Scientific Computing, Modeling, and Simulation
Savitribai Phule Pune University

Master of Technology (M.Tech.)
Programme in Modeling and Simulation

Internship Project Report

Exploring Entomological Knowledge using GBIF datasets

Akshay Ghatage
CMS1905

Academic Year 2020-21

Scientific Computing, Modeling, and Simulation
Savitribai Phule Pune University

Certificate

This is certify that this report, titled

Exploring Entomological Knowledge using GBIF datasets,

authored by

Akshay Ghatage (CMS1905),

describes the project work carried out by the author under our supervision during the period from January 2021 to June 2021. This work represents the project component of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling, and Simulation, Savitribai Phule Pune University.

Bhalchandra Pujari, Asst. Professor
SCMS-SPPU, Pune, India

Mihir Arjunwadkar, Head
SCMS-SPPU, Pune, India

Scientific Computing, Modeling, and Simulation
Savitribai Phule Pune University

Author's Declaration

This document, titled

Exploring Entomological Knowledge using GBIF datasets,

authored by me, is an authentic report of the project work carried out by me as part of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling, and Simulation, Savitribai Phule Pune University. In writing this report, I have taken reasonable and adequate care to ensure that material borrowed from sources such as books, research papers, internet, etc., is acknowledged as per accepted academic norms and practices in this regard. I have read and understood the University's policy on plagiarism (`http://unipune.ac.in/administration_files/pdf/Plagiarism_Policy_University_14-5-12.pdf`).

Akshay Ghatage
CMS1905

Abstract

Madagascar is the earth's fourth largest island and can be seen as an intermediate between a continent and island. Madagascar is large enough to have a number of very distinct biomes. The large number of species at high risk of extinction that makes Madagascar one of the most important sites for biodiversity conservation worldwide[5]. Madagascar is one of the most important biodiversity hotspots in the world. Over last 20 years of research on insects in Madagascar, there is lot progress happened. In the project, we tried to supply some observations on the different orders of insects. In this contribution, GBIF dataset is used which belongs to Madagascar. Insects are part of a higher classification group of invertebrates. This GBIF dataset includes 2,31,604 records referred to the almost 24 insect orders(Coleoptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera, Neuroptera, Odonata, Orthoptera, Trichoptera, etc. This will help us to go through the Entomological Knowledge of Madagascar. When we go through the dataset, Hymenoptera, Diptera, Lepidoptera, Coleoptera were the most occurred orders of insect. On the basis of GBIF dataset, we followed the spatial distribution of these orders of insect[7]. We see that Eastern Madagascar is having most of the localities which are representing lot of orders of insect. This research is only to save these species and secure their localities. Because these invertebrates are very important for the human well-being. We need to maintain our ecosystem and this will help us. There are lot of protected areas in the Madagascar to protect these insect species. But some orders of insect are not in the protected areas. So, we need to include these areas into protected areas. This will help us to save the rare orders of insect. This whole project is based on the Spatial Data Analysis. This will give us lot of information about their localities. By using this information, we create sustainable environment for these orders of insect in their localities. Our focus is to conserve these species to conserve Biodiversity.

Acknowledgments

It gives me immense pleasure to have this opportunity to thank and acknowledge all those who made a difference in this endeavor. First and foremost, my sincere thanks to Bhalchand Pujari for being a wonderful guide. He made it possible for me to dive into Entomology for Spatial Data Analysis. He was ever so patient and encouraging in all our discussions. It has been a wonderful experience throughout – thank you so much!

I would also like to thank the initial designers of this wonderful and unique M.Tech. course in modeling and simulation. It is their vision and implementation which inspired me to take up this program.

Most importantly, I thank my family for it is through their love, support and encouragement I stand where I am today!

Contents

Chapter 1

Introduction

The study area considered in this analysis is the island of Madagascar. The whole study based on the spatial information of the localities of orders of insect. Biodiversity hotspots are defined as areas with exceptional species richness and concentrations of endemic species, and the loss of ¿70 per cent of the original primary vegetation. Madagascar is one of eight hottest biodiversity hotspots[4]. Madagascar has been designated one of the world's most important biodiversity hotspots, and it is characterised by disproportionally diverse flora and fauna. Getting a good knowledge of the megadiverse fauna of the Madagascar is still a far-reaching goal, especially for the invertebrates.

To increase the entomological knowledge of Madagascar, I went through GBIF dataset. This dataset provides the brief information of different orders of insect, their localities. According to the GBIF, there are 24 orders of insect. To work on the GBIF dataset, GBIF provides us their own libraries. GBIF created different libraries for different languages, for example, rgbif for R programming, pygbif for Python programming, gbifrb for Ruby programming. This project is based on Spatial Data Analysis. Working on this dataset gives me lot of geographical results. These geographical results are totally based on the different orders of insect and their localities. Analysis of large GBIF dataset referred to the 2,31,604 records of 24 orders of insects, to estimate the current coverage in the knowledge of the entomofauna in Madagascar, trying to evaluate:

1. Areas of highest number of records;
2. Correlation between records of insects and localities;
3. Which order of insect having highest records;
4. Which order of insect having lowest records;
5. Which areas we need to protect;
6. Which area having highest density of records;

[7]

For this Spatial Data Analysis, I plotted lot maps like heatmaps, marker cluster maps, density map, scatter maps, bar graph, etc. For this, I used Python programming. I used different libraries like pandas, numpy, matplotlib, geopandas, folium, plotly.

The result we got is fully based on the different geographical maps. My work is analyse the GBIF data and get geographical results which helps us to conserve the different orders of insect and their localities. These results will help us to gain the entomological knowledge of Madagascar.

Chapter 2

Spatial Data analysis

Spatial analysis is a part of GIS (Geographical Information System). The true power of GIS lies in the ability to perform analysis. Spatial analysis is a process in which you model problems geographically, derive results by computer processing, and then explore and examine those results. This type of analysis has proven to be highly effective for evaluating the geographic suitability of certain locations for specific purposes, estimating and predicting outcomes, interpreting and understanding change, detecting important patterns hidden in your information, and much more[3].

Spatial analysis in GIS involves three types of operations: Attribute Query also known as non-spatial (or spatial) query, Spatial Query and Generation of new data sets from the original database. The scope of spatial analysis ranges from a simple query about the spatial phenomenon to complicated combinations of attribute queries, spatial queries, and alterations of original data.Spatial analysis is a vital part of GIS and can be used for many applications like site suitability, natural resource monitoring, environmental disaster management and many more[8].

Spatial analysis allows you to solve complex location-oriented problems and better understand where and what is occurring in your world. Through spatial analysis you can interact with a GIS to answer questions, support decisions, and reveal patterns. Spatial analysis is in many ways the crux of a GIS, because it includes all of the transformations, manipulations, and methods that can be applied to geographic data to turn them into useful information. There are many ways of defining spatial analysis, but all in one way or another express the fundamental idea that information on locations is essential. Basically, think of spatial analysis as "a set of methods whose results change when the locations of the objects being analysed change"[2].

2.1 Libraries used in Python for Spatial Data Analysis

2.1.1 Folium

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. I used this library to plot the basemap, heatmaps and marker cluster maps of different orders of insect present in the dataset.

2.1.2 Plotly

Plotly's Python graphing library makes interactive, publication-quality graphs. The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart

types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases. I used this library to plot scatter mapbox and density map.

2.1.3 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

2.1.4 Geopandas

GeoPandas is a project to add support for geographic data to pandas objects. It currently implements GeoSeries and GeoDataFrame types which are subclasses of pandas.Series and pandas.DataFrame respectively. Geopandas objects can act on shapely geometry objects and perform geometric operations.

2.2 Geographical Maps

1. Base Map -
   I created base map of the Madagascar. This will help me to see the Madagascar on the world map.

2. Heat Map -
   Geographic Heat Map is an interactive visualization that displays your data points on a real map and signifies areas of low and high density.

3. Marker Cluster -
   The marker clustering utility helps you to manage multiple markers at different zoom levels. When a user views the map at a high zoom level, the individual markers show on the map. When the user zooms out, the markers gather together into clusters, to make viewing the map easier.

4. Density Map -
   Density mapping is simply a way to show where points or lines may be concentrated in a given area. Often, such maps utilize interpolation methods to estimate, across a given surface, where concentration of a given feature might be.

When we go through results, we will see these maps for different orders that help us to know about their spatial distribution.

Chapter 3

GBIF Dataset

[6]

GBIF stands for Global Biodiversity Information Facility. GBIF is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all type of life on Earth. It focuses on making scientific data on biodiversity available via the internet using web services. The data are provided by many institutions from around the world; GBIF's information architecture makes these data accessible and searchable through a single portal. Data available through the GBIF single portal are primarily distribution data on plants, animals, fungi, and microbes for the world, and scientific names data. The mission of the GBIF is to facilitate free and open access to biodiversity data worldwide to underpin sustainable development.

There are many online services that collect and maintain specimen records. However, Global Biodiversity 36 Information Facility is the largest collection of biodiversity 37 records globally, currently with 820 million records, roughly 5.9 million taxa, 36,000 datasets from 38 1,300 publishers. Many large biodiversity warehouses such as iNaturalist, VertNet, and USGS's Biodiversity Information Serving 40 Our Nation (BISON) all feed into GBIF. The most important organizational level in GBIF occurrence data is the occurrence record. The organization of GBIF matters because you can navigate GBIF data through these hierarchical 51 organizational levels - it helps to be familiar with the terminology and how each group relates to another[1].

GBIF data which I worked is based on the occurrences of insects of Madagascar and it includes all the coordinates of different localities in Madagascar. When I went through the data, I found that there are 24 orders of insects present in the Madagascar. These 24 orders of insects are Lepidoptera(15080), Coleoptera(11177), Odonata(1184), Hemiptera(5237), Hymenoptera(170473), Orthoptera(826), Diptera(19572), Neuroptera(67), Cnemidolestodea(1), Psocodea(191), Blattodea(179), Thysanoptera(45), Dermaptera(29), Mantodea(202), Phasmida(40), Trichoptera(1597), Ephemeroptera(4316), Plecoptera(127), Strepsiptera(2), Siphonaptera(54), Embioptera(2), Archaeognatha(2), Mecoptera(1), Protorthoptera(1). The most occurred orders are Hymenoptera, Diptera, Lepidoptera, Coleoptera.

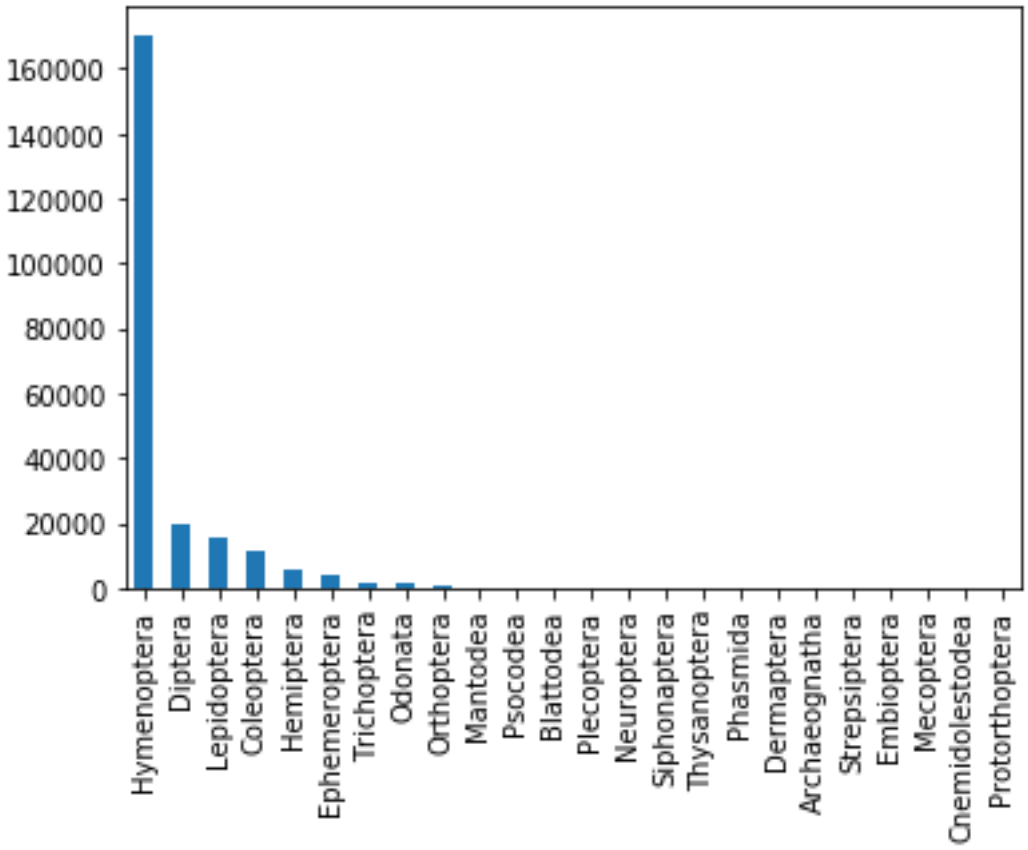

Figure 3.1: Bar plot of insect orders

The above figure shows that the occurrences of different orders of insect. The total records of all orders of insects are 2,31,604. When we go through the dataset we get this count of all orders.

Figure 3.2 is showing the map of all orders present in the dataset. This will tell us why entomological knowledge of Madagascar is very important to protect these orders.

All maps that we plotted are based on these orders of insect.

Figure 3.2 is having 2,31,604 records. But these records are occurred in lot of same localities. So, the total localities we found for these records are 7383.

