Department of Computer Science

BSCCS Final Year Project 2017-2018
Interim Report I

(17CS095)

Data Mining on Ecological Data

(Volume 1 of 1)

Student Name : **Lam Wing Ting**

Student No. : **54038173**

Programme Code : **BSC4**

Supervisor : **Dr WONG, Ka Chun**

Date : **6 November 2017**

For Official Use Only

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. Introduction

1.1 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.

1.2 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.

1.3 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.

2. Literature Review

2.1 Overview of Ecological Data

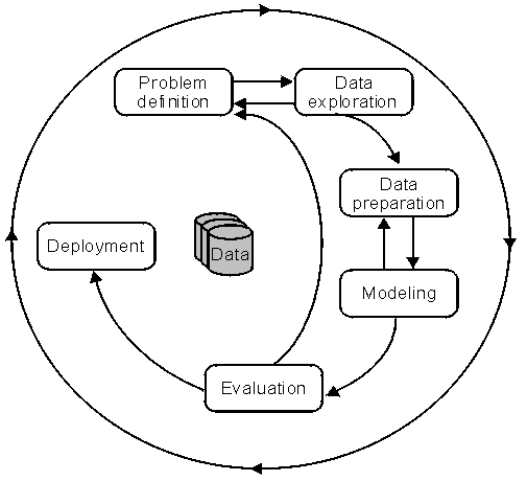
Ecology is a science concerned with the interactions of living organisms with each other and with their environment (Biology Online Dictionary, n.d.). According to the estimation described by scientists and published in BBC News (Black, 2011), there are about 8.7 million species in the world. It has 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 data base 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.

2.2 Data Mining Process

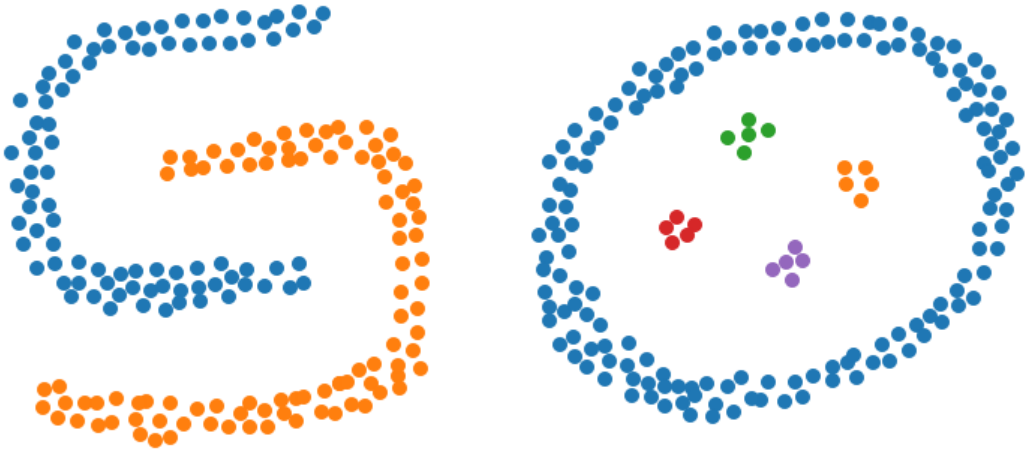
CRISP DM process model is usually adopted by scientists which is Cross Industry Standard Process. This data process is typical and widely used for big data analysis in which the details are introduced in IBM knowledge center website. It contains six phases – Problem Definition, Data Exploration, Data 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



2.3 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.

3. Major Technical Components

3.1 Data Mining

Weka
Weka is a 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.

3.2 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 – MySQL
Data sets will be managed by MySQL since some data formats are designed for MySQL.

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.

4. Weka Data Mining Analysis Findings

4.1 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.

