outcome returns the highest expected bee abundance or economic gain. However, combining them to be used in a utility which contains both economic gain and bee abundance is difficult. This is because it is hard to give a quantifiable monetary value to bee abundance as differing perspectives value it differently.

The first step for this problem is to define the probability space $(\Omega, \mathcal{F}, P)$, where Ω is the set of all possible outcomes for all variables at all time points $\{V_{ij}, i \in \{1,2,3\}, A_i, i \in \{1,2,3\}, B_i, i \in \{1,2,3\}, D_{ij}, i \in \{1,2,3\}, j \in \{1,2\}, Y_i, i \in \{1,\ldots,32\}\}$, $\mathcal{F}$ is the powerset of Ω, and P is a probability measure, returning the probability of an event occurring, which in this case is the number of orchards in which that route is followed, out of all orchards surveyed.

There are many ways to elicit the outcome of an orchard using a utility with multiple attributes, where the weighting of each attribute allows the utility function to vary depending on the end user's objectives. The general structure across all combinations will be the same,

with values of α and β varying the weighting of the utility function.

The outcome U_j is the utility of a realised path j within the decision tree and can be summarised by using the following three variables:

1. The utility gain from bee abundance, which is dependent on the bee abundances observed at each time point: $(B_j = \frac{1}{3}\sum_{i=1}^{n} B_{ij})$
2. Apple yield monetary gain: (Y_j)
3. Total cost of pesticides: $(C_j = \sum_{i=1}^{3} C_{ij})$

Then, using α and β to allow for the weightings in the utility function, the formula becomes:

$$U_j = \alpha(B_j) + \beta(Y_j - C_j)$$

The expected utility of a node is then the summation of the probabilities of all the paths from that node multiplied by their associated utility:

$$\mathbb{E}[U] = \sum_{j=1}^{n} P(U_j) * U_j$$

A similar methodology will be used as in the economist's section. The subscript j will reference the values associated with the route taken for outcome utility j, for node 1 in figure 18, the expected utility is:

$$\begin{aligned}
\mathbb{E}[U] = &\sum_{j=1}^{32} P(V_3 = V_{3j}, D_2 = D_{2j}, A_2 = a_2, B_2 = b_2, V_2 = V_{2j}, D_1 = D_{1j}, \\
&A_1 = a_1, B_1 = b_1, V_1 = V_{1j}) * U_j \\
&= \sum_{j=1}^{32} P(V_3 = V_{3j}, D_2 = D_{2j}, V_2 = V_{2j}, D_1 = D_{1j}, V_1 = V_{1j}) * U_j \\
&= \sum_{j=1}^{32} P(V_3 = V_{3j} | D_2 = D_{2j}) * P(D_2 = D_{2j} | A_2 = a_2, B_2 = b_2) * \\
&P(A_2 = a_2 | V_2 = V_{2j}) * P(B_2 = b_2 | V_2 = V_{2j}) * P(D_1 = D_{1j} | A_1 = a_1, B_1 = b_1) * \\
&P(V_2 = V_{2j} | D_1 = D_{1j}) * P(A_1 = a_1 | V_1 = V_{1j}) * P(B_1 = b_1 | V_1 = V_{1j}) * P(v_1 = V_{1j}) * U_j
\end{aligned}$$

In a similar way, the expected utility can also be expanded for n time points, where time points $t \in (1, \ldots, n-1)$, as summarised by:

$$\mathbb{E}[U] = \sum_{j=1}^{2^{(n+1-t)}} P(V_n = V_{nj}|D_{n-1} = D_{(n-1)j}) * P(D_{n-1} = D_{(n-1)j}|A_{n-1} = a_{n-1}, B_{n-1} = b_{n-1})$$
$$* P(A_{n-1} = a_{(n-1)j}|V_{n-1} = V_{(n-1)j}) * P(B_{n-1} = b_{(n-1)j}|V_{n-1} = V_{(n-1)j}) \cdots * P(V_t = V_{tj}) * U_j$$

It can be seen that the individual ecologist and economist perspectives are special cases of the economist vs ecologist approach. Setting values of α or β equal to zero returns the economist and ecologist perspectives respectively. This demonstrates the additive property of the utility model considered and highlights the potential flaws when a decision maker follows a utility close to the boundaries, as it allows for the variation of one attribute to have little affect on the end utility output. The benefits, when the values are not at the extremities, is it allows for differing perspectives to be elicited easily and accurately. To account for the edge cases, the utility function could be adapted to have minimum thresholds for both bee abundance and apple yield before a positive utility can be obtained, but these values would need to be decided upon by the decision makers in advance of any statistical analysis. The independence assumption allows for further attributes to be added in the future to accommodate other decision factors in building a more complex utility function.