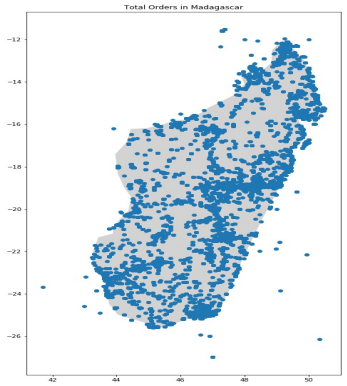

Figure 3.2: Map of all orders of insect in Madagascar

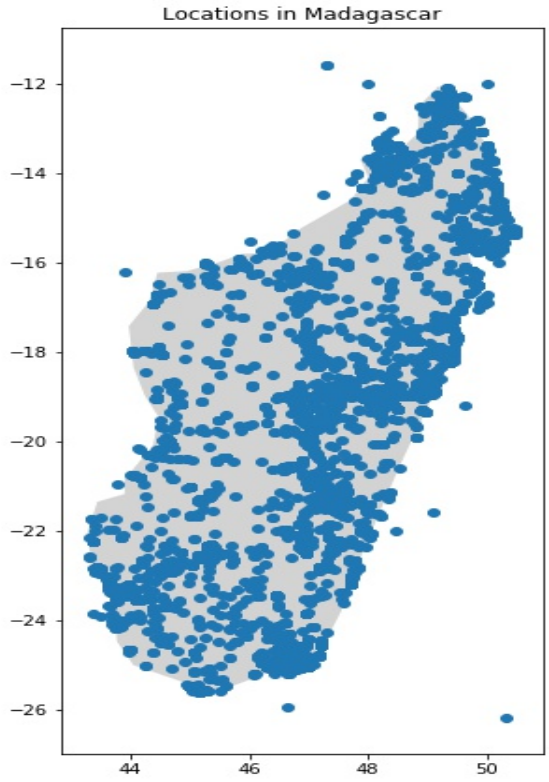

Figure 3.3: Locations in Madagascar

Figure 3.3 will describe these localities in Madagascar. During the data processing, we got to know that there are 7383 localities present in the Madagascar which are having these 2,31,604 records.

3.1 GBIF Libraries

Although we discuss libraries for R, Python, and Ruby here, we focus mostly on the R library rgbif as it has seen the most developer and user attention, and is the most mature [1].

3.1.1 rgbif

Herein, we describe the rgbif software packag for working with GBIF data in the R programming environment. R is a widely used language in academia, as well as non-profit and private sectors. Importantly, R makes it easy to execute all steps of the research process, including data management, data manipulation and cleaning, statistics, and visualization. Thus, an R client for getting GBIF data is a powerful tool to facilitate reproducible research.

The rgbif package is nearly completely written in R (a small Javascript library is included for reading well known text), uses an MIT license to maximize use everywhere. rgbif is developed publicly on GitHub, where development versions of the package can be installed, and bugs and feature requests reported. Stable versions of rgbif can be installed from CRAN, the distribution

network for R packages. rgbif is part of the rOpenSci project, a developer network making R software to facilitate reproducible research[1].

3.1.2 pygbif

pygbif (Chamberlain) is a Python library for working with GBIF data in the Python programming environment. Python is a general purpose programming language used widely in all sectors, and for all parts of software development including server and client side use cases. Python is used exclusively in some scientific disciplines (e.g., astronomy), and has partial usage in other disciplines. A Python client for GBIF data is an important tool given the even wider usage of Python than R, though maybe slightly less than R for ecology/biology disciplines.

The pygbif library is less mature and complete than the R package. It also uses an MIT license to maximize use everywhere. pygbif is developed publicly on GitHub, where development versions of the package can be installed, and bugs and feature requests reported. Stable versions of pygbif can be installed from pypi, the distribution network for Python libraries[1].

3.1.3 gbifrb

gbifrb (Chamberlain) is a library for working with GBIF data in the Ruby programming environment. Like Python, Ruby is a general purpose programming language used widely in all sectors. Unlike Python, Ruby is not used extensively in scientific disciplines. However, a Ruby client for GBIF data can be an important tool given how widely Ruby is used for web and web service development[1].

My whole work is in python. So, I went through the pygbif. There are lot of functions to understand. But the problem is that GBIF imposes for any given search a limit of 200,000 records in the search service, after which point you can't download any more records for that search. However, you can download more records for different searches. My search for the GBIF dataset is having more than 200,000 records. So, I dropped the idea to use pygbif library.

Chapter 4

Results

1. **Base Map**

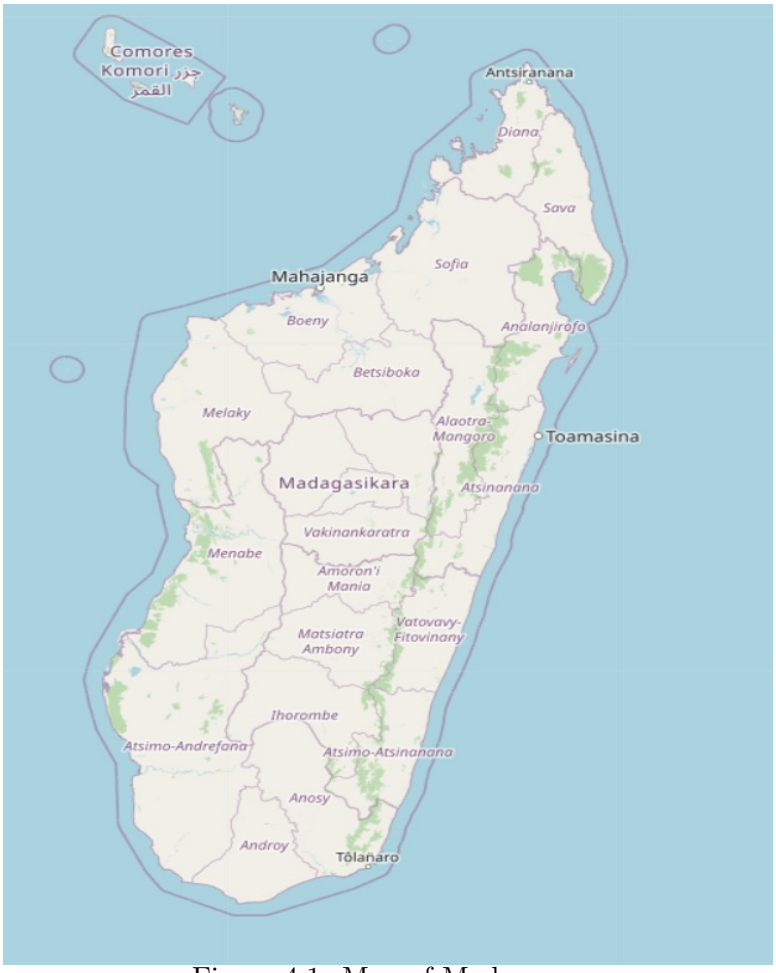

Figure 4.1: Map of Madagascar

This is the base map of the Madagascar. We are using this base map to plot heatmaps and marker clusters of different orders of insects.

2. **Hymenoptera**

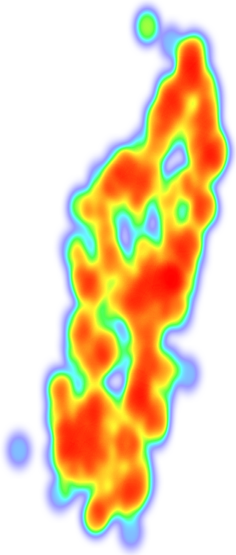

Figure 4.2: Heatmap of the order Hymenoptera

This is the heat map of the Hymnoptera order of insects. Here we will see that records of this order are almost present in the all localities in the Madagascar. Therefore this is the order which is having most of the records.

Figure 4.3: Marker clusters of Hymenoptera

Above map shows that the large amount of Hymenoptera records are on the Eastern Madagascar.

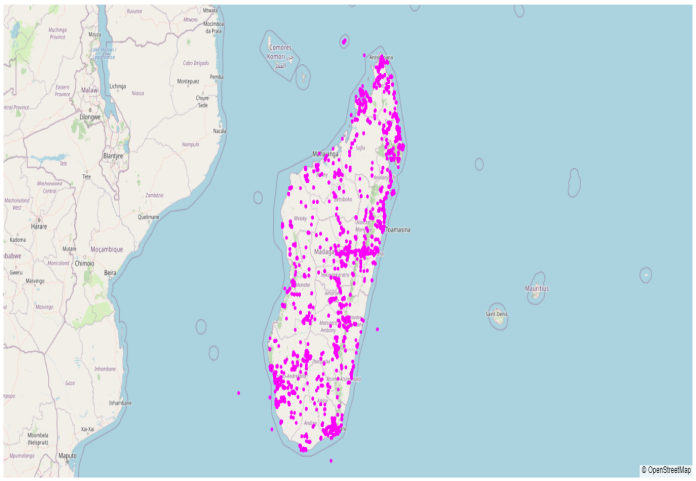

Figure 4.4: Scatter Map of Hymenoptera

Above map shows that total records of Hymenoptera (170473) how distributed in Madagascar.

3. **Diptera**

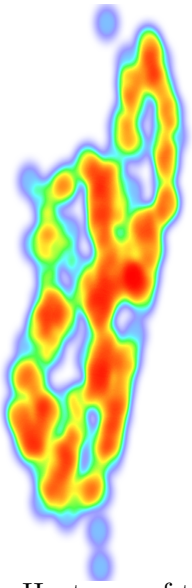

Figure 4.5: Heatmap of the order Diptera

In this heatmap, we will see that small areas of Madagascar are not having the records of order Diptera.

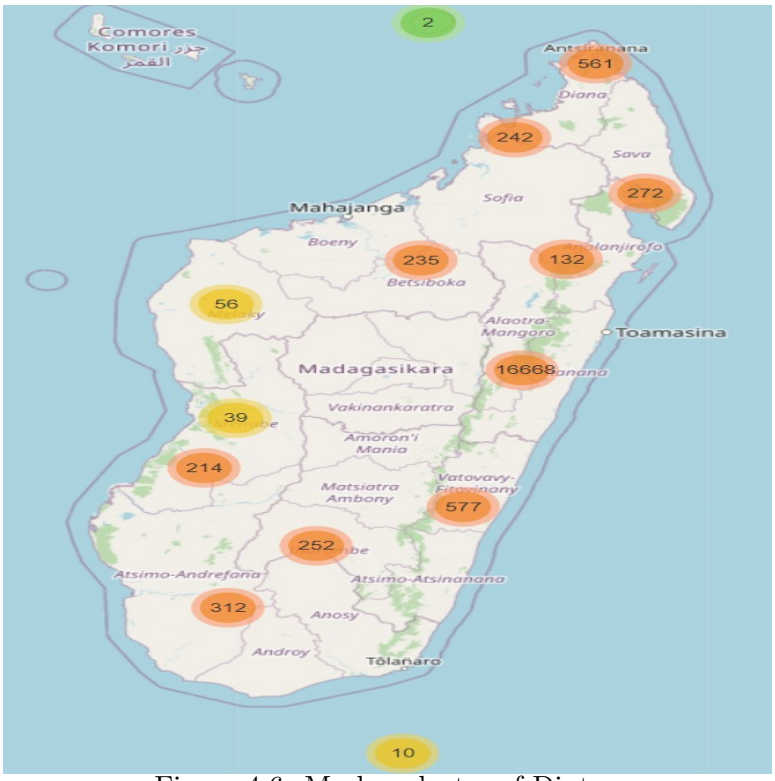

Figure 4.6: Marker cluster of Diptera

Here also we see that Eastern Madagascar is having large number of records of Diptera. Diptera is the second largest order of records. This map will help us to know that which areas of Madagascar having maximum records.

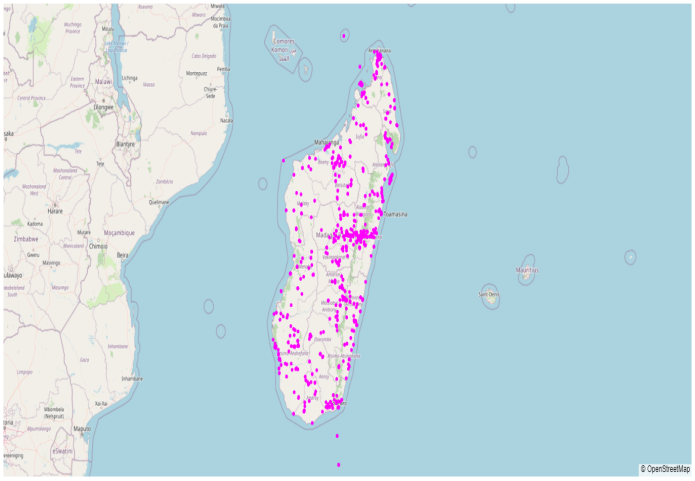

Figure 4.7: Scatter Map of Diptera

This map will show us the distribution of total records of Diptera (19527) in different localities of Madagascar.

4. **Lepidoptera**

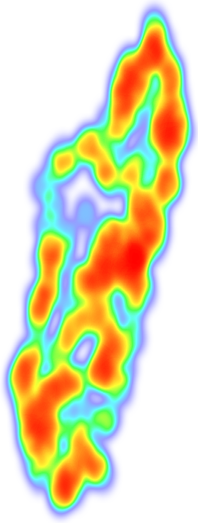

Figure 4.8: Heatmap of the order Lepidoptera

Above map will help us to find out large density areas of Lepidoptera. This is plotted by using coordinates of all Lepidoptera records.

Figure 4.9: Marker cluster of Lepidoptera

When we go through this map, we will see that large number of records of Lepidoptera are representing the Eastern Madagascar localities.

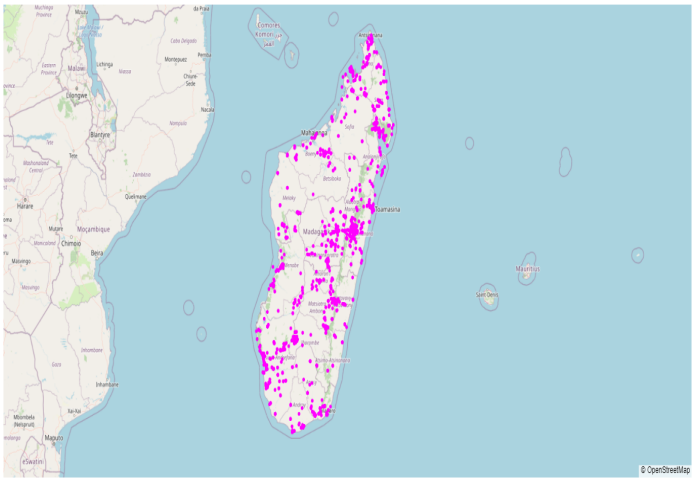

Figure 4.10: Scatter Map of Lepidoptera

Above map will show us the distribution of total records of Lepidoptera (15080) in different localities in Madagascar.

