

**Department of Computer Science**

**BSCCS Final Year Project 2017-2018**

**Interim Report II**

|  |  |
| --- | --- |
| **(17CS095)** |  |
| **Data Mining on Ecological Data** |  |
| **(Volume 1 of 1)** |  |

**For Official Use Only**

|  |  |  |
| --- | --- | --- |
| Student Name | **:** | **Lam Wing Ting** |
| Student No. | **:** | **54038173** |
| Programme Code | **:** | **BSC4** |
|  |  |  |
| Supervisor | **:** | **Dr WONG, Ka Chun** |
| Date | **:** | **5 February 2018** |

**Student Final Year Project Declaration**

I have read the project guidelines and I understand the meaning of academic dishonesty, in particular plagiarism and collusion. I hereby declare that the work I submitted for my final year project, entitled:

Data Mining on Ecological Data

does not involve academic dishonesty. I give permission for my final year project work to be electronically scanned and if found to involve academic dishonesty, I am aware of the consequences as stated in the Project Guidelines.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student Name: | Lam Wing Ting |  | Signature: |  |
| Student ID: | 54038173 |  | Date: |  |

Table of Contents

[1. Summary of Revisions and Additions since Interim Report I 4](#_Toc505402513)

[2. Abstract 4](#_Toc505402514)

[3. Introduction 5](#_Toc505402515)

[3.1 Background Information 5](#_Toc505402516)

[3.2 Aims and Objectives 5](#_Toc505402517)

[3.3 Scope 5](#_Toc505402518)

[4. Literature Review 6](#_Toc505402519)

[4.1 Overview of Ecological Data 6](#_Toc505402520)

[4.2 Data Mining Process 6](#_Toc505402521)

[4.3 Data Visualization 8](#_Toc505402522)

[5. Methodology 9](#_Toc505402523)

[5.1 Data Collection 9](#_Toc505402524)

[5.2 Data Analysis 9](#_Toc505402525)

[6. Major Technical Components 10](#_Toc505402526)

[6.1 Data Mining 10](#_Toc505402527)

[6.2 Web Application 10](#_Toc505402528)

[7. Expected Results and Deliverables 11](#_Toc505402529)

[8. Weka Data Mining Analysis Findings 11](#_Toc505402530)

[8.1 Description 11](#_Toc505402531)

[8.2 Data Preprocessing - Clustering 11](#_Toc505402532)

[8.3 Data Preprocessing - Association 17](#_Toc505402533)

[9. Data processing in Python 17](#_Toc505402534)

[9.1 Clustering by DBSCAN 17](#_Toc505402535)

[9.2 Association 17](#_Toc505402536)

[10. Design of Web Application 17](#_Toc505402537)

[11. Preliminary Result and Future Improvement 17](#_Toc505402538)

[11.1 Data Mining Analysis 17](#_Toc505402539)

[11.2 Data visualization system (Web Application) 17](#_Toc505402540)

[12. Further Work 17](#_Toc505402541)

[12.1 Data Processing 17](#_Toc505402542)

[12.2 UI Design 18](#_Toc505402543)

[13. Problems to be encountered 18](#_Toc505402544)

[13.1 Time 18](#_Toc505402545)

[13.2 Integration of API 18](#_Toc505402546)

[14. Alterative Solution 19](#_Toc505402547)

[14.1 Test-driven approach 19](#_Toc505402548)

[14.2 Google Map API 19](#_Toc505402549)

[15. Project Schedule 20](#_Toc505402550)

[16. Monthly Log 21](#_Toc505402551)

[17. Reference 22](#_Toc505402552)

# Summary of Revisions and Additions since Interim Report I

1. Based on the “Within cluster sum of squared errors”, plotted and updated the clustering results with different number of clustering groups.
2. Association analysis findings are added.
3. DBSCAN is used for clustering analysis.
4. Figures and tables are given the original links.

# Abstract

# Introduction

## Background Information

Refer to Biology online dictionary, ecology is a science concerned with the interactions of living organisms with each other and with their environment. The statistics published by Census of Marine Life in 2011, estimated that there are around 3 million to 100 million species ranging from the land to ocean depths on earth. Besides, there are still unknown living organisms not seen by eyes. Investigation of ecosystem has been conducted by researchers for decades. Researchers spent their life long time to study any external, internal or even unknown factors affecting the ecosystem.

As a result, sinking in thousands of sparse and large data sets, it is not easy to conduct the research and obtain the results in a short period of time. In fact, pattern of researching on species can be very similar. Since the species belong to the same classification, they have similar features and also their attributes in the data sets. For example, several studies focus on birds but the attributes of data are quite similar or even the same e.g. temperature, humidity, types of bird, weather, etc.