14.3.5 Applications and Conclusion

The applications and potential benefits of decision theory are clear. As demonstrated, there exist statistical methodologies that could be used to model the ecologist vs economist trade off, and provide insight into the optimal strategies under varying utilities and priorities.

This section further demonstrates the benefits to statistical analysis in the agricultural sector if better collection standards were in place. If future analysis followed some of the standards suggested, it would allow data collected from multiple studies to be collated and used to enhance understanding further.

The collection of more environmental variables such as rainfall would allow for extra dependencies to be built into the decision tree, with decisions then following the higher order Markov process with $m > 2$. It would also enable an end-user to be risk-seeking or risk-adverse when evaluating their future strategies.

The possibility to enable orchard owners the option to decide their pesticide protocol based on maximised utility theory would enable them to hit their profit targets in a way which is least harmful to the wild bee populations. As a concept, taking the application developed further to incorporate the expected bee abundances and profit, based on their utility preferences, would be widely beneficial and could help start a trend towards more sustainable apple growing. Apple growing is a small subsection of the agricultural industry and if proven to be successful here, it could potentially be expanded to the agricultural industry as a whole.

15 Final Conclusions

The dissertation was a success. The original questions posed in the specification were answered, as far as they could be with the data provided. The potential benefits of decision theory in the field of agriculture were clearly demonstrated, from a theoretical perspective (Section 14.3.5) and it was shown that with an expanded dataset, optimal strategies for varying utility perspectives, could also be demonstrated using this methodology.

Due to the constraints the original research put on the data, a shift in focus also demonstrated the use of machine learning could be used to replace field expertise and associated domain knowledge in orchards (Section 13). It would be expected this learning could be generalised to the agricultural industry in general, although there would likely be some scenarios in which machine learning could not replace human knowledge. The machine learning also demonstrated that when considering datasets with a low number of data points, the effect of sparsity is a big influence. The analysis undertaken showed, of the methods considered, Agglomerative hierarchical clustering using the Ward distance metric with the Maximum Linkage functions, performed best under these conditions.

This dissertation also provides good evidence that a future study with an increased number of orchards, and data collection methods, as outlined in Section(14.1), could provide real steps towards the goal of sustainable agriculture. In particular, looking at optimising decision theory processes for orchards, to allow optimised strategies to be determined more dynamically, could allow decision theory to reach an end-user on a more personal level. If this concept could be standardised across the agriculture sector, it would make a significant, positive contribution to how agriculture is approached.

References

Ah-Pine, J. and X. Wang (2016). Similarity based hierarchical clustering with an application to text collections. *15th International Symposium on Intelligent Data Analysis (IDA 2016)*, 320–331.

Brownsey, S. (2019a). Wild pollinators application. [online] Available from: accessed 01/03/2020: `https://brownsey.shinyapps.io/shinyapp/`.

Brownsey, S. (2019b). Wild pollinators github. [online] Available from: accessed 26/03/2020: `https://github.com/Brownsey/wild_pollinators`.

Documentation, R. (2019). nbclust function. *R Documentation*. [online] Available from: accessed 26/01/2020: `https://www.rdocumentation.org/packages/`.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Park, M. G., E. Blitzer, J. Gibbs, J. E. Losey, and B. N. Danforth (2015). Negative effects of pesticides on wild bee communities can be buffered by landscape context. *Proceedings of the Royal Society B: Biological Sciences 282*(1809), 20150299.