5. Coleoptera

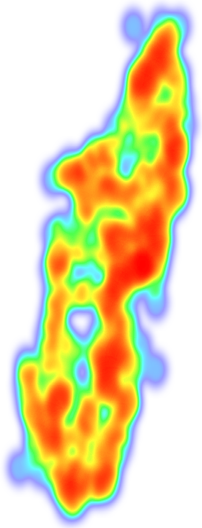

Figure 4.11: Heatmap of the order Coleoptera

In above map, we will see that where the large amount of records take place in Madagascar map.

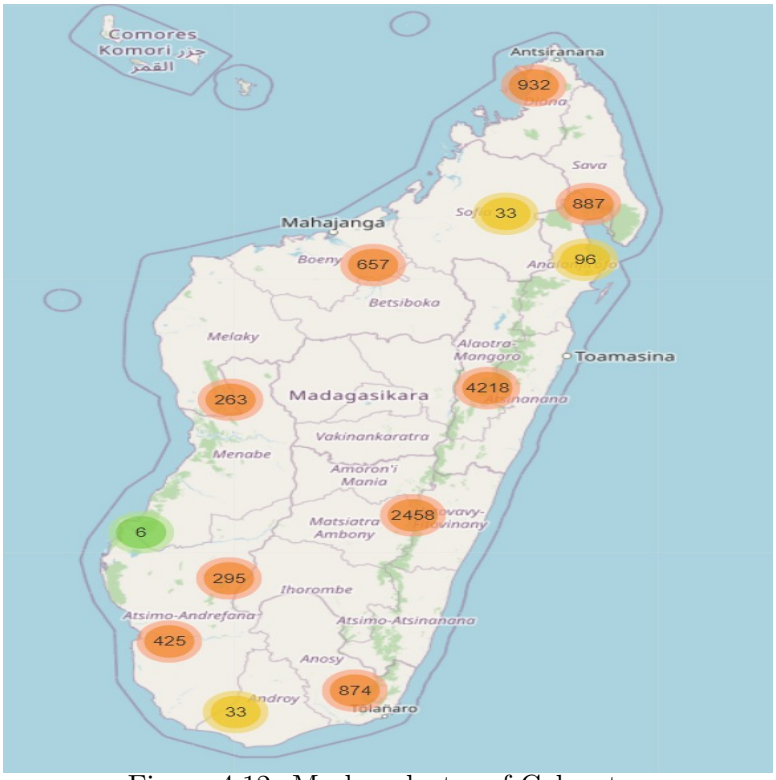

Figure 4.12: Marker cluster of Coleoptera

In above map, we will see that areas in Eastern Madagascar are having large number of records of Coleoptera.

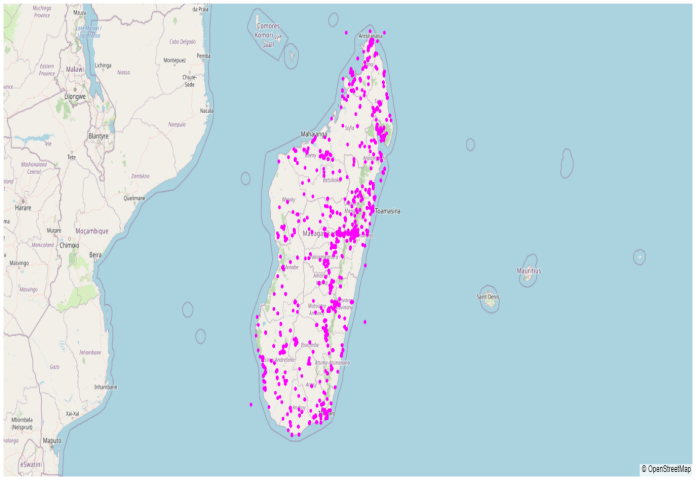

Figure 4.13: Scatter Map of Coleoptera

Above map will show us the distribution of total records of Coleoptera (11177) in Madagascar.