## Aims and Objectives

The project aims to identify the similar and common data analysis practice among ecological data sets. By performing data mining on different data types like compilation, experimental, observational and time series, similar and common data mining models might be identified.

To visualize the data analysis result, a web application will be developed with API supports i.e. deck.gl developed by Uber and Google Map API. Using the APIs, the application will compare data among different data sets and display on the map. User is able to select the timespan of the data for comparison. Then, depending on the types of enquiries, display the result using maps or the charts.

When the data mining model is identified, the web application can provide some guidance to researchers. Thus, researchers can spare the time for further and deeper data analysis in their study. Also, deck.gl is newly developed by Uber since 2009 and made available in 2016. If the web application can integrate this API, it can improve the data visualization with large data sets and some interactive user events.

## Scope

Since there are more than thousands of data sets, not all data sets will be explored. Mainly focuses on the data sets of similar species with similar attributes. The data sets also include the location and coordinates for map visualization and route tracking.

# Literature Review

## Overview of Ecological Data

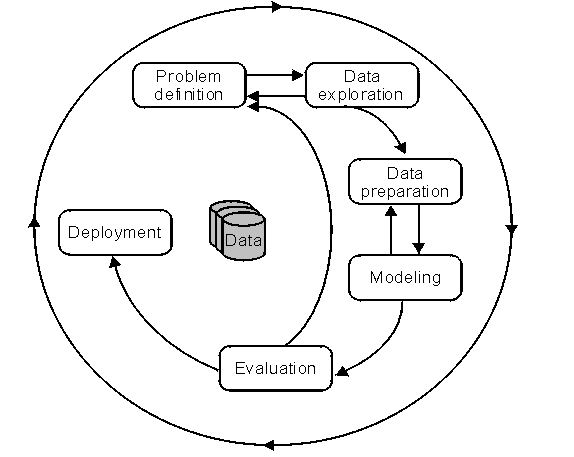
Ecology is a science concerned with the interactions of living organisms with each other and with their environment (Biology Online Dictionary, n.d.). According to the estimation described by scientists and published in BBC News (Black, 2011), there are about 8.7 million species in the world. It has been more than 1,000 years to study about ecosystem for identification and classification by scientists. Observed by Lord (Robert) May, PLoS Biology, former Royal Society president (Black, 2011), we cannot tell the number of distinct species of plants and animals even the books in US Library of Congress in 2011 reached more than 22,000,000. It showed that there is a large potential to explore more on ecological data.

Ecological Data Wiki is a popular database which scientists share their data sets about their studies. Everyone can access data sets and understand more on ecological data easily, in which in this research paper, the data sets are mainly from Ecological Data Wiki. It classifies the data sets into various categories which is close to public standard by taxonomic group, biome, data types, ecological level and scale. In big data analysis, it is of paramount importance on analysis which types of data. Different data analysis models are used for each data types. In general, the types are compilation, experimental, meta-analysis, observational and time series. UMass Landscape Ecology Lab of University of Massachusetts (Design and Analysis of Ecological Data, n.d.) introduced the design model to focus on each sampling data. It facilitates the identification of relationship between variables across data sets.

## Data Mining Process

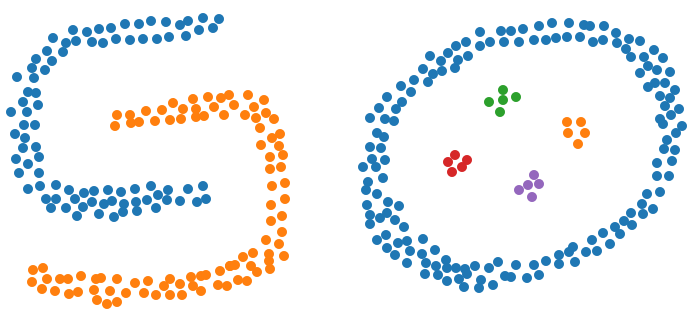
CRISP DM process model is usually adopted by scientists which is Cross Industry Standard Process. This data process is typical and widely used for big data analysis in which the details are introduced in IBM knowledge center website. It contains six phases – Problem Definition, Data Exploration, Data Preparation, Modeling, Evaluation and Deployment (The data mining process, n.d.). The following is the flow chart of data mining process:

**Figure 1 – The CRISP DM process model**



For data processing, DBSCAN is applied. Since the data is belong to spherical data such as coordinates of observation points, it is not accurate and sufficient to cater for different shapes of clustering groups (Figure 2) if using simple K-Means. But for DBSCAN, the clustering starts from the core data and connect other data not limited to their closeness. The clustering groups are dense groups of data points or coordinates. minPoint is set beforehand and check if the group include more than minPoints of data points. If yes, then it will grow the clustering recursively until all data are clustered.