Reynolds, S. (2019). Insects: species that prefer crops prosper while majority decline. [online] Available from: accessed 12/02/2020: `https://theconversation.com/insects-species-that-prefer-crops-prosper-while-majority-decline-114206`.

Ross, S. (2019). Stochastic processes.

RStudio, Inc (2013). *Easy web applications in R.* [online] Available from: accessed 06/11/2020: `http://www.rstudio.com/shiny/`.

Speake, C. (2019). Declining Bee Population, How Important are They? [online] Available from: accessed 26/01/2020: `https://thegardeningcook.com/the-importance-of-bees-in-nature/`.

Thomson, D. (2019). *Effects of long-term variation in pollinator abundance and diversity on reproduction of a generalise plant.*

Time (2015). White house unveils plan to save the honeybees. [online] Available from: accessed 26/01/2020: `https://time.com/3888147/white-house-honeybees-plan/`.

Time (2017). More than 700 North American Bee Species Are Headed Toward Extinction. [online] Available from: accessed 20/01/2020: `https://time.com/4688417/north-american-bee-population-extinction/`.

USDA (2019). Report on the national stakeholders conference on honey bee health. [online] Available from: accessed 15/01/2020: `https://www.usda.gov/sites/default/files/documents/ReportHoneyBeeHealth.pdf`.

Wikipedia. The elbow method.

16 Appendix - Code

There are about 3,000 lines of code used for the investigation and analysis of this project. For this reason, no code will be shown in this final report. The repository containing the code can be found on the project GitHub page **(GitHub)** and key findings are presented in the application: **(App)**. The key files are also located in the code .zip file associated with this submission.

17 Appendix - Specification

The original project specification is attached to demonstrate how the progressed according to the original specifications.

Data Science Project: CS350

Application of Decision Theory to Wild Pollinators in Apple Orchards

Author

Stephen Brownsey

Supervisor

Julia Brettschneider
Department of Statistics

November 18, 2019

Contents

1 Introduction

The research will be based on a white paper by Mia G. Park et al 'Negative effects of pesticides on wild bee communities can be buffered by landscape context'[1], in which the impact of conventional pesticide use and percentage of natural area treated, affected pollinating bee communities. This paper evaluated overall pesticide use, as well as the specific impact of pesticide compound class (e.g., fungicide, insecticide), and the timing of application on wild bee abundance, wild bee species richness and honeybee abundance. The authors also assessed the effects of landscape context, and the effects of temperature on wild bee communities.

Due to agricultural intensification over the past century, the current approach, by growers of the 19 conventional apple orchards included in the research article, was to apply multiple classes of compounds before, during, and after the orchard blooming season. Fungicides are usually applied just before or during bloom, when the rainfall is highest, to prevent the spread of fungal pathogens. Insecticides are applied after the flowering window to minimise the effects on pollinators. Thirdly, plant growth regulators (PGRs, thinners), are commonly applied after the 'June drop', when the trees naturally shed some fruitlets, to thin the crop, thereby avoiding branch damage and increasing fruit size. In the conventional setting investigated, pollination was by strategic placement of honeybee hives throughout the orchard. When this is the case, insecticides and PGRs are only applied after the hives have been removed from the orchard. However, as the trees are rarely perfectly synchronised, timing of this application can only be approximate, affecting wild pollinators and their services. Understanding exactly how this affects the wild pollinators could lead to having a better grasp on how to maximise their use and longevity in the farming industry.

The analyses in the current project will be based on a decision theory approach to the data, with the goal of answering additional questions about optimal management, risk assessment and other aspects relating to the impact of various land management policies. Whilst most models typically adopt an industrial perspective, focussed primarily on maximising the yield with minimal costs, and little or no thought for supporting the ecosystems, the approach proposed for this project will have a different focus. The analyses will investigate whether taking a neutral stand will allow the construction of decision models that take both the industrial and environmental perspectives into account.

2 Objectives/Problem

Mathematical and biological objectives will be considered in the analyses presented in this project.