6. **Hemiptera (total records - 5237)**

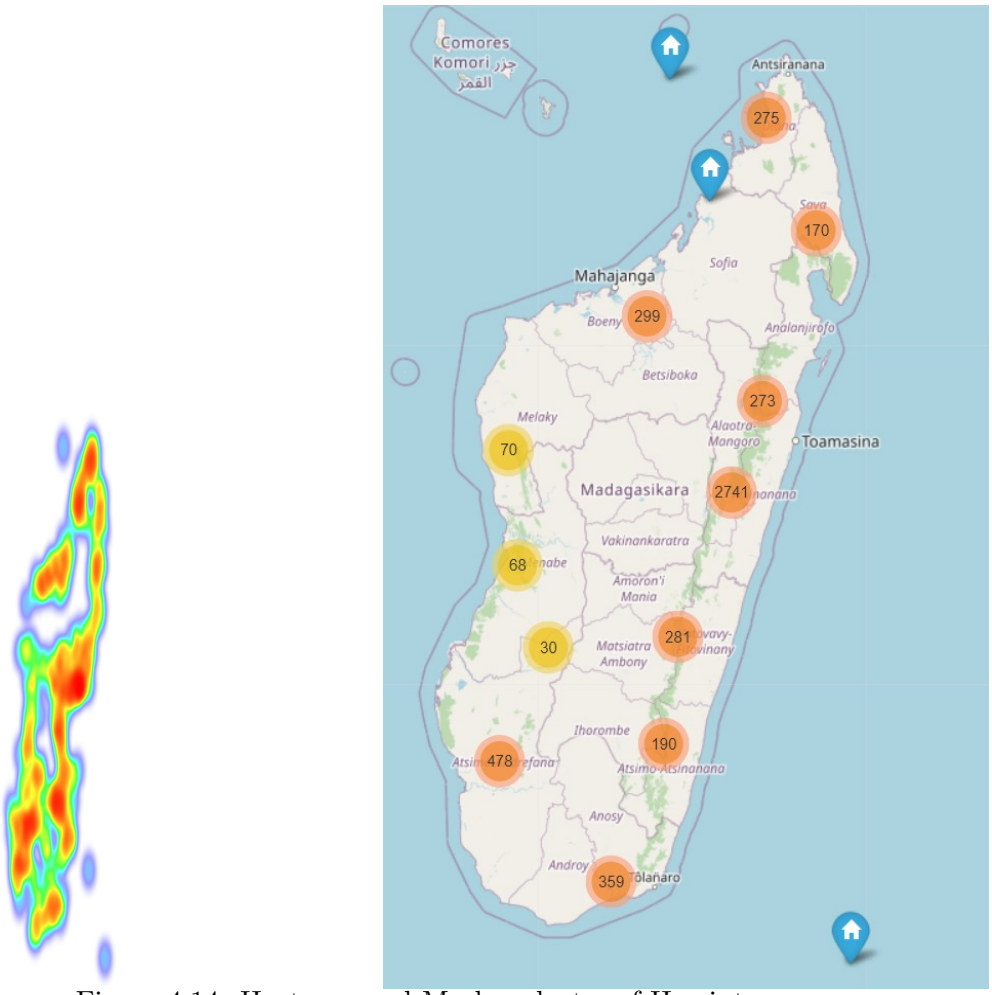

Figure 4.14: Heatmap and Marker cluster of Hemiptera

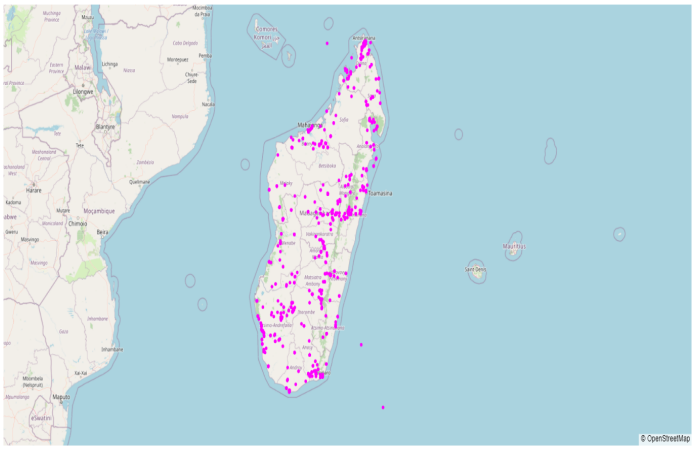

Figure 4.15: Scatter Map of Hemiptera

7. Ephemeroptera (total records - 4316)

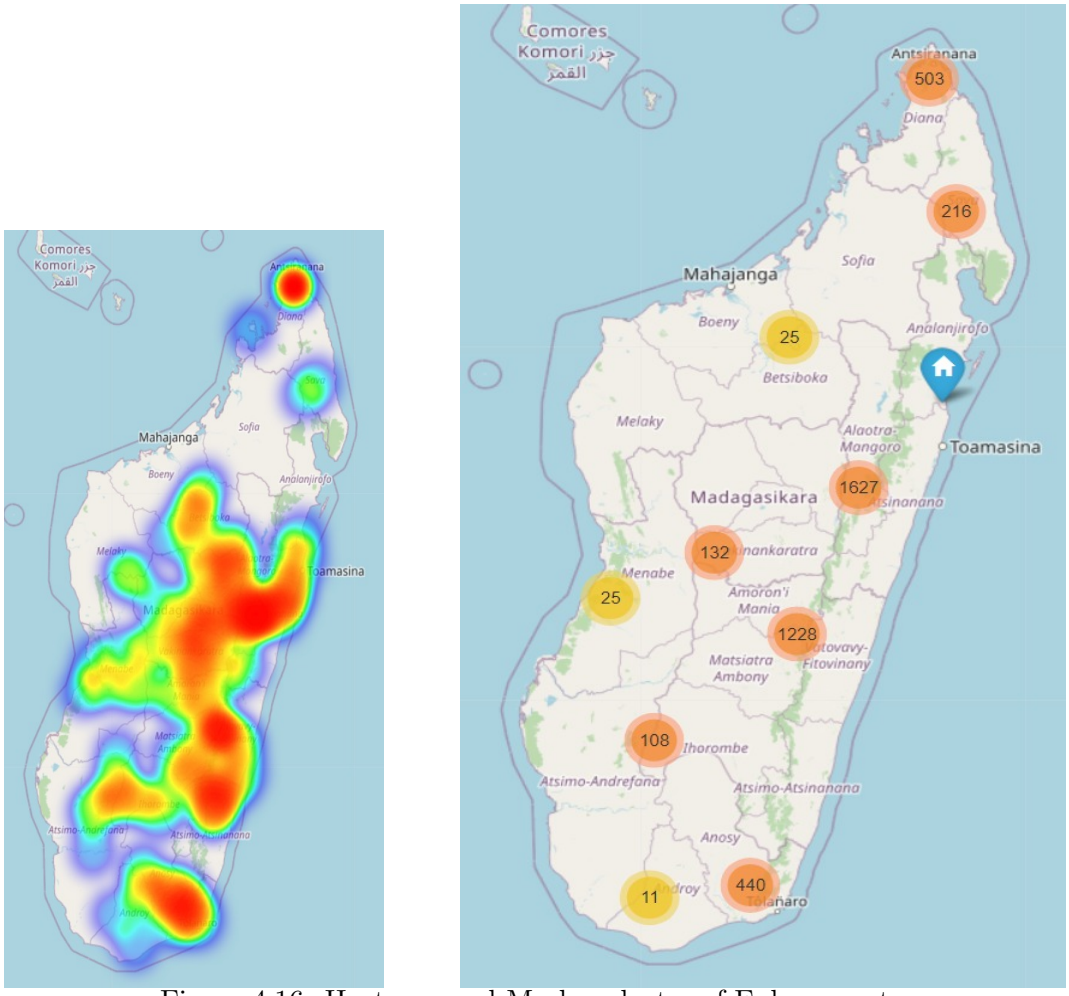

Figure 4.16: Heatmap and Marker cluster of Ephemeroptera

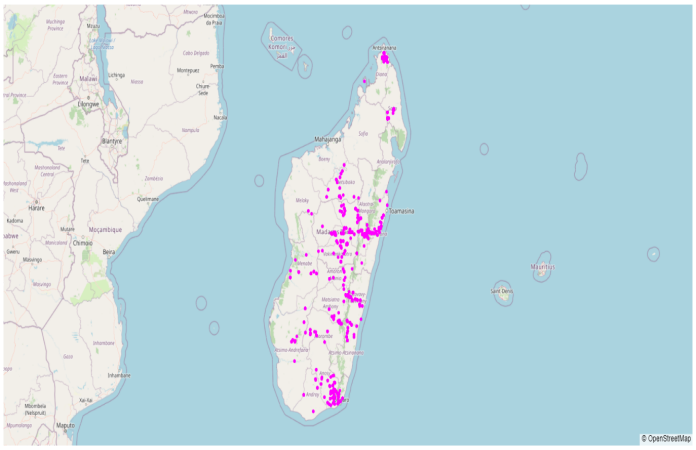

Figure 4.17: Scatter Map of Ephemeroptera

8. Trichoptera (total records - 1597)

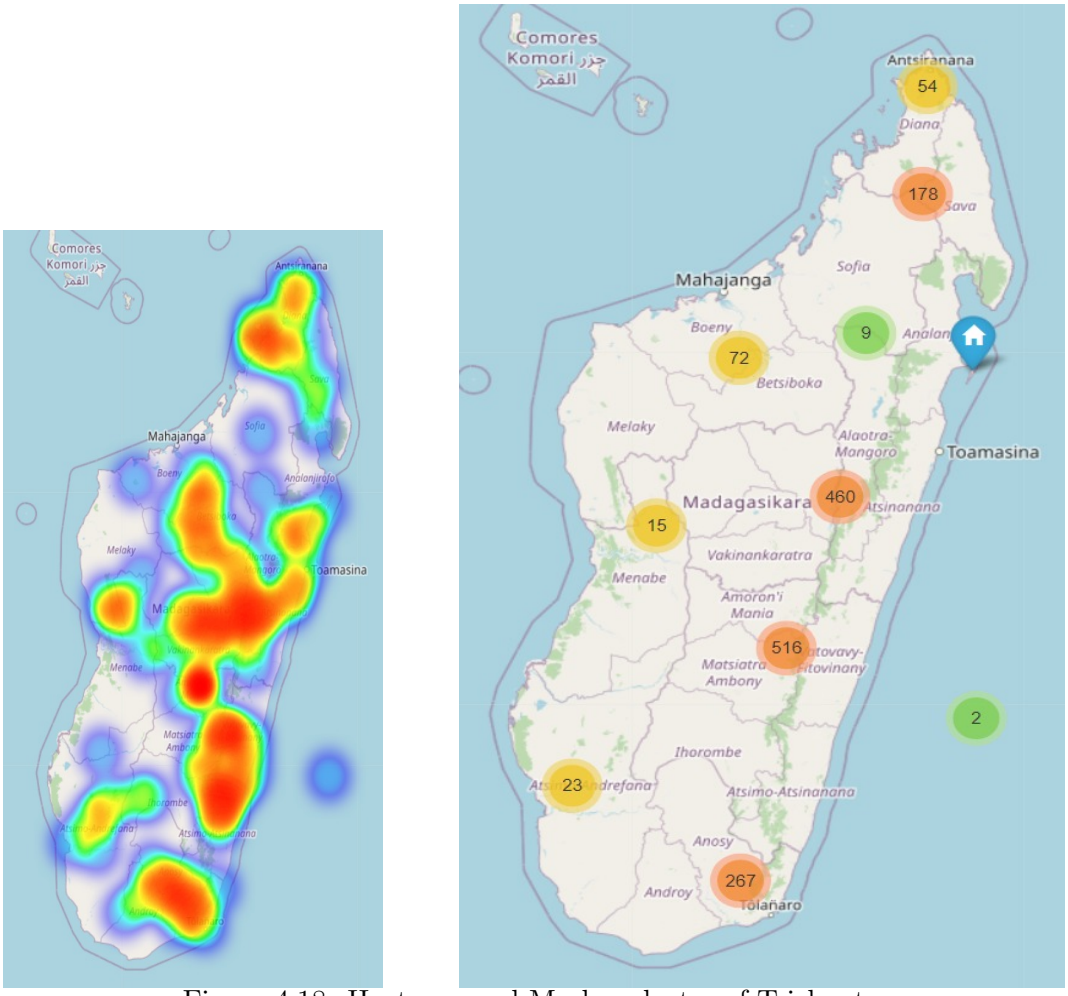

Figure 4.18: Heatmap and Marker cluster of Trichoptera

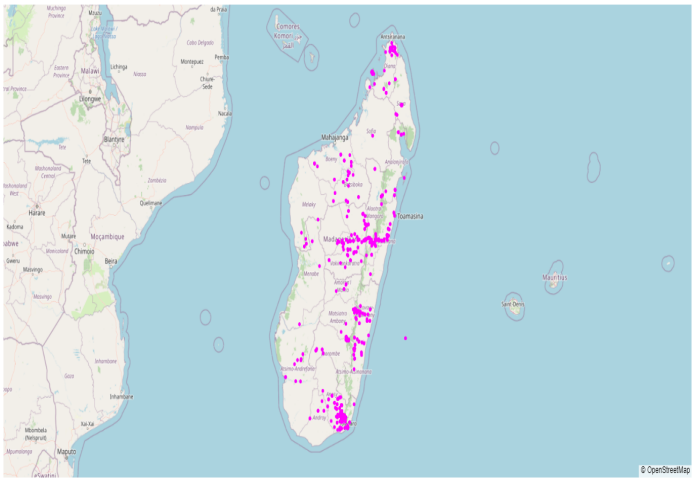

Figure 4.19: Scatter Map of Trichoptera

9. Odonata (total records - 1184)

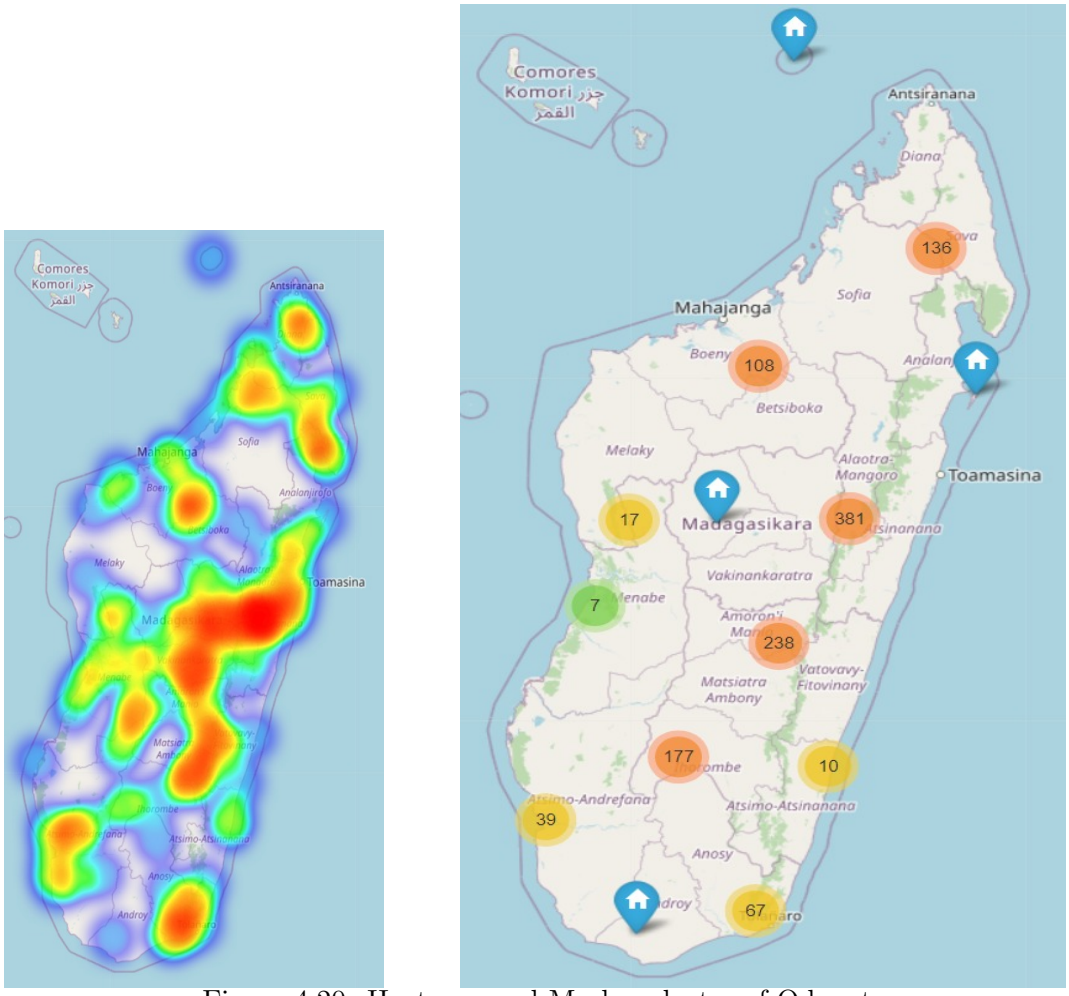

Figure 4.20: Heatmap and Marker cluster of Odonata