**Figure 2 – shapes of DBSCAN clustering groups**



## Data Visualization

After perform data analysis, scientists use open source tool or self-developed tools for data representation. There are online data visualization tools to show the analysis results by charts, scatterplots, map, etc e.g. MATLAB, IBM Watson Data Platform and Spark. These tools are very convenient and effective to display the analysis result in a graphic way. Custom the visualization using Java, Python, etc in different representations to show trend, phenomenon and sets of rules.

However, lots of time are spent on data cleaning, preparation and learning before using the tools. The tools usually support same format of data sets and since every study has its own attributes and parameters, the tool design is to cater for general usage. Therefore, the researchers will spend some time to fit their data to the tool. This might waste some time and those time can be used for deeper learning of data. Nevertheless, everyone is using similar data analysis technique and the ways of data visualization are also almost the same. In GBIF, the website mainly visualizes data using maps. It provides support for visualizing geographical data and visitors can select the time range to display part of the data. But the functions are limited to show occurrence and location.

As a result, in this research will first investigate the data analysis work flow and algorithm more to compromise the most popular and common pattern and algorithm used in ecological data analysis. When the pattern can be identified, make a template using common attributes or parameters e.g. temperature, humidity and location. After that, using a newly developed data visualization API - deck.gl by Uber (Etherington, 2017), develop an online data visualization system. By using the template, users only need to fit their data in the template. Then, upload the template to the system. Users can select the ways to represent the data and time range to display on the charts or maps. Also, the system will include the algorithm to perform prediction. On the other hand, deck.gl has more support on user events on visualization which can help on interaction features in the system. This system will be able to perform more interaction events with users than other tools.

# Methodology

## Data Collection

Data are mainly selected from Ecological Data Wiki and Data.Gov.HK for analysis.

Ecological Data Wiki is a web platform for scientists providing their data sources around the world. In the website, ecological datasets can be found quickly with brief description of each data set e.g. timespan of the data, label of the attributes, objectives of the study, etc. Users submit their data files or any quick links on the platform which visitors can easily access to the ecological data. Wide variety of species is available ranging from amphibians to microbes.

Another source is from Data.Gov.HK. It is public sector information portal of HKSAR. Datasets are provided by government departments and public/private organizations. The portal divides several data categories and in this project, will select the data sets under environment which is provided by Agriculture, Fisheries and Conservation Department.

## Data Analysis

Python is used for data mining with open source tool called IPython with Jupyter Notebook. Association and clustering analysis will be performed in every selected data sets.

For association analysis, frequent itemset generation will be used to identify high occurrence of data attributes. For the study about route tracking, high occurrence of combination of observation points can be identified and figure out the migration route of species.

For clustering analysis, DBSCAN will be adopted. In the ecological studies, it is very common to set up observation points in different locations. With the location coordinates, observe which types of species will concentrate at the same point. DBSCAN can support different shapes of clustering groups by setting minPoints.

Upon completion of individual data analysis, evaluate the flow of data analysis of each data set. Then, figure out a data analysis model which can cater for similar data sets. Frequent itemset generation is also used which can quickly identify the common attributes used by the researchers. It helps to figure out the methodology of common ecological study.

# Major Technical Components

## Data Mining

Weka

Weka is an open source software developed by The University of Waikato and in full name which is “Waikato Environment for Knowledge Analysis”. The project team has included several standard machine learning techniques into Weka. It can be used as data pre-processing, classification, clustering, association and visualization.

IPython with Jupyter Notebook

IPython is common to use in machine learning and provide a comprehensive environment for interactive computing. Jupyter Notebook is open source web application to share documents containing code and visualizations. It also handles data cleaning, data transformation and machine learning. It supports over 40 programming languages e.g. Python. This is the reason why using IPython with Jupyter Notebook.

## Web Application

Client Side – HTML, CSS and Javascript

The web application is implemented by HTML and Javascript and CSS for the style of web pages. Basic elements of websites can be achieved by HTML, CSS and Javascript.

Server Side – Python

Since Python is also used for data mining, server side will also be written in Python. It also helps the integration of Python library.

Database Server – Oracle Database 12c

Data sets will be managed by Oracle database. It will be used for two purposes. Oracle database helps to generate different sets of data samples for testing during the development. It also stores the data imported by users and maintains the web application.

Data Visualization API - Google Map API and deck.gl by Uber

Google Map API is developed by Google and supports Javascript which is convenient for the web application. Deck.gl is developed by Uber since 2009 and designs for visualization of large data sets. It gives impressive visual results with limited effort and provides a complete architecture with Javascript support. These two APIs will be used for map visualization and display the data analysis result.