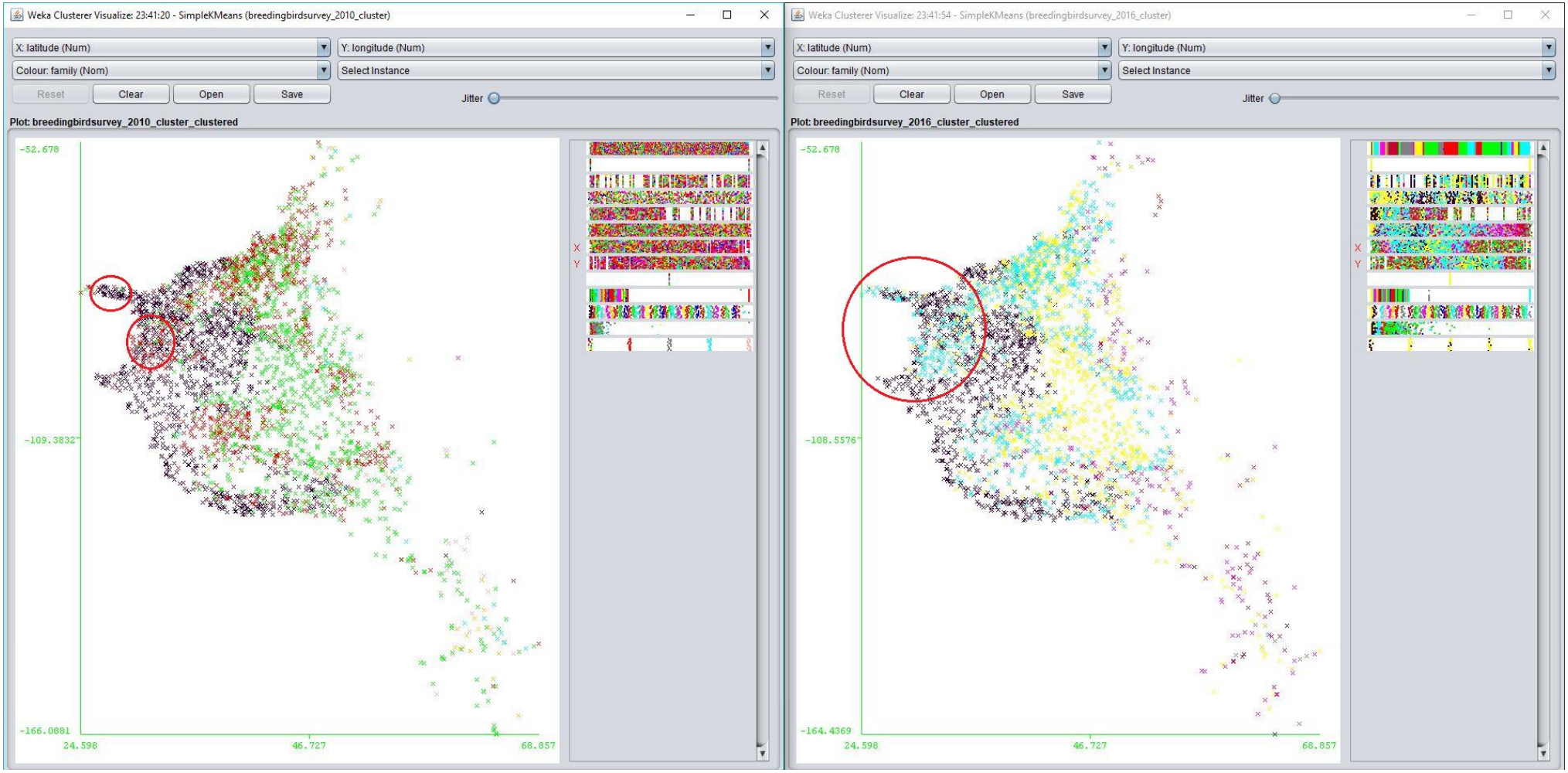
4.2 Data Preprocessing

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

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(151067.0)	(45325.0)	(24090.0)	(41286.0)	(14279.0)	(26087.0)
countrynum	723.6139	810.0962	769.8861	831.7277	835.9384	298.039
statenum	52.3871	32.6347	58.1634	68.8141	60.7364	50.8042
statename	TEXAS	CALIFORNIA	WASHINGTON	TEXAS	TEXAS	Ontario
Route	147.117	127.3277	110.7318	83.3138	459.5649	145.0546
routename	HEBRON	BERWICK	GLEN	TOWN BLUFF	ST VINCENT I	STRATTON
latitude	40.9822	39.0519	44.3549	38.947	37.4895	46.3542
longitude	-93.9409	-95.1992	-102.0368	-89.0728	-93.0568	-92.4669
Year	2010	2010	2010	2010	2010	2010
Aou	5631.6109	6000.5045	5613.2968	5423.2486	5791.9826	5249.5652
family	Parulidae	Parulidae	Emberizidae	Icteridae	Hirundinidae	Picidae
SpeciesTotal	13.9527	12.7138	13.9213	17.3638	14.9106	10.2118

Figure 5 - 2016 Breeding Bird Survey Clustering Result

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(145444.0)	(37106.0)	(23642.0)	(20936.0)	(27589.0)	(36171.0)
countrynum	712.3552	800.5009	510.9525	600.0566	717.7384	814.4646
statenum	52.7947	35.6322	61.802	34.8172	54.7409	73.4347
statename	Ontario	COLORADO	PENNSYLVANIA	CALIFORNIA	MINNESOTA	TEXAS