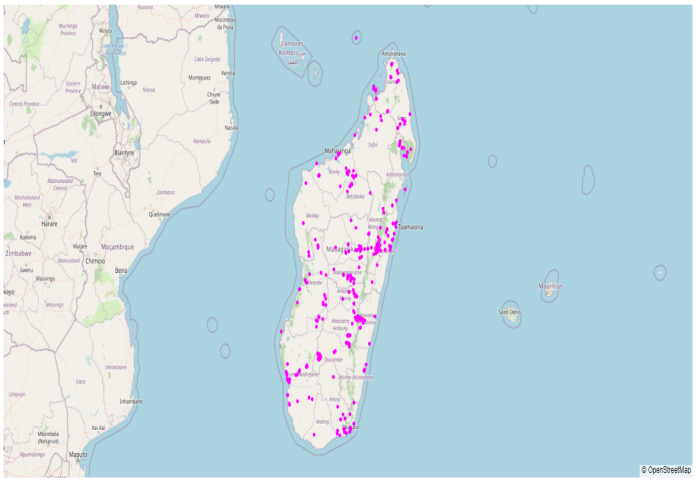

Figure 4.21: Scatter Map of Odonata

10. Orthoptera (total records - 826)

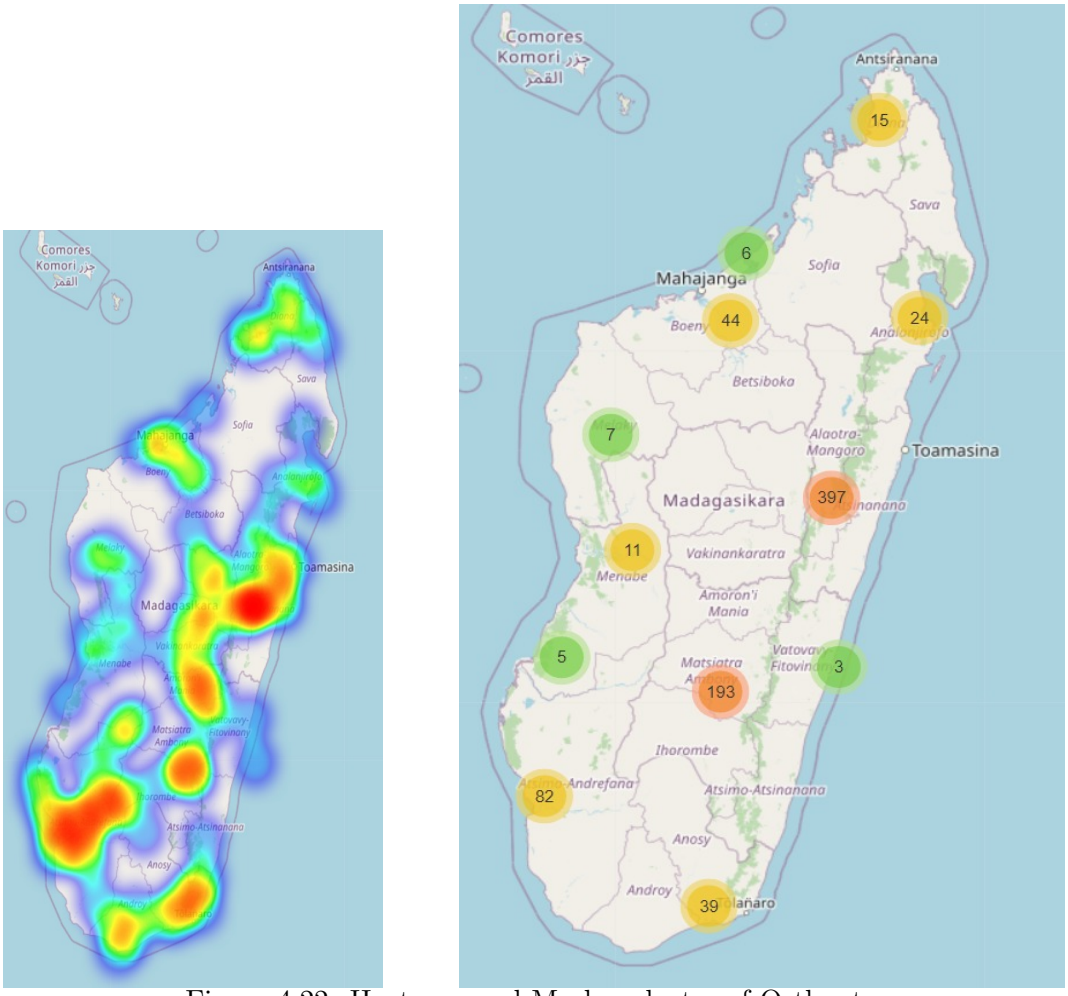

Figure 4.22: Heatmap and Marker cluster of Orthoptera

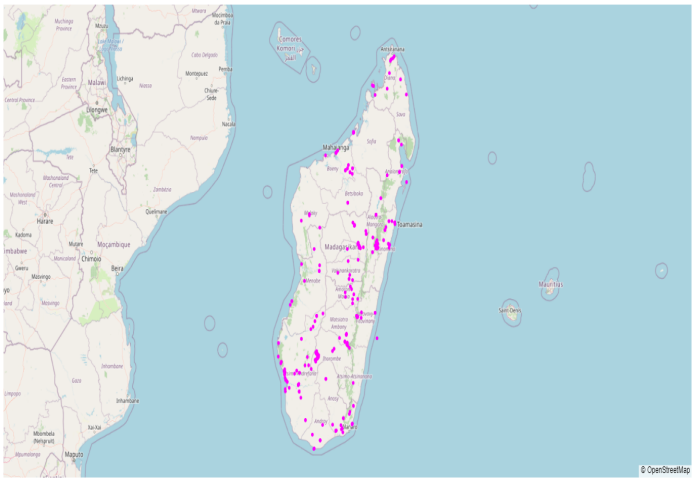

Figure 4.23: Scatter Map of Orthoptera

11. Neuroptera (total records - 67)

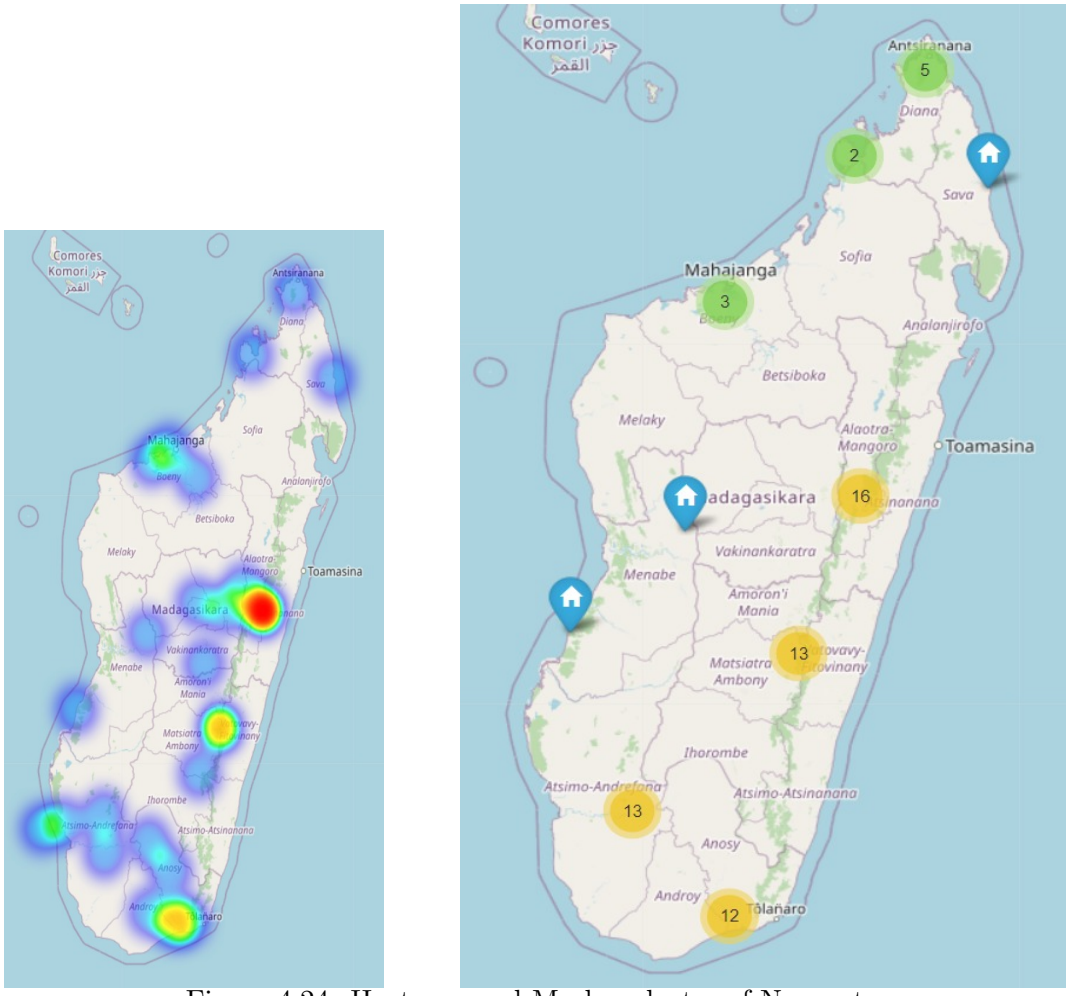

Figure 4.24: Heatmap and Marker cluster of Neuroptera

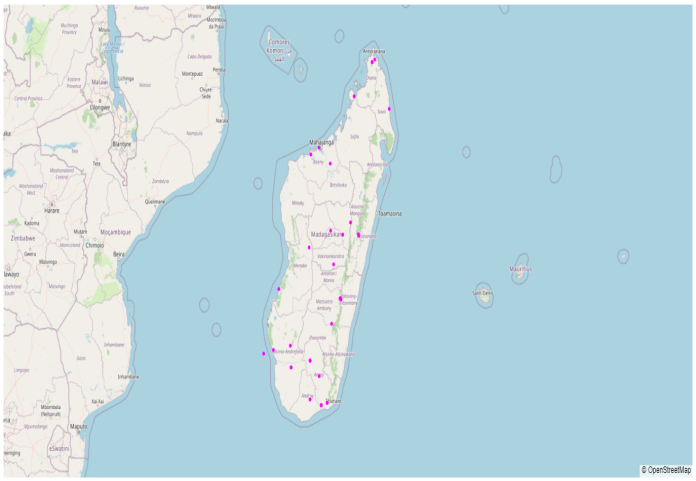

Figure 4.25: Scatter Map of Neuroptera

12. Mantodea (total records - 202)

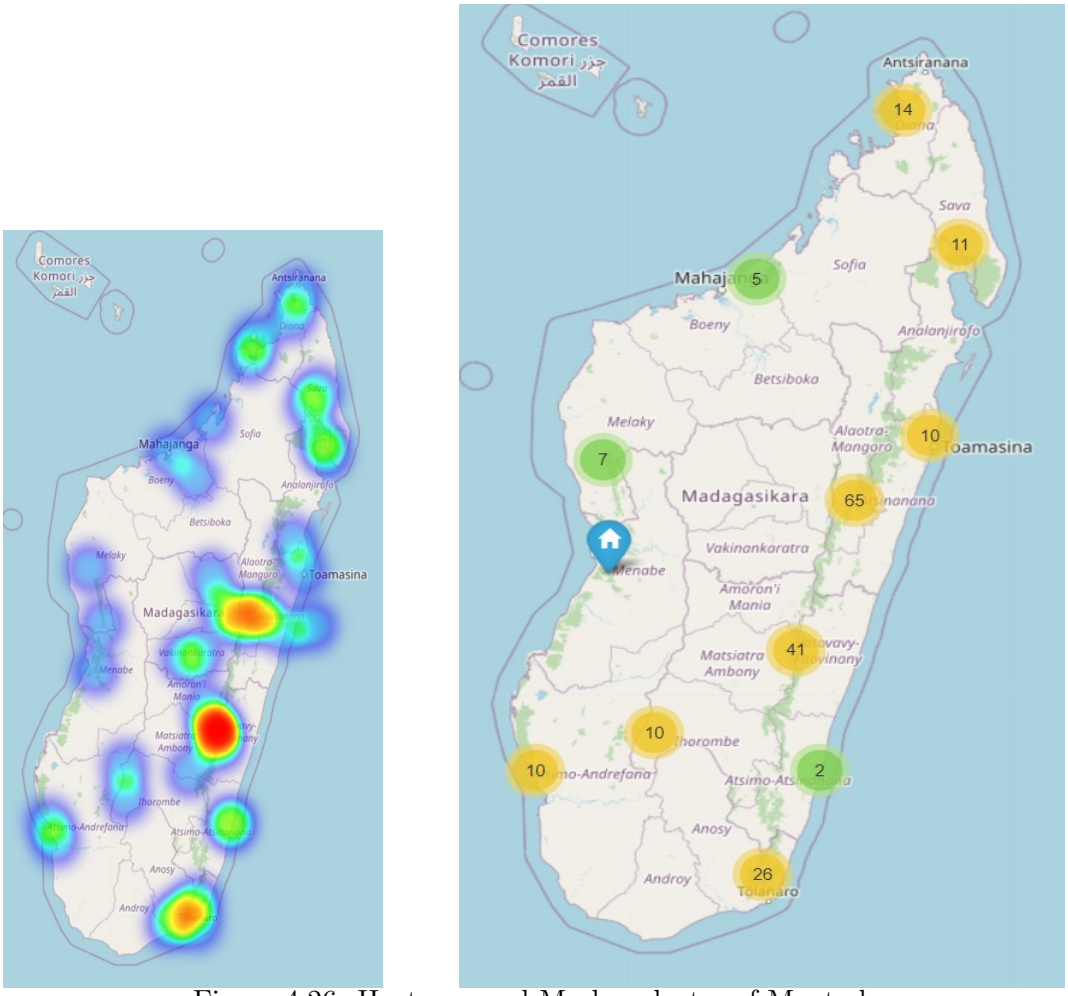

Figure 4.26: Heatmap and Marker cluster of Mantodea

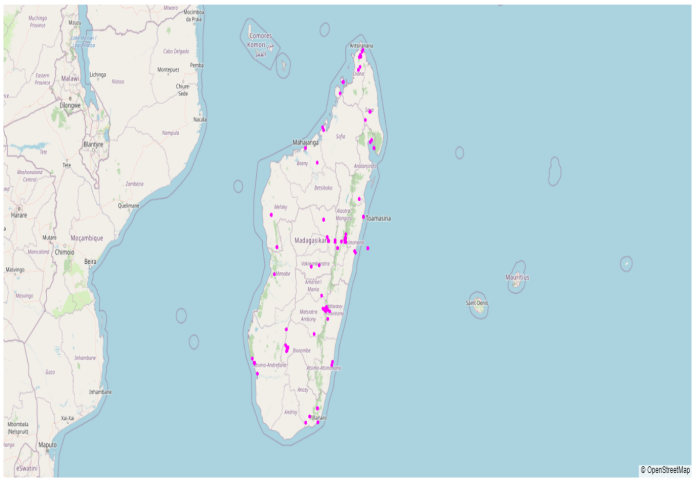

Figure 4.27: Scatter Map of Mantodea

13. Psocodea (total records - 191)

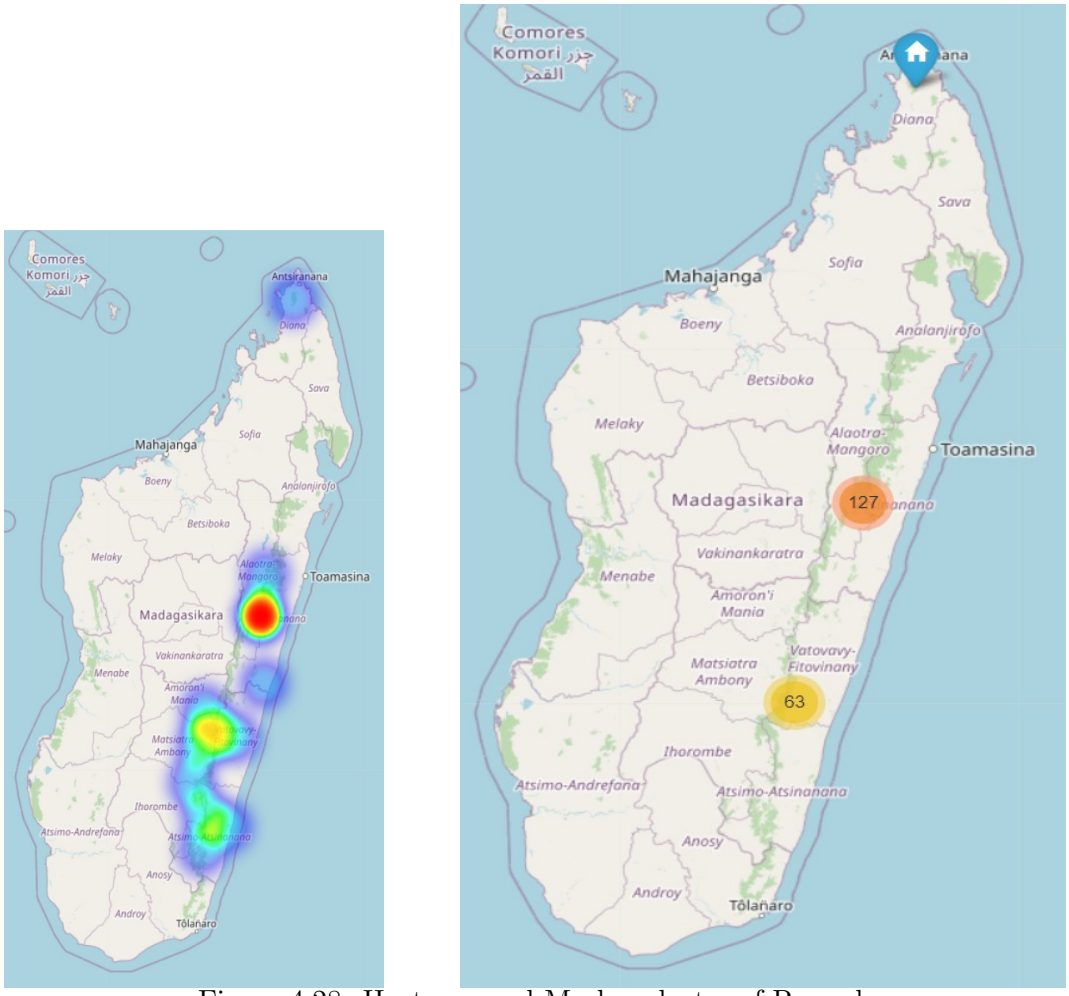