# Expected Results and Deliverables

A common work flow of data analysis used in ecological study will be identified. A web application is developed for data visualization and give guidance with the common work flow. The application also integrates with deck.gl or Google Map API to display on maps. Some interactive user events will be added, for example, users select the timespan of data and the application will generate some comparison results using charts or maps.

# Weka Data Mining Analysis Findings

## Description

Before data processing using python, data preprocessing was performed in Weka to make sure the data sets are relevant and suitable for trend prediction. In general, the aim is to compare the abundance and distribution of families of species at different locations. Then, identify the trend of the species movement and predict their intentions.

The data sets chosen mainly focus on living things such as birds, beetles, landbirds and small mammals. The species are counted at different routes or observations. After Weka data preprocessing, by comparing the results among different years, the changes in distribution and abundance of species might give a tendency of their migration.

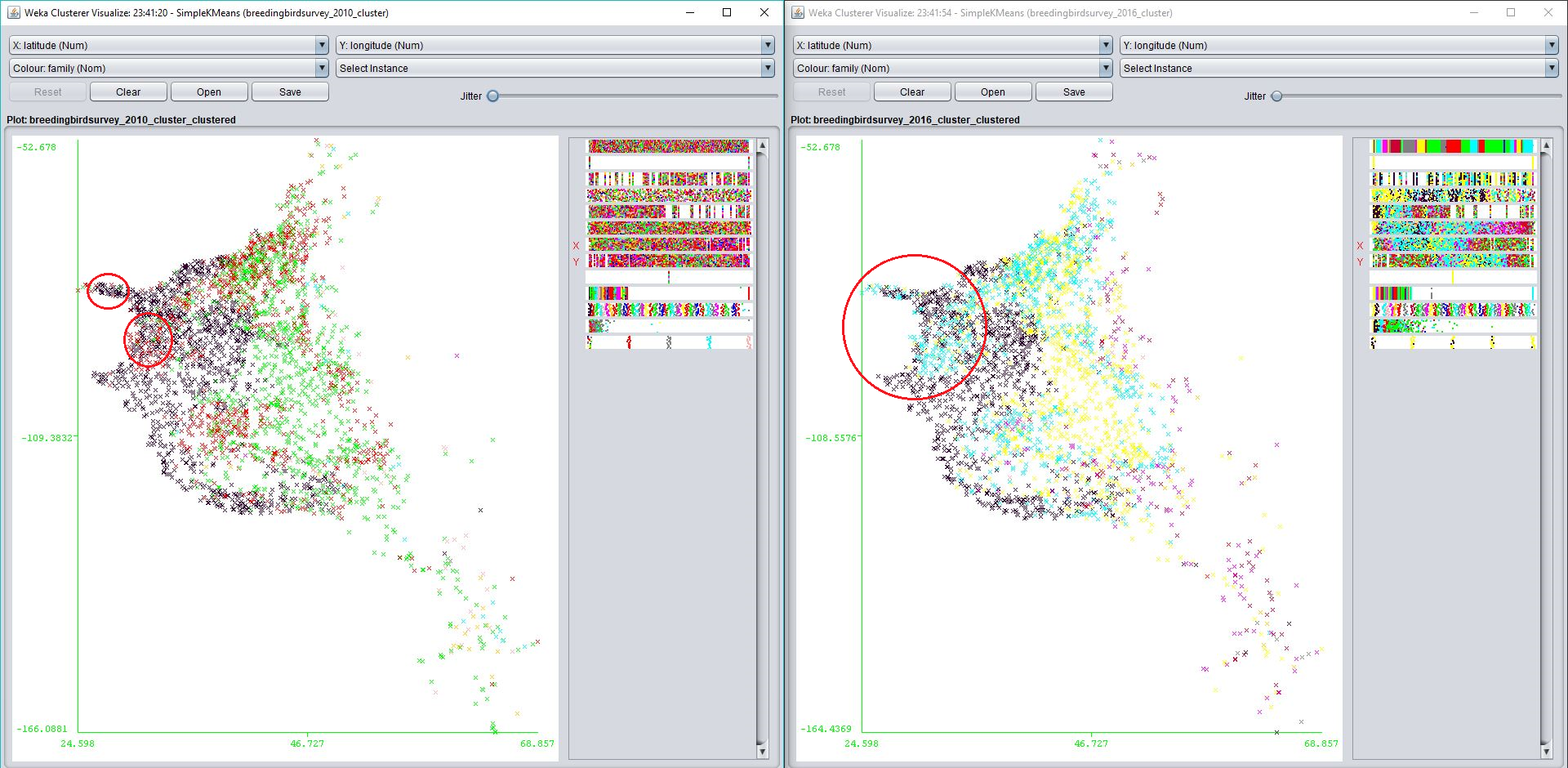
## Data Preprocessing - Clustering

In this section, data set of breeding bird survey carried in Canada and America is used for explanation. Before importing the data, integrated all data in each csv file by years and families of species.