1. The mathematical and statistical goal is to analyse the data and use the findings to build decision models in an agri-environmental context.
   (a) Organise the variables and conduct comprehensive exploratory data analysis.
   (b) Develop decision models to analyse how the choice of pesticides, and the order in which they are deployed, can impact on the richness and abundance of wild bees.
   (c) Fit the models using the available data.
   (d) Compare the results obtained by alternative decision rules.
2. From an agri-environmental context the goal is to convey the decisions in a way that others can understand.
   (a) Quantify the implications of different priorities such as high yield (land manager's perspective) versus biodiversity (ecologist's perspective), and suggest trade-off strategies. In an ideal world the strategies would lead to both a high yield of apples and minimal harm to wild bees.
   (b) Develop a user-friendly prototype for a web application that a grower could use involving the decision process, with variables that may be altered to allow visualisation of how varying decision criteria would affect both bees and crop yield.

If time permits, extensions to the project may focus on use intensity, weather conditions, and orchard scaling factors. For a programming extension, a CRAN ready R package could be developed in addition to the web application.

3 Data and Methods

3.1 Data

The data collected in the paper[1] contain compound class (fungicide, insecticide, thinner), intensity of compound applied, broken down by bloom timings (before, during, after), temperature levels, flowering levels and characteristics of the orchard and nearby natural habitats. There were nineteen orchards involved in the analysis, sixteen of which had data collected over two consecutive years. To start with we can consider the three timings (before, during and after) and several options for applying the compounds (fungicides, insecticides and thinners). From this models will be built to study how order and choice of pesticides impact the richness and abundance of wild bees

3.2 Methods

For the exploratory data analysis summaries and various plots of all the variables will be generated to try and understand how they interact and their underlying distributions.

The methodology will involve Bayesian decision theory, which, unlike multi-criteria decision analysis (MCDA), will address the complex ecological networks within the model itself, by organising sufficiently flexible structures into decision trees.These will consist of arranged stages and choice sets.

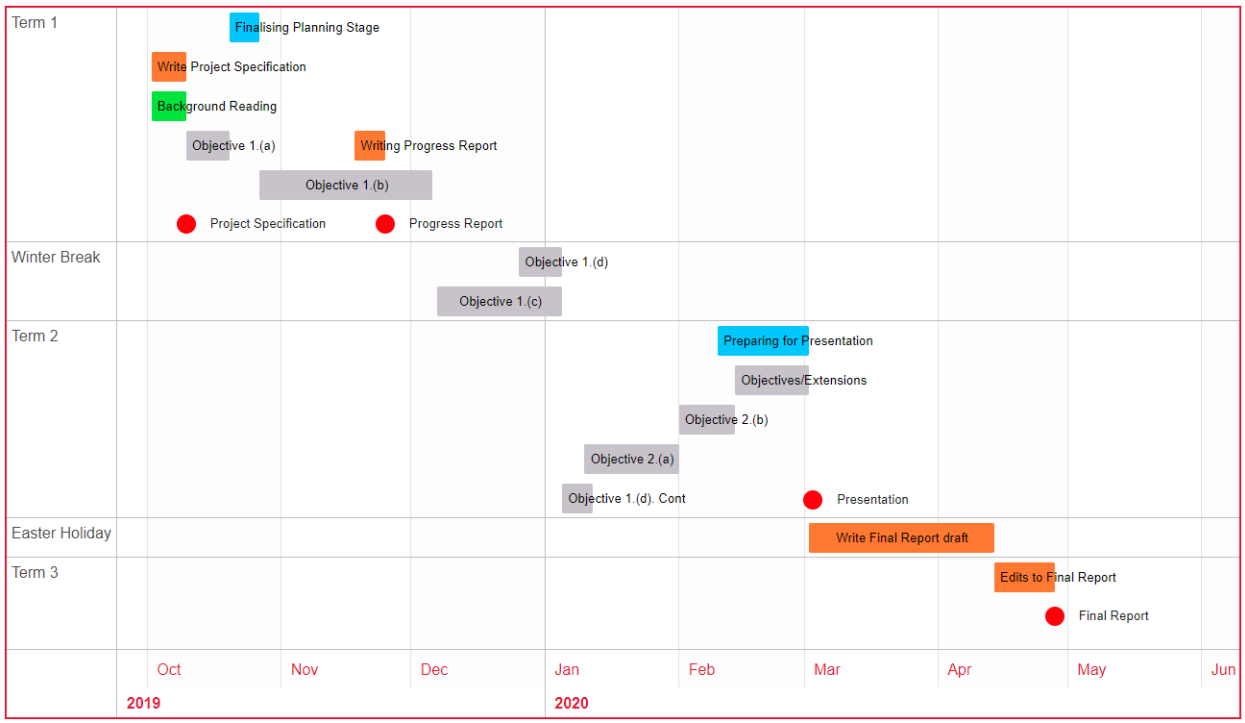
These models will then be used to study how the choice of pesticides, and the order in which they are applied, impact the richness and abundance of wild bee populations. As the project progresses the models will be iteratively developed, to take more considerations into account.