Figure 4.28: Heatmap and Marker cluster of Psocodea

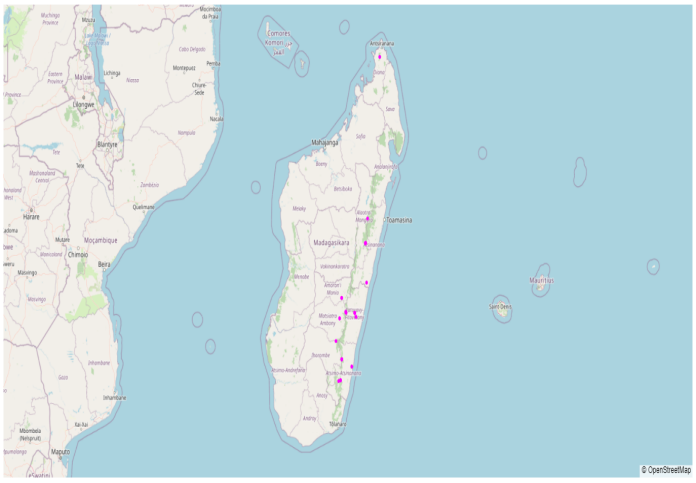

Figure 4.29: Scatter Map of Psocodea

14. Blattodea (total orders - 179)

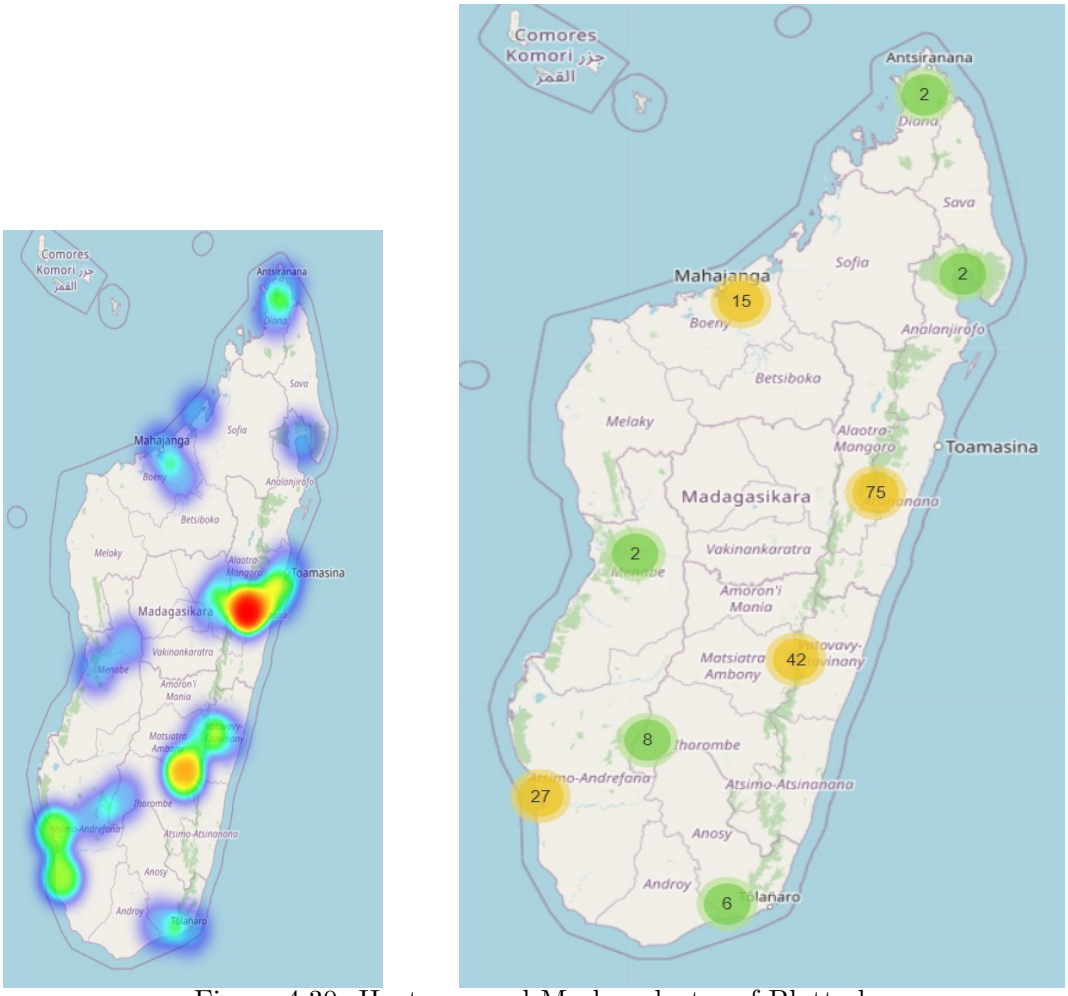

Figure 4.30: Heatmap and Marker cluster of Blattodea

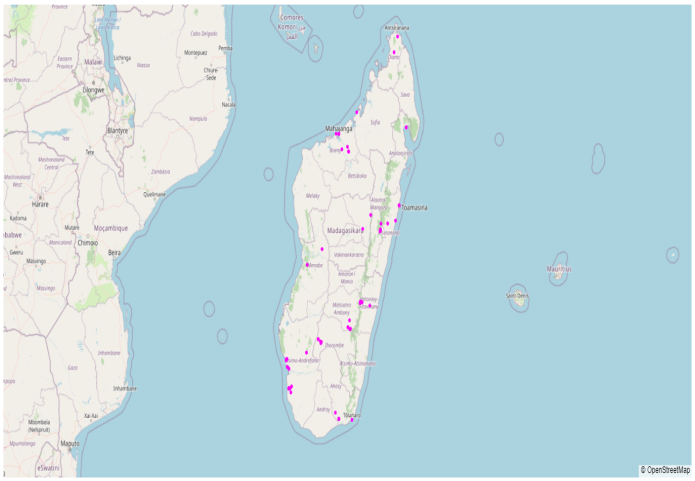

Figure 4.31: Scatter Map of Blattodea

15. Plecoptera (total orders - 127)

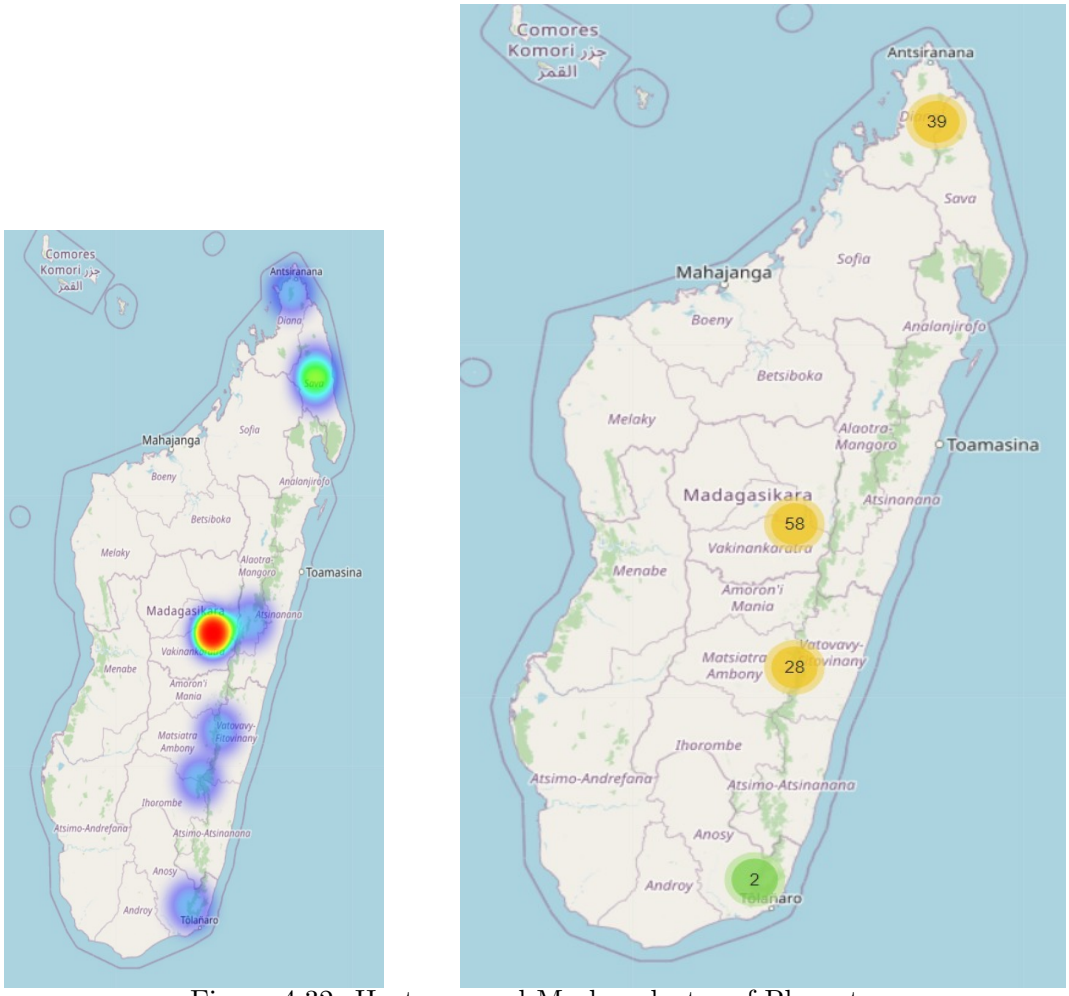

Figure 4.32: Heatmap and Marker cluster of Plecoptera

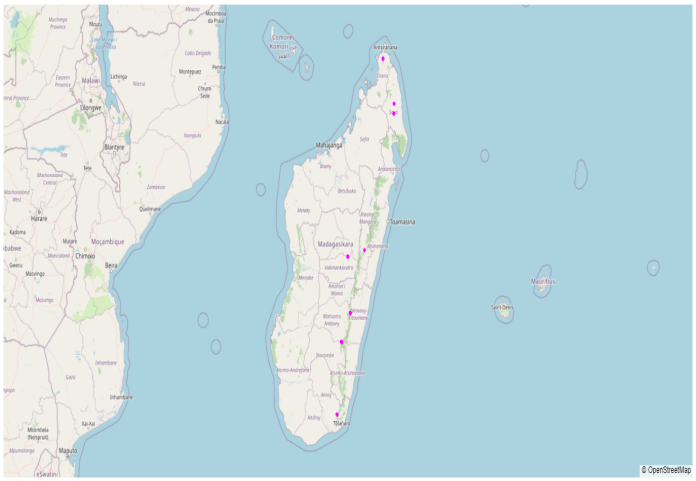

Figure 4.33: Scatter Map of Plecoptera

16. Siphonaptera (total records - 54)

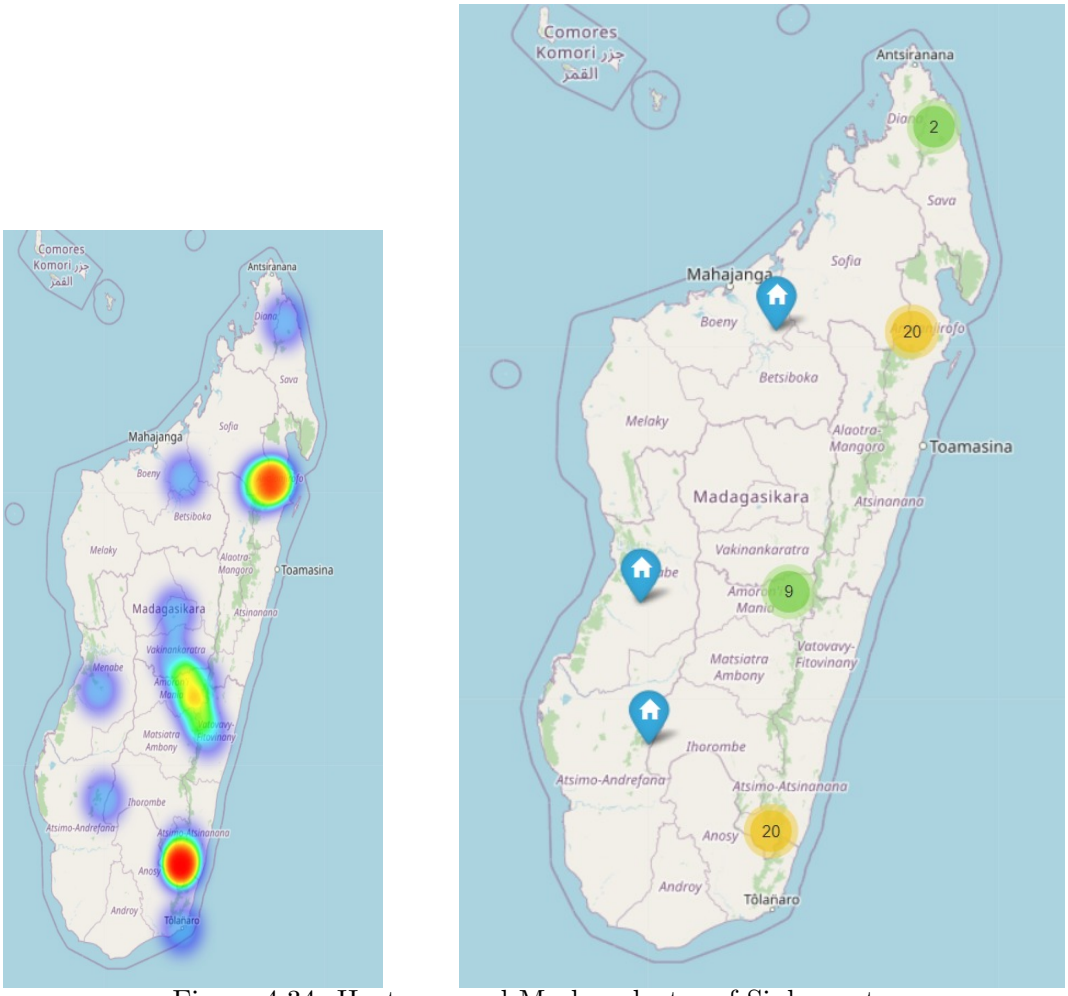

Figure 4.34: Heatmap and Marker cluster of Siphonaptera

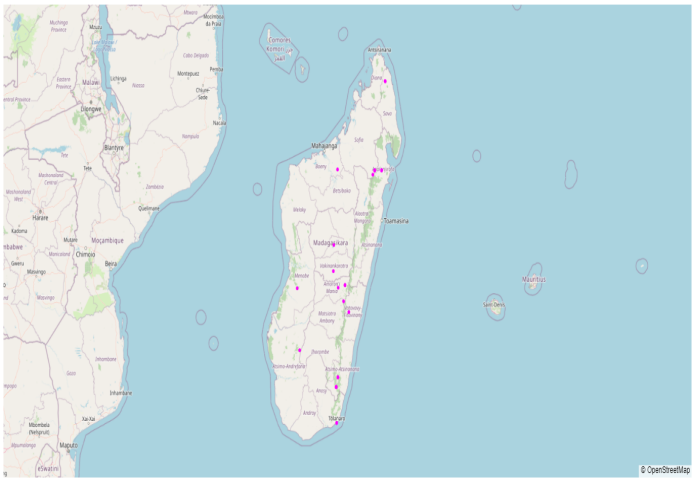

Figure 4.35: Scatter Map of Siphonaptera