First, compared the distribution of all families in America. The data imported are the species count of all families in 2010 and 2016. Figure 3 is comparing the distribution of all families in 2010 and 2016. Generated two clustering graphs using Weka and simple K-means clustering was selected.

Black spot in the graph represents the family – Odontophoridae. X-axis and y-axis represents the latitude and longitude of the observation stops. In 2010 (graph on the left), the black spots were denser than that in 2016 (circled in red). By comparing several years’ result, the trend of families of species can be concluded. This study has been conducted since 1966, therefore, a prediction can be made and estimate why they are moving.

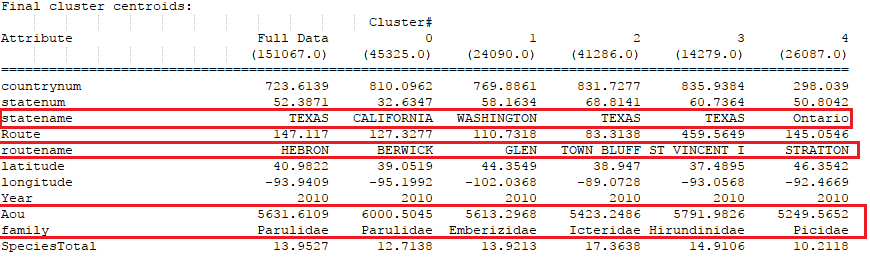
**Figure 3 – Weka Cluster Visualize for breeding bird survey in 2010 and 2016**



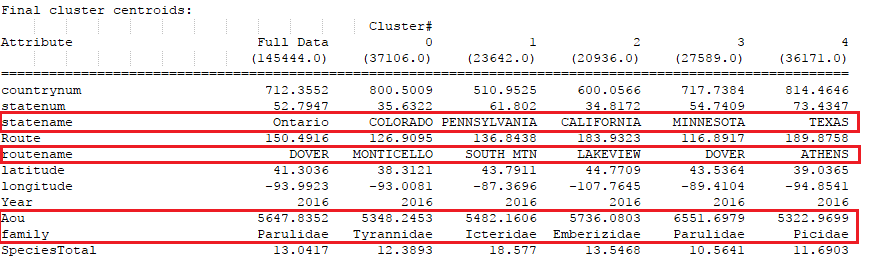
In addition, in the clustering result, we can see which state/ route concentrate which types of families. AOU is an ID assigned to each species. At that stage, number of cluster groups is 5. To determine the appropriate number of clustering groups, several numbers starting from 5 to 30 were tested several times. Finally, 5 gives a more specific conclusion among the data sets while the others did not have significant difference comparing with 5 cluster groups.

By comparing two years’ clustering result, there was a change in distribution. For example, in 2010, the family Picidae was closer to Ontario but in 2016, the family Parulidae was closed to Ontario rather than Texas.