Route	150.4916	126.9095	136.8438	183.9323	116.8917	189.8758
routename	DOVER	MONTICELLO	SOUTH MTN	LAKEVIEW	DOVER	ATHENS
latitude	41.3036	38.3121	43.7911	44.7709	43.5364	39.0365
longitude	-93.9923	-93.0081	-87.3696	-107.7645	-89.4104	-94.8541
Year	2016	2016	2016	2016	2016	2016
Aou	5647.8352	5348.2453	5482.1606	5736.0803	6551.6979	5322.9699
family	Parulidae	Tyrannidae	Icteridae	Emberizidae	Parulidae	Picidae
SpeciesTotal	13.0417	12.3893	18.577	13.5468	10.5641	11.6903

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

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(4271.0)	(1258.0)	(538.0)	(573.0)	(1329.0)	(573.0)
countrynum	743.941	840	840	124	840	840
statenum	51.435	22.1844	84.7695	42.0244	64.1776	64.2112
statename	TEXAS	COLORADO	WYOMING	Alberta	TEXAS	VIRGINIA
Route	157.8492	153.1097	76.8513	178.5218	181.4409	168.9145
routename	LAKEVIEW	LAKEVIEW	LORAINE	CLEAR LAKE	PULASKI	JORDAN
latitude	40.4701	37.045	43.3319	50.0853	38.0502	41.3
longitude	-97.2528	-98.6956	-103.1883	-100.9418	-92.871	-94.9861
Year	2016	2016	2016	2016	2016	2016
Aou	3384.3184	3376.6383	3376.9424	3390.5218	3359.4364	3459.6126
family	Accipitridae	Accipitridae	Accipitridae	Accipitridae	Accipitridae	Accipitridae
SpeciesTotal	2.6188	3.1208	2.5446	2.3805	2.5041	2.0908