17. Thysanoptera (total records - 45)

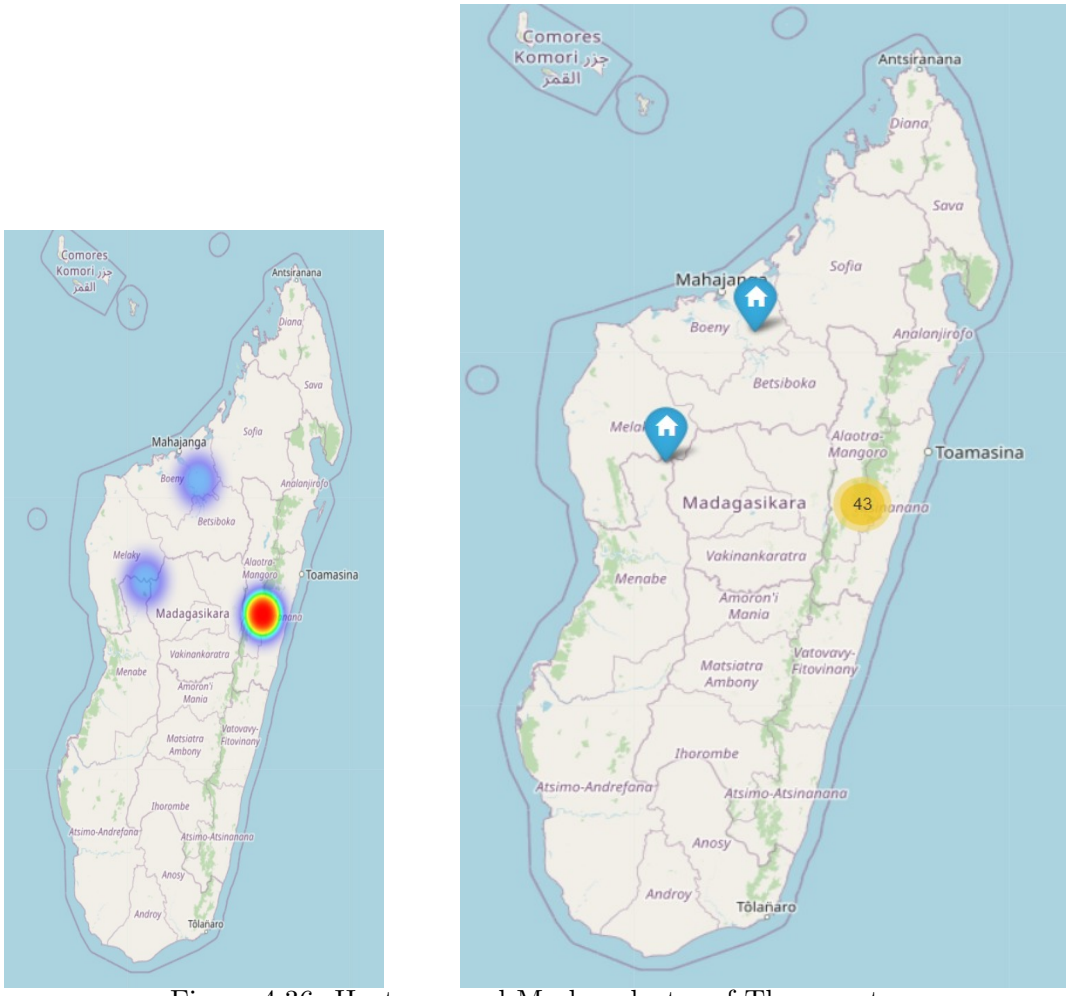

Figure 4.36: Heatmap and Marker cluster of Thysanoptera

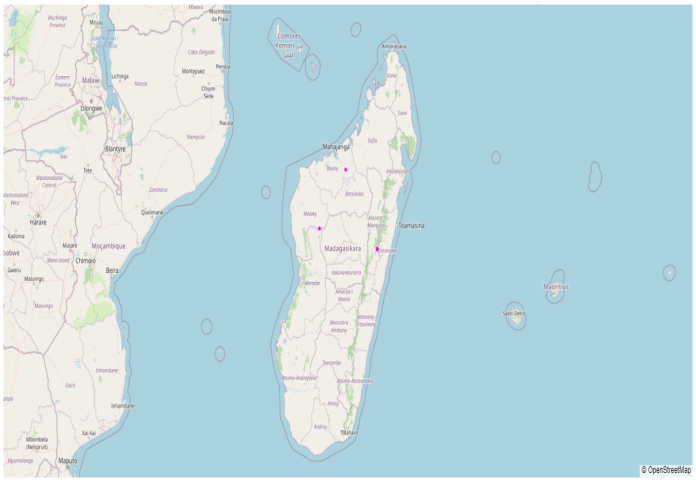

Figure 4.37: Scatter Map of Thysanoptera

18. Phasmida (total records - 40)

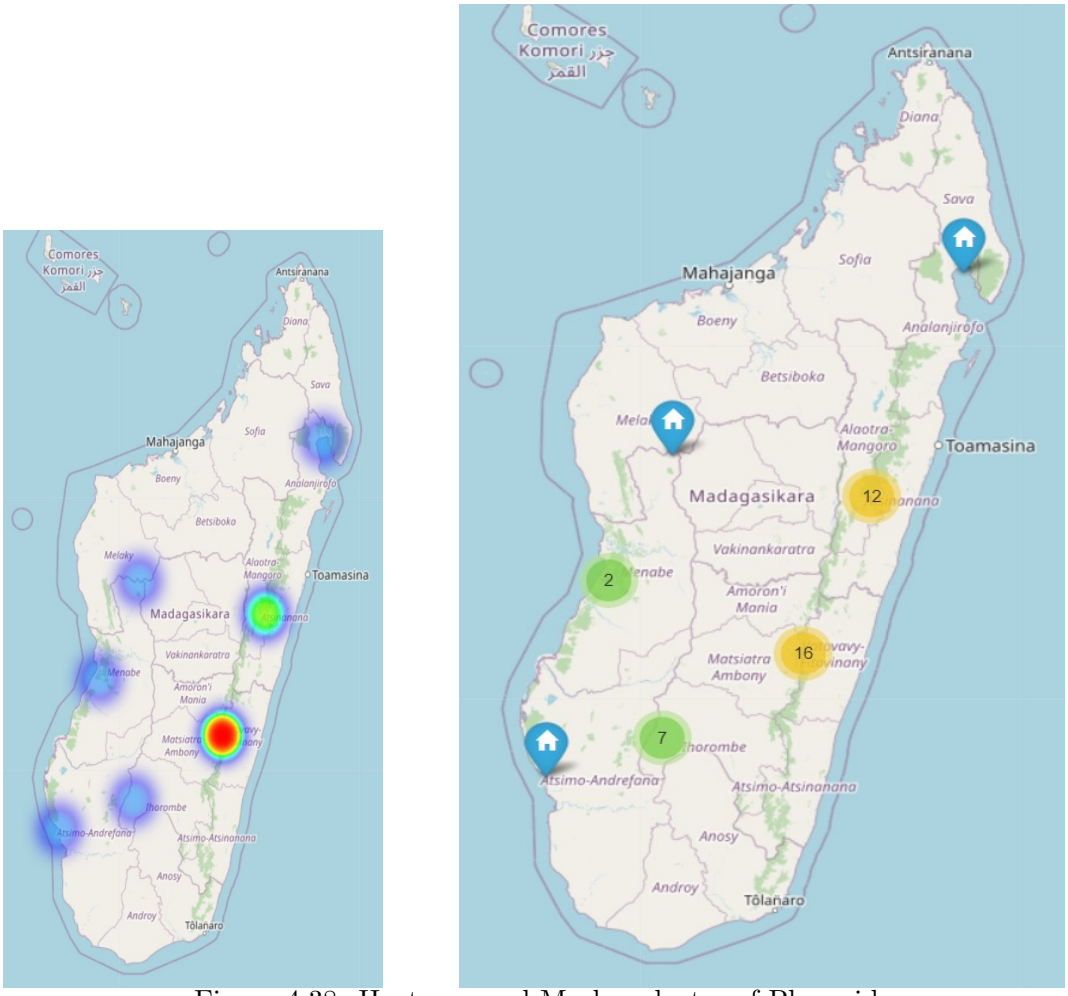

Figure 4.38: Heatmap and Marker cluster of Phasmida

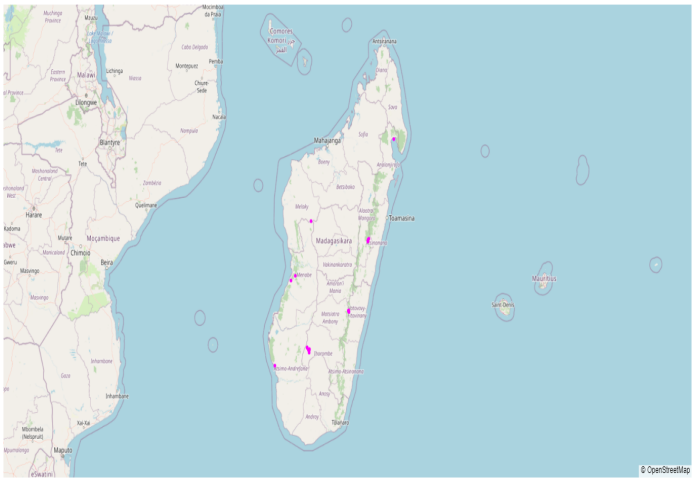

Figure 4.39: Scatter Map of Phasmida

19. Dermaptera (total records - 29)

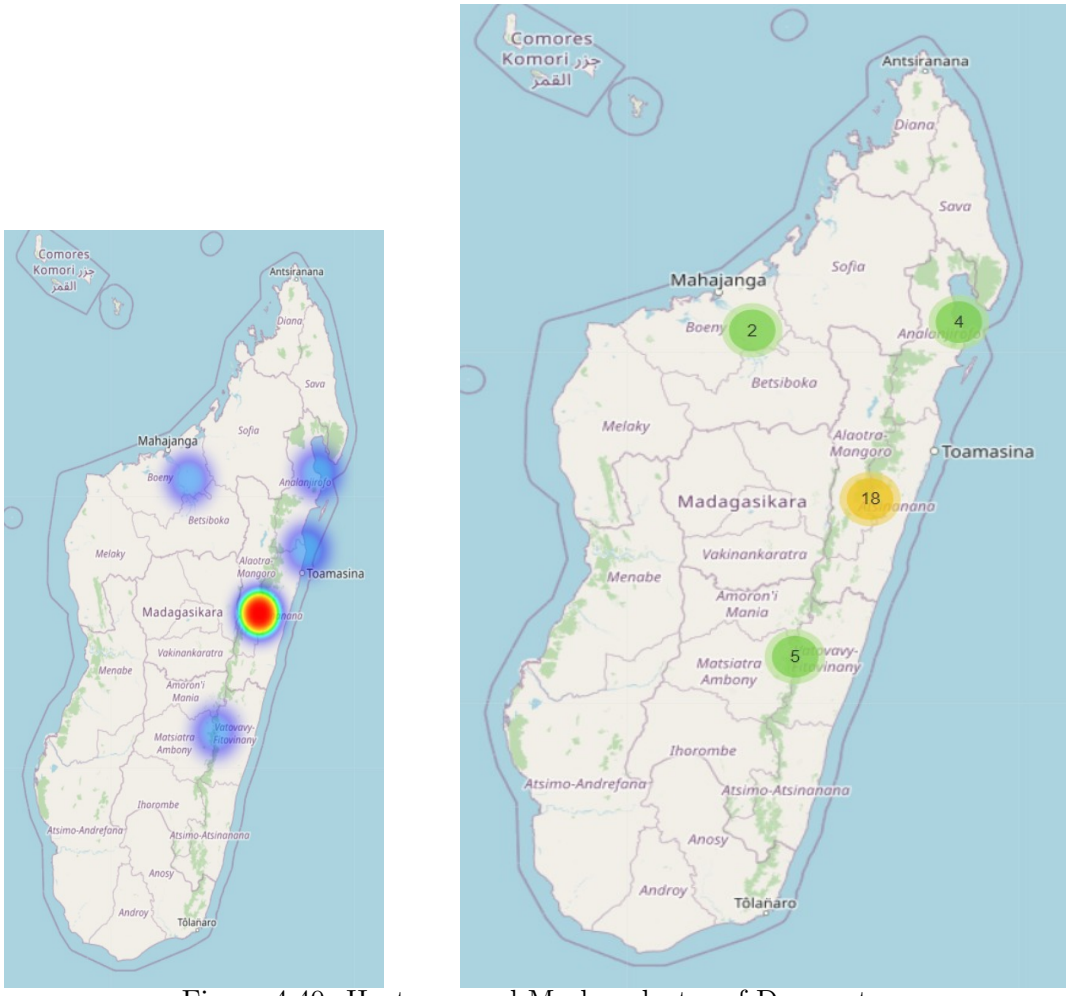

Figure 4.40: Heatmap and Marker cluster of Dermaptera

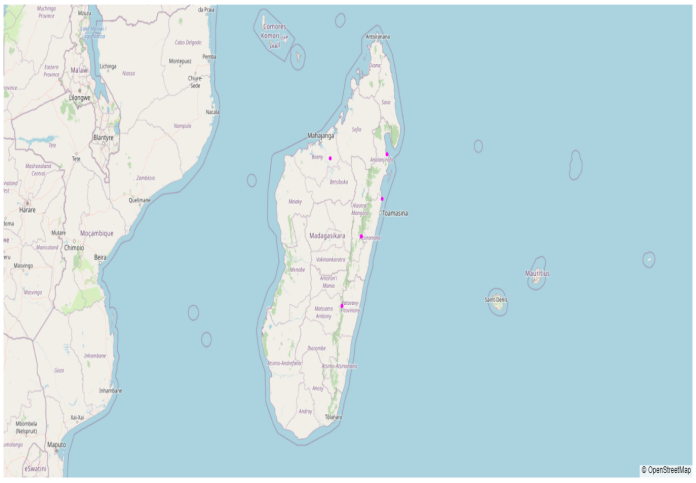

Figure 4.41: Scatter Map of Dermaptera

20. Embioptera(2), Archaeognatha(2), Strepsiptera(2), Cnemidolestodea(1), Mecoptera(1), Protorthoptera(1) (total records - 9)(in bracket their individual records)