To create a web application (app), Rshiny [2] will be used, as it is a great way to embed R code into an interactive webpage that potential users can access on their machine via the free app sharing platform; shinyapps.io. This app will have a user facing panel with variables that the end-users will be able to vary. The backend R code will then apply the decision model, taking the userâs choices into account and visually showing them the results on their screen.

4 Project Road Map and Timetable

1. Complete background reading on the original study to understand the authors'original study goals and analysis.[1]
2. Complete background reading on decision theory models using books and materials [3] .
3. Undertake some exploratory analysis, generating a variety of summaries and graphics to help visualise the data and get a better feel for the dataset (Objective 1.(a)).
4. With this new understanding, the plan for developing the decision theory models will be formalised.
5. With the plan formalised, work will begin on objective 1.(b).
6. The first progress report will be written to summarise the work undertaken in points 1-5 above.
7. The results and findings from objective 1.(b) will be used and built on, to bring in measures for comparing alternative decision rules.
8. These models will then be applied to the data (objective 1.(c)), and the results analysed (objective 1.(d)).
9. Research will continue by running varying decision models with differing weights on each priority. These models will be used to convey the differences between perspectives 2.(a).
10. All code used in the above steps will be optimised to allow for the building of the prototype Rshiny application. This will then be web-hosted on shinyapps.io (objective 2.(b)).
11. Time will be factored in to go over previous research and finalise any unfinished areas or allow additional research.
12. All notes will be collated and used in preparation for the project presentation.
13. Draft final project report will be written.
14. Gather feedback on the report from the project supervisor and implement any changes.
15. Note: Extra reading will be required to implement these steps but for the purposes of the Gantt chart and task planning, this will be included within each task objective.

For the purpose of creating a Gantt chart, I developed an RShiny application[4]: The reasons for this were: unlike the online Gantt creators, the Rshiny app uses a .csv as the input file. This will make the comparison of how my work was planned, vs how it actually happened, much easier. It is also fully interactive, which helps visualise how the workload will be managed. If the deadlines and workloads change, it is very quick to edit the .csv and generate an updated Gantt chart. Finally, I have the option to add the current date onto the chart via an Rshiny button which should help me to keep track of where I need to be and ensure that I do not fall behind schedule.



5 Risk Analysis

Since I will be using both university computers and my own personal laptop throughout this project, there are risks involved when accessing files from different locations. In particular, over-writing research data and code from different machines. To mitigate this risk, I will be using GitHub as a central repository for all my code and overleaf for my reports to allow me to access the latest versions of my work wherever I am. I will also take regular backups of my work to be stored on dropbox. I will be using R and RShiny throughout the project, to mitigate package version issues. I will ensure all my package versions are the same on both machines.

References

[1] M. G. Park, E. Blitzer, J. Gibbs, J. E. Losey, and B. N. Danforth, "Negative effects of pesticides on wild bee communities can be buffered by landscape context," *Proceedings of the Royal Society B: Biological Sciences*, vol. 282, no. 1809, p. 20150299, 2015.

[2] Rshiny, "https://shiny.rstudio.com/."

[3] M. Peterson, *An Introduction to Decision Theory*. Cambridge Introductions to Philosophy, Cambridge University Press, 2 ed., 2017.

[4] S. Brownsey, "Github: https://github.com/brownsey/brownsey_gantt, hosted app: https://brownsey.shinyapps.io/brownsey_gantt/."