Figure 7 - Clustering Result of Accipitridae family in 2016

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(4108.0)	(454.0)	(1495.0)	(392.0)	(535.0)	(1232.0)
countrynum	750.2386	486.7313	840	308.4796	840	840
statenum	50.2992	10.2952	70.2154	66.4413	70.7364	26.862
statename	TEXAS	COLORADO	WYOMING	PENNSYLVANIA	TEXAS	CALIFORNIA
Route	152.5686	199.8612	64.8749	114.7832	512.2486	97.3856
routename	BERWICK	SUMMERVIEW	ESTHER	ELVA	GLEN	BERWICK
latitude	40.0604	47.3263	41.6097	46.341	35.2791	35.5809
longitude	-97.0597	-114.0872	-96.7843	-87.8771	-96.4814	-94.2921
Year	2010	2010	2010	2010	2010	2010
Aou	3382.0041	3405.2093	3385.1365	3383.2423	3381.2393	3369.5901
family	Accipitridae	Accipitridae	Accipitridae	Accipitridae	Accipitridae	Accipitridae
SpeciesTotal	2.5285	3.1718	2.3485	1.898	2.4636	2.7386

Figure 8 – Clustering Graph of Accipitridae family in 2010

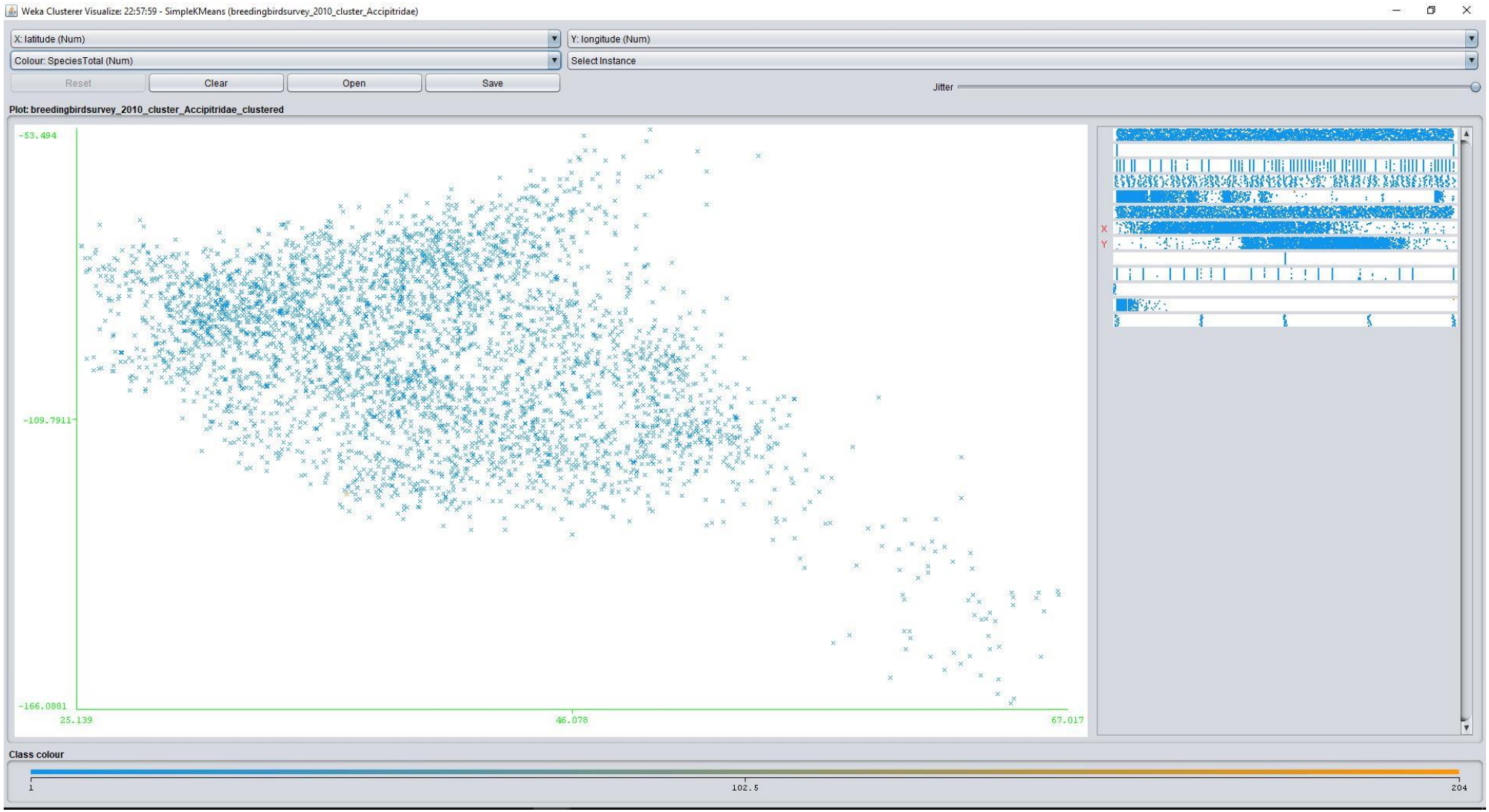
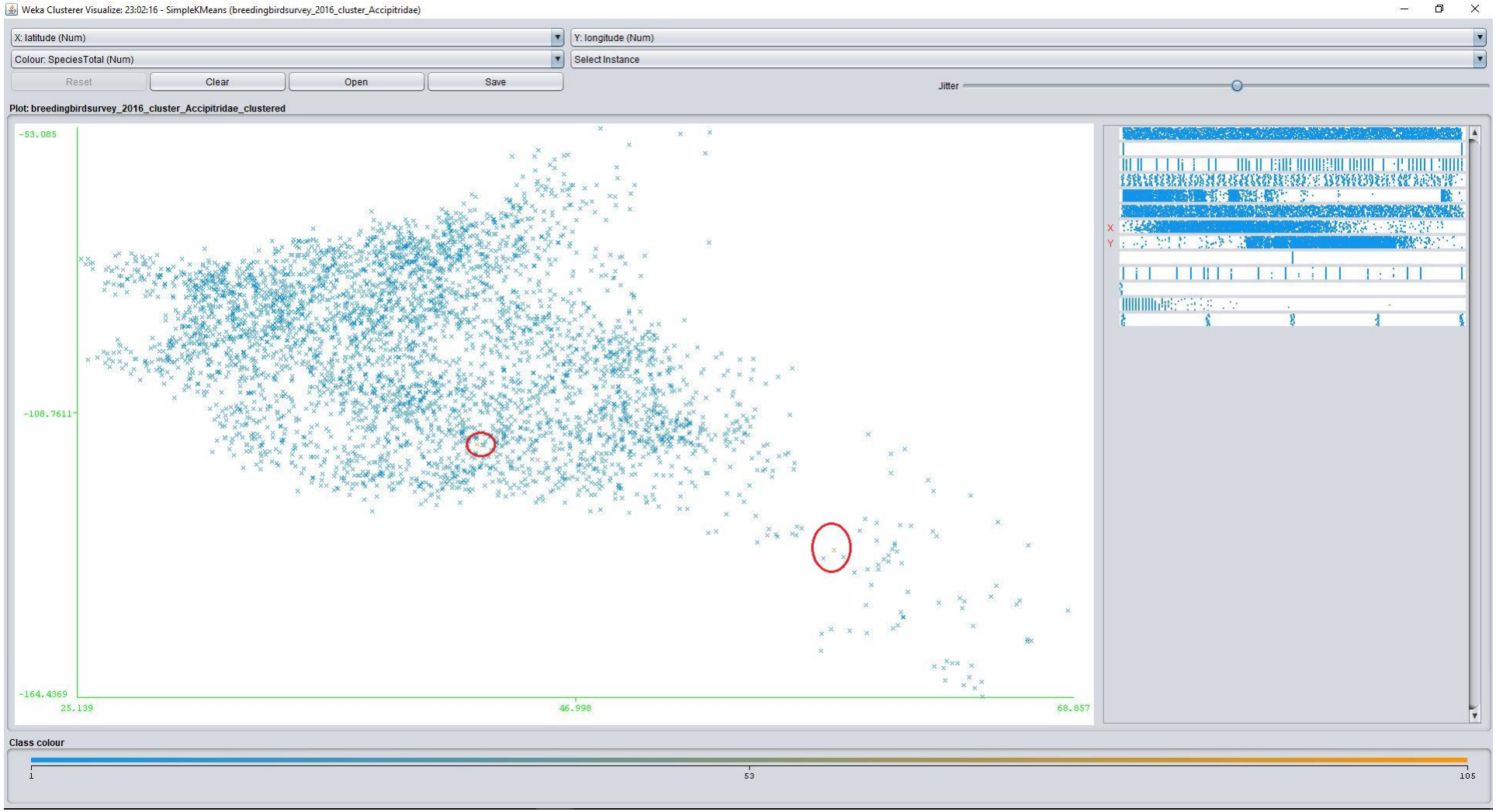