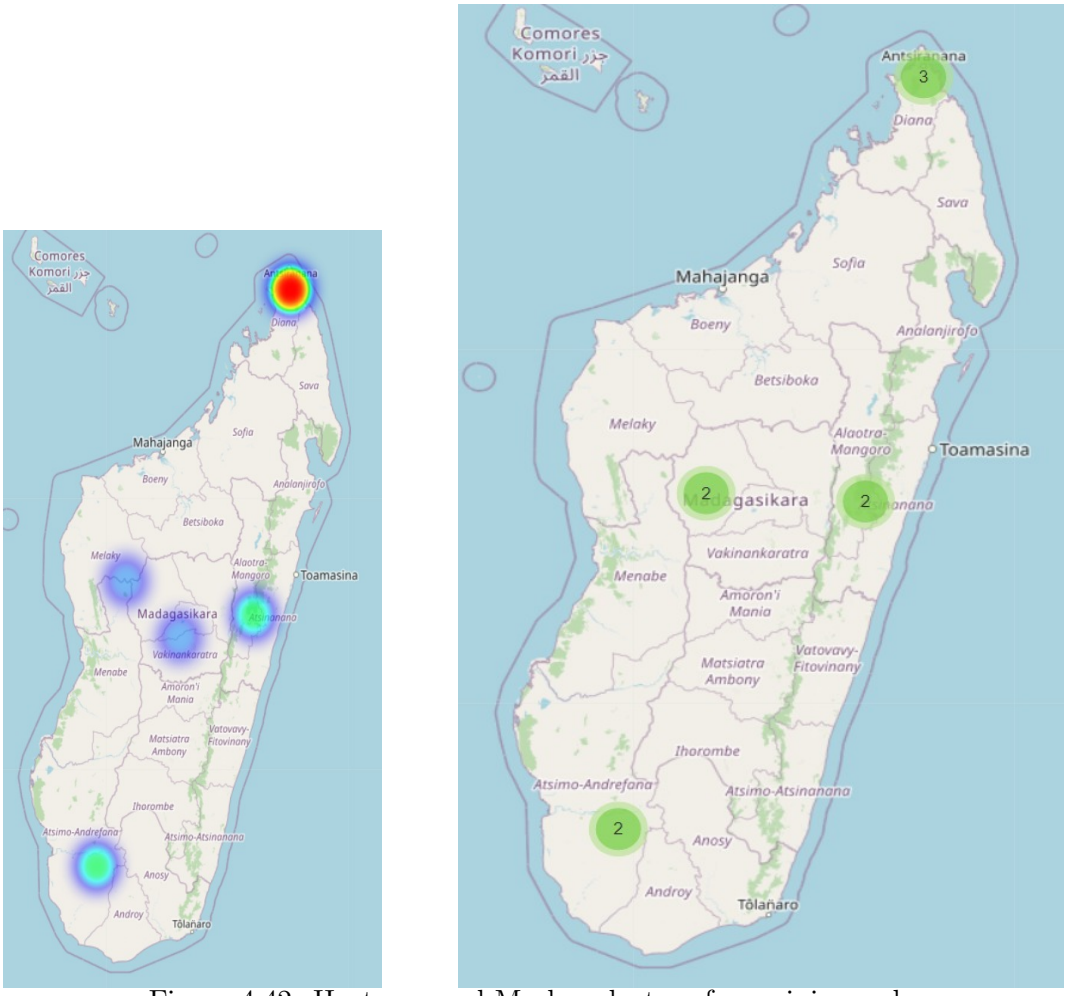

Figure 4.42: Heatmap and Marker cluster of remaining orders

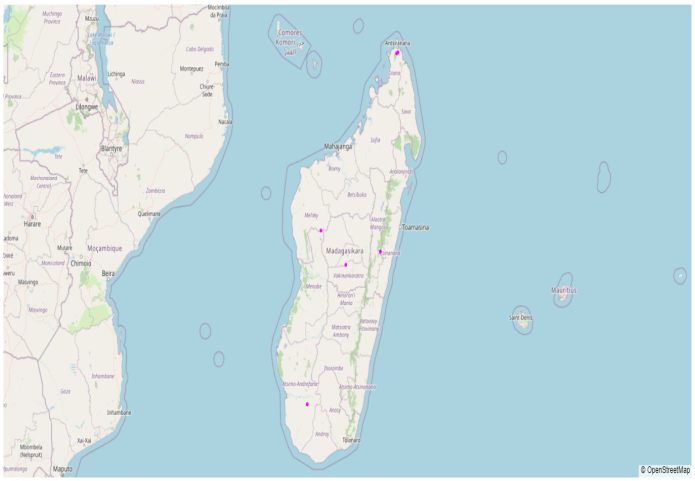

Figure 4.43: Scatter Map of remaining orders

21. Some maps related to the total records of all orders

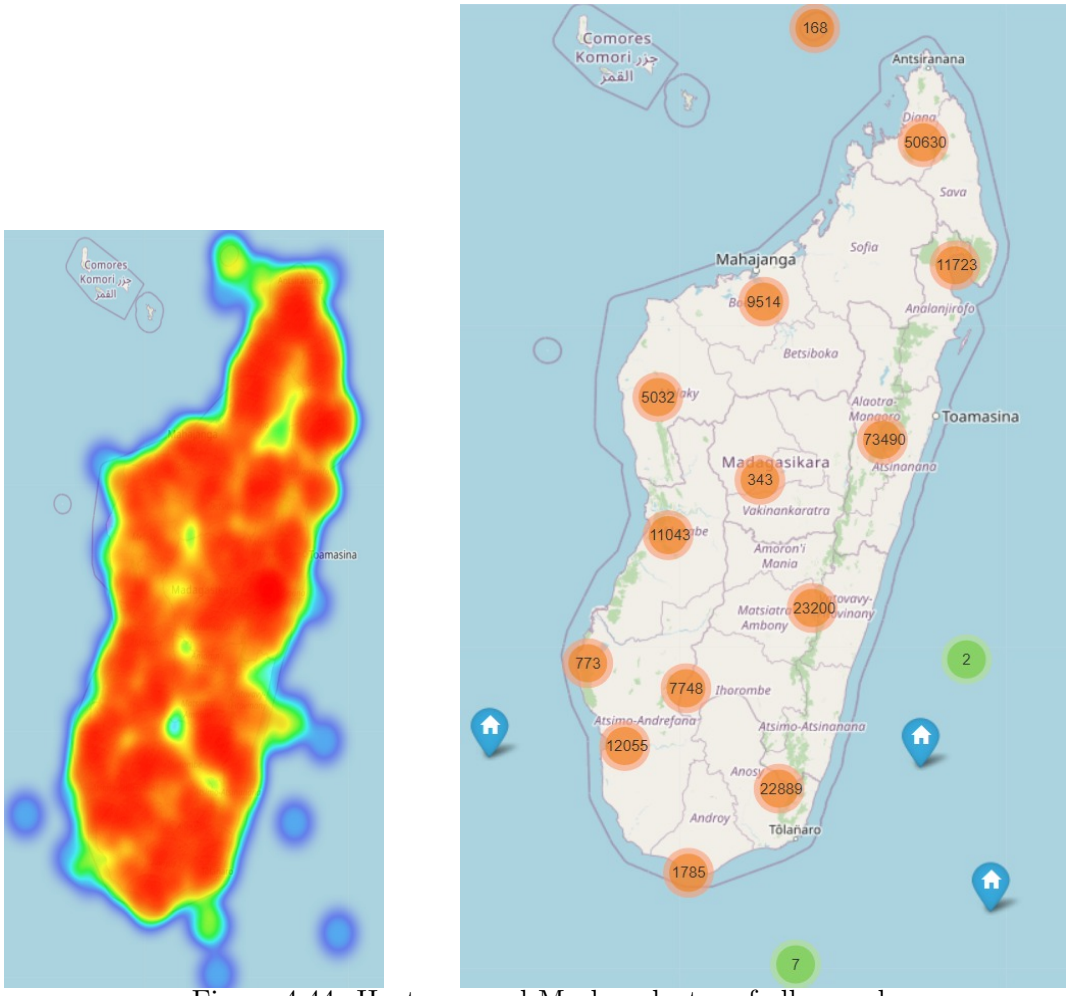

Figure 4.44: Heatmap and Marker cluster of all records

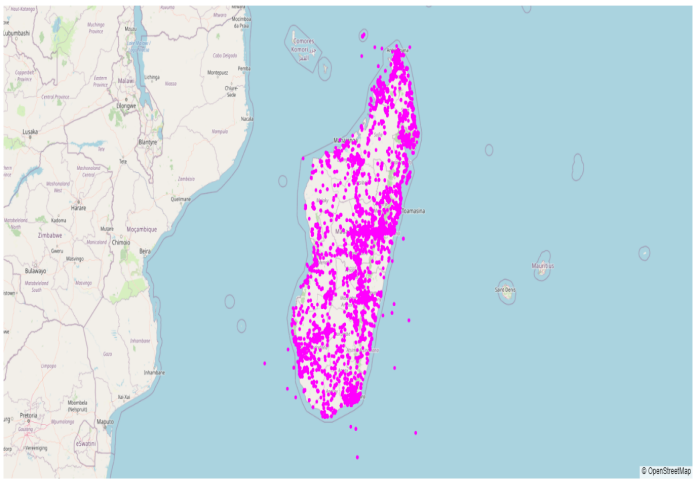

Figure 4.45: Scatter Map of all records

These maps describe the importance of Eastern Madagascar.

22. Maps related to localities (total localities - 7383)

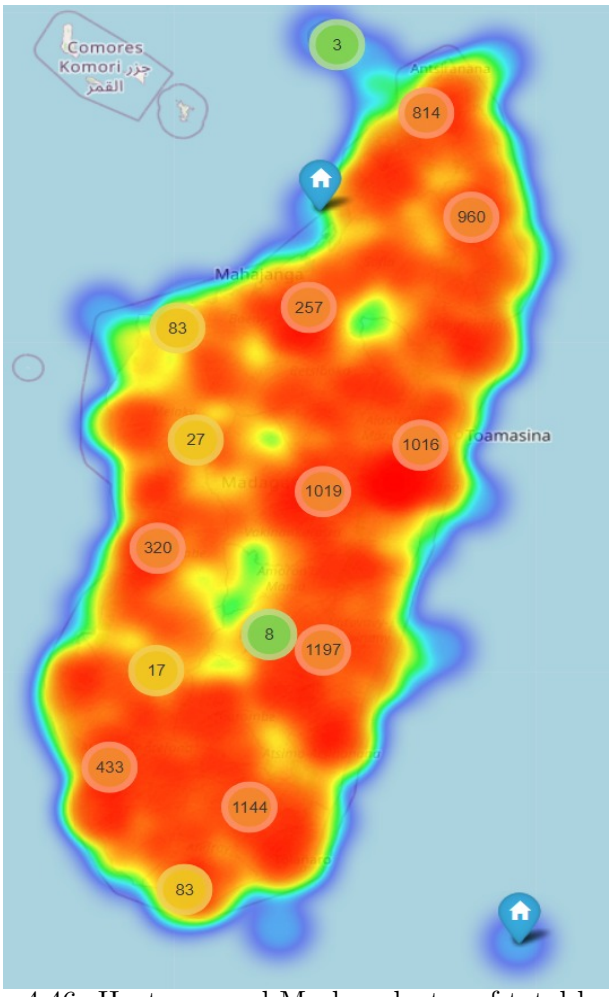

Figure 4.46: Heatmap and Marker cluster of total localities

The above map will help us to know about the localities in the Madagascar. There are total 7383 localities which having all records of all orders. Here we will see that maximum area is covered by the Eastern Madagascar.

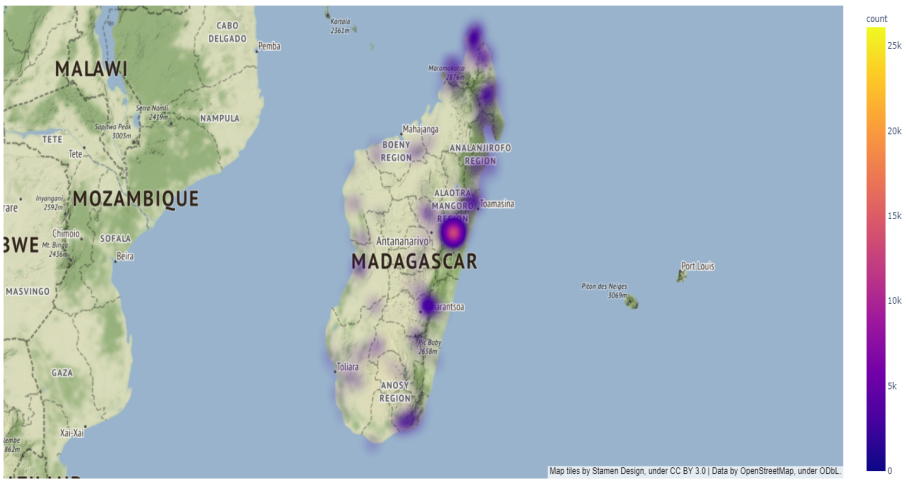

Figure 4.47: Density map of localities containing all records

The above map is representing the density of all records of all orders in different localities of Madagascar. When we see this map, we found that maximum number of records of all orders are occurred in the Eastern Madagascar.

Figure 3.3 is representing the each and every locality where we found records of all orders of insects. This information is necessary because here we talked about the localities in Madagascar.

Now we move to the conclusion on the basis all results we get in this chapter.

Chapter 5

Conclusions and Future work

In this chapter we summarize our work. We further analyze some results produced earlier and reason about why certain things worked the way they did. We also discuss some future research directions.

We can conclude that analyses on the level of entomological knowledge of a given area, even if based on partial databases, provide useful results with highly relevant geographical plots when performed on a very large number of records.

When we got above results in the geographical format, we definitely made one observation that Eastern Madagascar is playing very important role in the biodiversity of Madagascar. The most of the records of the different type of orders of insects are present in the Eastern Madagascar.

We observed that the orders of insects namely Hymenoptera, Diptera, Lepidoptera, Coleoptera are more occurred orders in the Madagascar. The occurrences of other orders are very low, so we need to conserve these species. To conserve these species there are protected areas in the Madagascar. But we need to increase those areas. When we see our results, we got know that where these species mostly occurs. So, on the basis of this knowledge, we will definitely help to increase protected areas. The localities which are having these records of all orders are mostly in the Eastern Madagascar.

When we think about future work regarding this project, GBIF organisation is playing very important role in this. GBIF provides us various datasets which are having the information about vertebrates, invertebrates, their oders, their kingdom at different location. Here we worked on the entomological information of Madagascar. We explore the entomological knowledge of Madagascar. By using GBIF datasets, we explore different continent's different species on the basis of their occurrences and we will get different geographical results on the basis of their occurrences. That means this will help us to conserve different different species in different continents.

Bibliography

[1] S. A. Chamberlain and C. Boettiger. R python, and ruby clients for gbif species occurrence data. Technical report, PeerJ Preprints, 2017.

[2] R. Chingkhei, A. Kumar, and R. N. Kumar. Block-4 gis analysis output and project design, 2017.

[3] esri. How to perform spatial analysis, 2018.

[4] J. U. Ganzhorn, P. P. Lowry, G. E. Schatz, and S. Sommer. The biodiversity of madagascar: one of the world's hottest hotspots on its way out. *Oryx*, 35(4):346–348, 2001.

[5] J. U. Ganzhorn, L. Wilmé, and J.-L. Mercier. Explaining madagascar's biodiversity. *Conservation and Environmental Management in Madagascar. IR Scales (ed.)*, pages 17–43, 2014.

[6] GBIF. Daraset used in this project, 2021.

[7] M. Iannella, P. D'Alessandro, and M. Biondi. Entomological knowledge in madagascar by gbif datasets: Estimates on the coverage and possible biases (insecta). *Fragmenta entomologica*, 51(1):1–10, 2019.

[8] P. Raju. Spatial data analysis. *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, page 151, 2003.