**Figure 4 - 2010 Breeding Bird Survey Clustering Result**

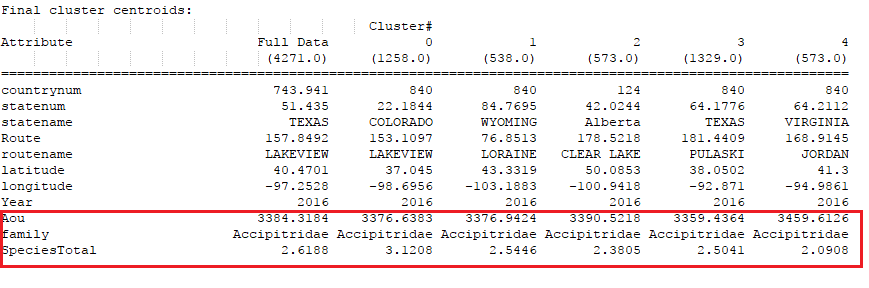


**Figure 5 - 2016 Breeding Bird Survey Clustering Result**

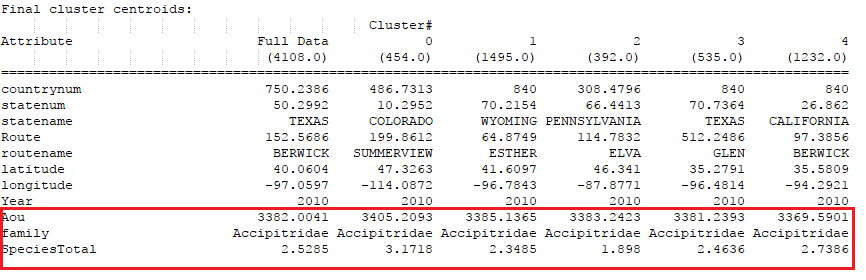
****

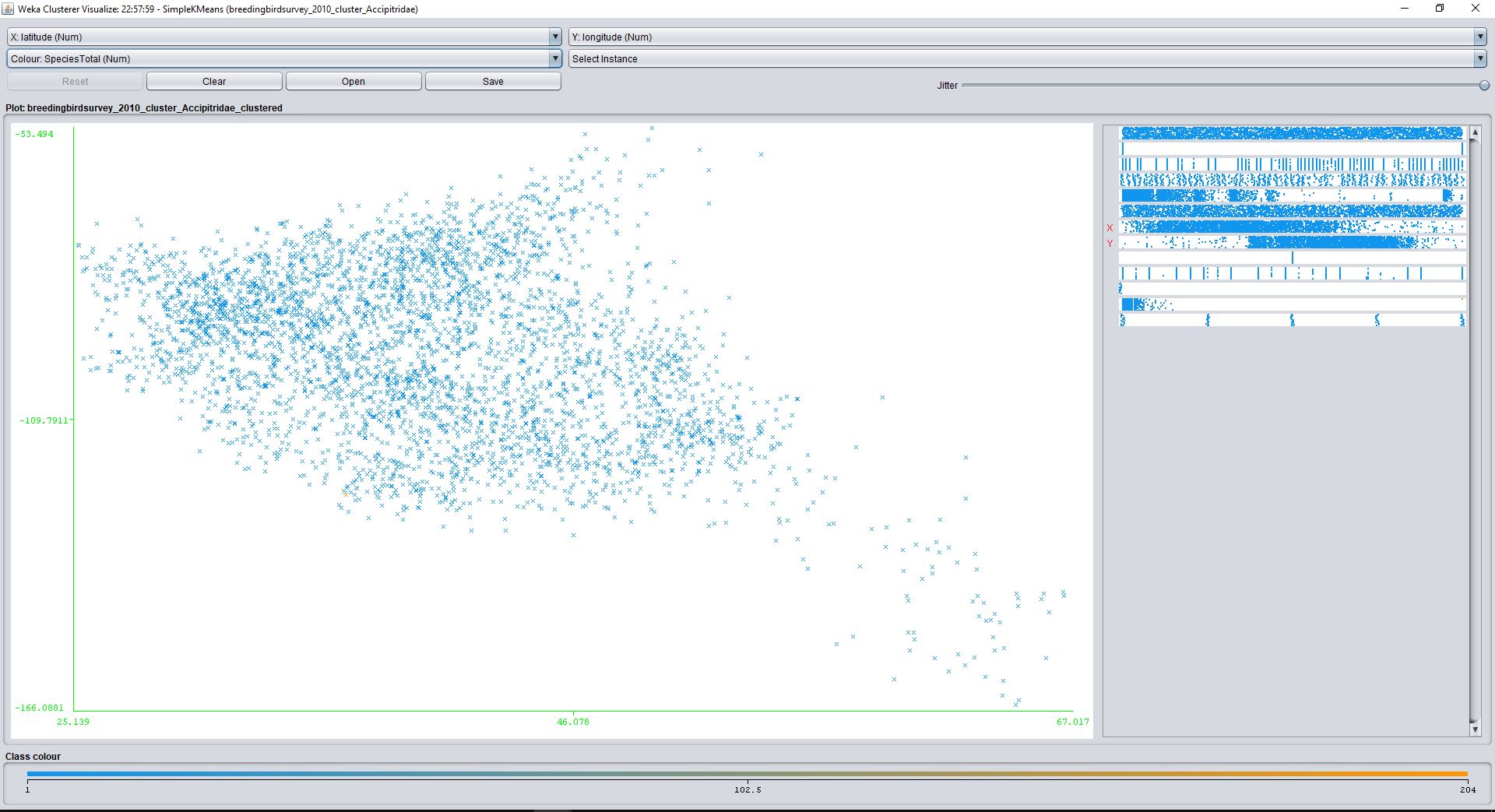
Second, compared the abundance of a family at all observation stops. Species counts of one family in 2010 and 2016 were imported and also used simple K-Means for clustering. Figure 6-9 are comparing the abundance of family Accipitridae in all observation stops in 2010 and 2016. The color of the spot represents the scale of species total count. The darker the color, less species is observed. Though more spots with light color in 2016, the total count was actually decreasing. The maximum of species total count is 105 in 2016 while its maximum is 204 in 2010. Hence, the total count of this family was decreased in 2016. Refer to the clustering group, there was a slight decrease in species total in 2016. Comparing more data in other years, the change in abundance of a family can be captured.