Figure 9 – Clustering Graph of Accipitridae family in 2016



5. Future Work

5.1 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)

countrynum	statenum	statename	Route	routename	latitude	longitude	Year	Aou	family	SpeciesTotal
124	4	Alberta	2	BOW ISLAND	49.87324	-111.4102	2016	2881	Phasianidae	13
124	4	Alberta	6	BINDLOSS	50.74668	-110.25869	2016	2881	Phasianidae	6
124	4	Alberta	9	BUFFALO HILL	50.60147	-113.02852	2016	2881	Phasianidae	2
124	4	Alberta	101	MANYBERRIES	49.26127	-110.69079	2016	2881	Phasianidae	5
124	4	Alberta	107	WARDLOW	50.92139	-111.11275	2016	2881	Phasianidae	1
124	4	Alberta	109	MAZEPPA	50.42635	-113.82728	2016	2881	Phasianidae	7

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.

5.2 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.

6. Test Plan

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

6.1 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.

6.2 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.

7. Problems to be encountered

7.1 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.

7.2 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.

8. Alterative Solution

8.1 Test-driven approach

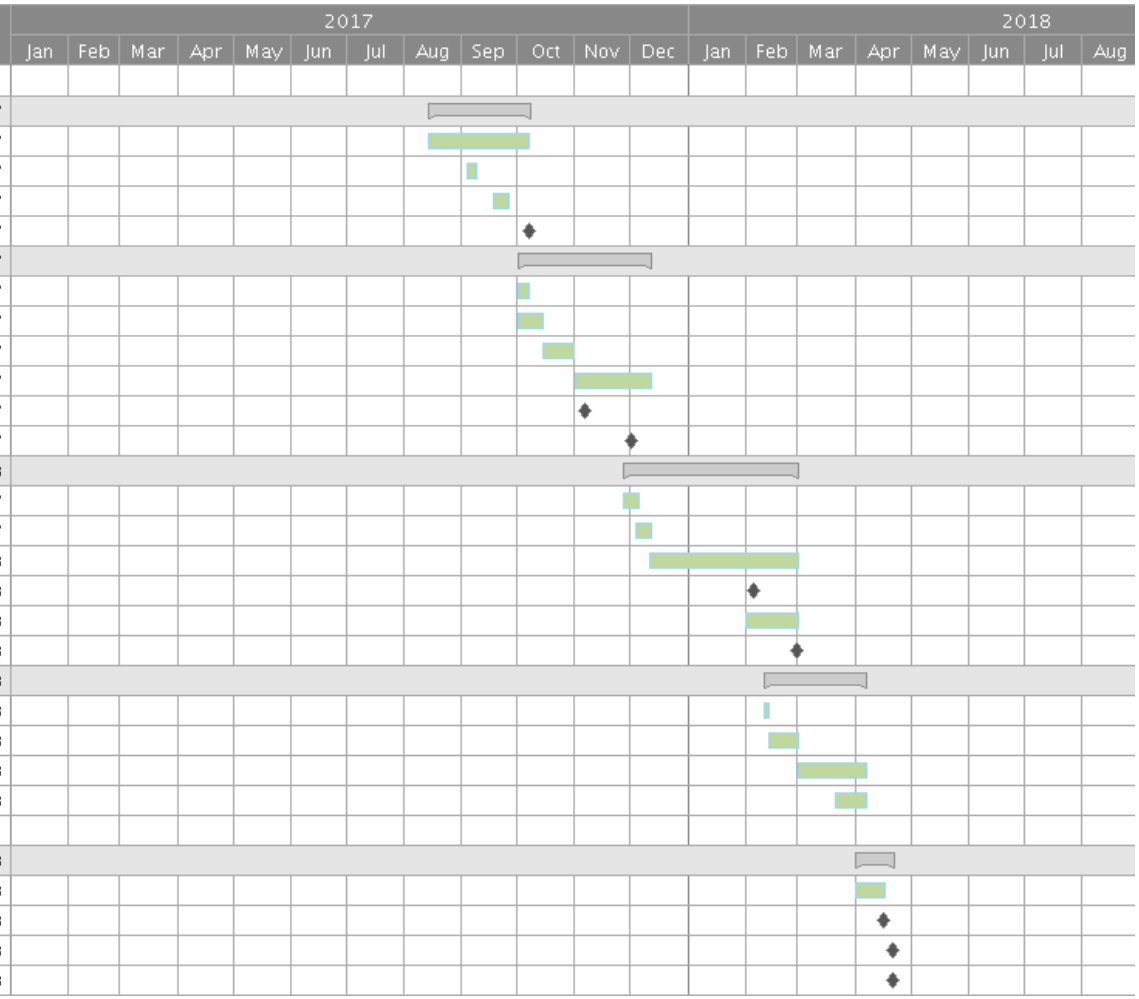
As mentioned in previous section, test-driven approach is adopted in this project. This is to minimize any complicated bug fixing before moving on next stage.

8.2 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.

9. Project Schedule

Task	Start	Due
Big Data Analysis on Ecological Data		
Planning	**08/14/17**	**10/07/17**