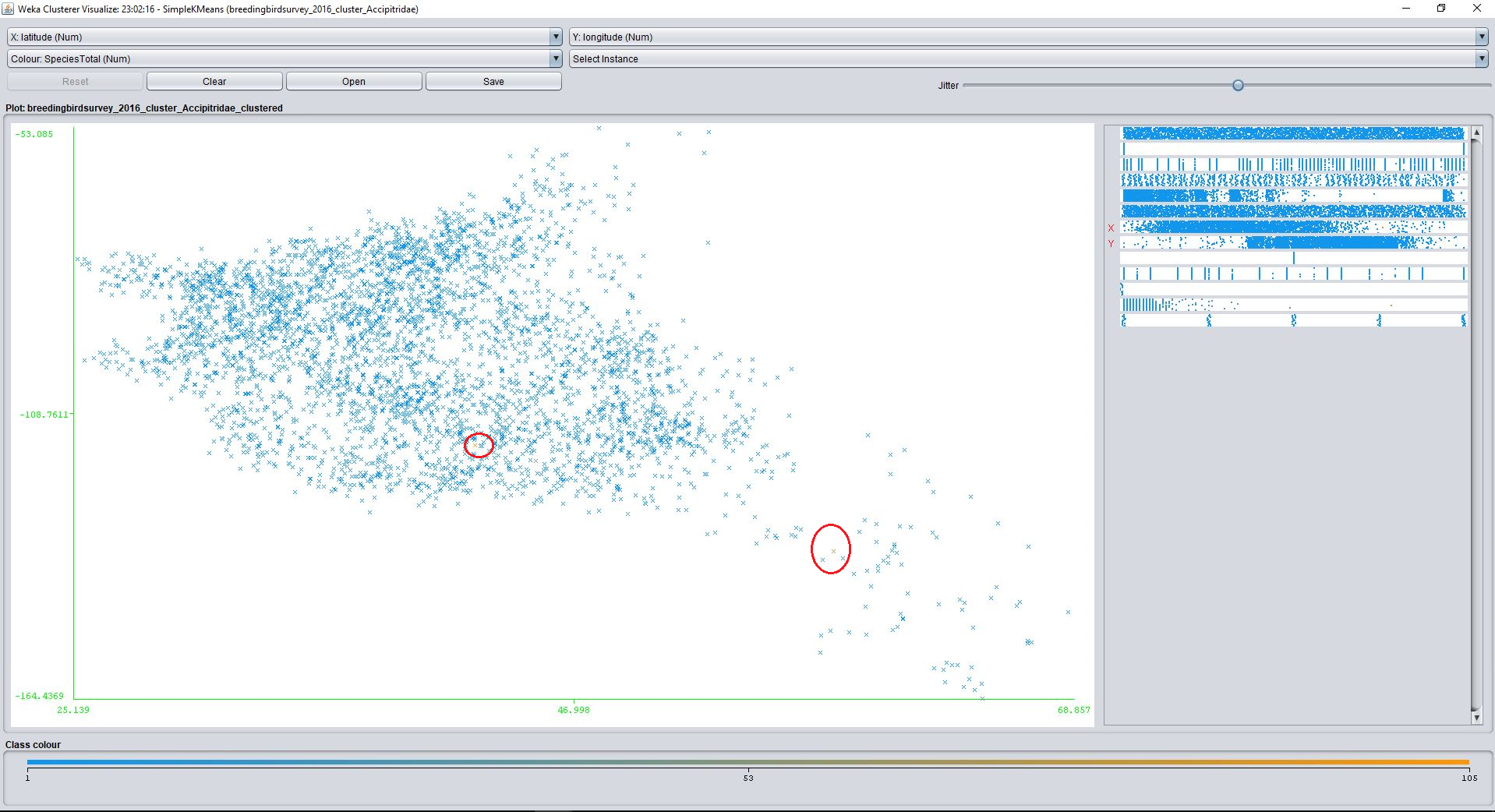
**Figure 6 - Clustering Result of Accipitridae family in 2010**



**Figure 7 - Clustering Result of Accipitridae family in 2016**



**Figure 8 – Clustering Graph of Accipitridae family in 2010 **

**Figure 9 – Clustering Graph of Accipitridae family in 2016 **

## Data Preprocessing - Association

# Data processing in Python

[Python code demo]

## Clustering by DBSCAN

## Association

# Design of Web Application

[use case diagram, sequence diagram, database diagram]

# Preliminary Result and Future Improvement

[screen cap of UI]

Mainly separate into two parts of testing – data mining analysis and web application.

## Data Mining Analysis

Use the data sets in Weka data preprocessing as trained set to test the model (DBSCAN) written in python. Though not the same algorithm is applied, the model should give similar results with data preprocessing. This is to ensure the model work fine before integrating into the visualization system.