Background Research and Study	08/14/17	10/06/17
Define Project Scope and Objectives	09/04/17	09/08/17
Documentation of Project Plan	09/18/17	09/25/17
Phase Complete (Milestone)	10/07/17	10/07/17
Implementation (Big Data Analysis)	**10/01/17**	**12/11/17**
Software Installation and Configuration	10/01/17	10/06/17
Data Collection	10/01/17	10/13/17
Data Pre-Processing and Remove Noise	10/15/17	10/31/17
Data Analysis with Weka and IPython	11/01/17	12/11/17
Interim Report I (Milestone)	11/06/17	11/06/17
Phase Complete (Milestone)	12/01/17	12/01/17
Web Application Development	**11/27/17**	**02/28/18**
Design System Architecture	11/27/17	12/04/17
System Set Up and Configuration	12/04/17	12/11/17
Code Development	12/11/17	02/28/18
Interim Report II (Milestone)	02/05/18	02/05/18
Unit Testing and Debugging	02/01/18	02/28/18
Phase Complete (Milestone)	02/28/18	02/28/18
Testing	**02/11/18**	**04/06/18**
Develop Test Plan	02/11/18	02/12/18
Test Case Preparation	02/13/18	02/28/18
Test Case Execution	03/01/18	04/06/18
Documentation of Testing Result	03/21/18	04/06/18
Phase Complete (Milestone)		
Closing	**04/01/18**	**04/21/18**
Preparation of Final Report and Demonstration	04/01/18	04/16/18
Final Report Submission	04/16/18	04/16/18
Presentation	04/21/18	04/21/18
Phase Complete (Milestone)	04/21/18	04/21/18



10.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
1 Nov 2017 – Present	In progress: - Development, debugging and testing of python code

11.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