## Data visualization system (Web Application)

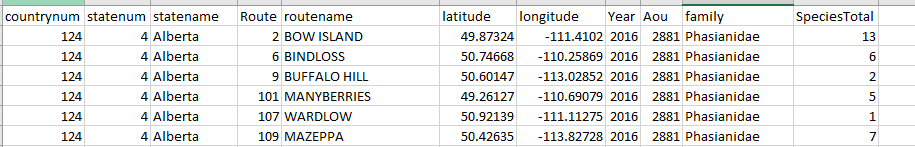
Test-driven approach will be adopted. First, test all basic components such as map visualization with dummy data imported and API functions. Second, test the visualization of data analysis result. Finally, test the trend prediction visualization. This approach can minimize complicated bug fixing and discover bugs as soon as possible before developing the next stage.

# Further Work

## Data Processing

After Weka data preprocessing analysis, captured the main attributes used to perform clustering. In data processing, attributes – coordinates of the route points, family, species total and AOU are used. Coordinates of the route points belong to spherical data which may not be so accurate if using simple K-Means clustering. Therefore, DBSCAN will be used for further analysis.

**Figure 10 – Sample data of breeding bird survey (extracted and integrated from all separated data files)**



Using DBSCAN, when investigating the distribution of each family, it can minimize the chances mixing with other families. Integrate the data sets in DBSCAN algorithm written in python and compare if similar results can be achieved in Weka. To ensure the model can be used in same types of data sets, other data sets are also prepared for further studies.

## UI Design

Further study on map visualization API is necessary. Completed Weka data mining analysis, foresee that layers of visualization might be required in the system. After completion of the data processing, integrate the mining algorithm to the system first before integration of API.

# Problems to be encountered

## Time

It is time-consuming to both development and testing on the data processing and web application. It might not have not enough time on development and testing of web application at later stage.

## Integration of API

Since the API is newly developed by Uber, within short duration of FYP, it might be difficult to study in a short period of time and apply deck.gl in the visualization system. Lots of time is required to code and test the API with the system. The analysis will require layers of data display and further study is also necessary.

# Alterative Solution

## Test-driven approach

As mentioned in previous section, test-driven approach is adopted in this project. This is to minimize any complicated bug fixing before moving on next stage.

## Google Map API

Other than deck.gl, google map API is also a choice for data visualization. In case, deck.gl is not sufficient to visualize the analysis result or encounter any technical issues. Integration of google map API is also possible to work with the web application.

# C:\Users\tcpsa\AppData\Local\Microsoft\Windows\INetCache\Content.Word\FYP Project Schedule (6).pngProject Schedule

# Monthly Log

|  |  |
| --- | --- |
| **Date** | **Description** |
| 2 October 2017 | Completed software installation (Weka and IPython) |
| 9-16 October 2017 | * Research on database and select appropriate data sets * Study the attribute of the selected data sets |
| 16-23 October 2017 | Perform Weka data mining (data pre-processing) |
| 16-31 October 2017 | * Debugging * Clean imported data to fit the Weka format * Integrate the data to python code |
| November 2017 | * Interim Report I * Further Analysis of Data Sets * Development of data mining model in IPython |
| December 2017 | * Development of data mining model in IPython * Software installation for website development * Data-preprocessing (Association) * Draft of UI design |
| January 2018 | * Set up of web application * Development of web application and data mining model * Study of API |

# Reference

A., F.Zuur, E., N.Ieno, & G., M.Smith. (2007). Analysing Ecological Data. Springer Science.

Black, R. (2011, August 23). Species count put at 8.7 million. Retrieved October 15, 2017, from <http://www.bbc.com/news/science-environment-14616161>

Design and Analysis of Ecological Data. (n.d.). Retrieved October 15, 2017, from <http://www.umass.edu/landeco/teaching/ecodata/schedule/ecological.data.pdf>

Ecology. (n.d.). Retrieved October 15, 2017, from <http://www.biology-online.org/dictionary/Ecology>

Etherington, D. (2017, April 06). Uber’s open source data visualization tool now goes beyond maps. Retrieved September 15, 2017, from <https://techcrunch.com/2017/04/06/ubers-open-source-data-visualization-tool-now-goes-beyond-maps/>

GBIF. (n.d.). Retrieved October 15, 2017, from <https://www.gbif.org/>

J. B. (2017, September 20). Classification And Regression Trees for Machine Learning. Retrieved October 15, 2017, from <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>

L., Naidoo et al. (2012). Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. ISPRS Journal of Photogrammetry and Remote Sensing, 69, 167-179.

The data mining. (n.d.). Retrieved October 13, 2017, from <https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.5.0/com.ibm.im.easy.doc/c_dm_process.html>