
Algorithms for Data Mining: Applications to Biodiversity

Thesis Submitted by
Moumita Ghosh

DOCTOR OF PHILOSOPHY (ENGINEERING)

Department of Information Technology
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India

2024

Name, Designation, and Institution of the Supervisors:

Name: Dr. Kartick Chandra Mondal

Designation: Assistant Professor

Institution of the Supervisor:

Department of Information Technology,

Jadavpur University, Salt Lake Campus

E-mail kartickjgec@gmail.com / kartick.mondal@jadavpuruniversity.in

Kolkata-700106

West Bengal

India

AND

Name: Dr. Anirban Roy

Designation: Research Officer

Institution of the Supervisor:

Department of Environment,

West Bengal Biodiversity Board,

Prani Sampad Bhawan (5th Floor) LB-2, Sector-III,

Salt Lake city

E-mail dr.anirbanroy@yahoo.co.in

Kolkata 700106

West Bengal

India

List of Publications

Journals

1. Moumita Ghosh, Sourav Mondal, Harshita Moondra, Dina Tri Utari, Anirban Roy, Kartick Chandra Mondal, An Irregular CLA-based Novel Frequent Pattern Mining Approach, *International Journal of Data Mining, Modelling and Management, Inter-science Journal*, 2023 (In press).
2. Moumita Ghosh, Kartick Chandra Mondal, Anirban Roy, Recognition of co-existence pattern of salt marshes and mangroves for littoral forest restoration, *Ecological Informatics*, Volume 71, 2022, 101769, ISSN 1574-9541, <https://doi.org/10.1016/j.ecoinf.2022.101769>. (<https://www.sciencedirect.com/science/article/pii/S1574954122002199>)
3. Moumita Ghosh, Pritam Sil, Anirban Roy, Rohmatul Fajriyah, Kartick Chandra Mondal, Frequent itemset mining using FP-Tree: A CLA Based Approach and its extended application in biodiversity data, *Innovations in Systems and Software Engineering - Springer*, 2022, <https://doi.org/10.1007/s11334-022-00500-3>. (<https://link.springer.com/article/10.1007/s11334-022-00500-3#citeas>)
4. Moumita Ghosh, Anirban Roy, Kartick Chandra Mondal, Knowledge Discovery of Sundarban Mangrove Species: A Way Forward for Managing Species Biodiversity, *SN Computer Science*, 3, Article number: 18 (2022), Electronic ISSN: 2661-8907, <https://doi.org/10.1007/s42979-021-00869-1>. (<https://link.springer.com/article/10.1007/s42979-021-00869-1>)
5. Kartick Chandra Mondal, Moumita Ghosh, Introducing Suffix Forest for Mining Tri-clusters from Time Series Data, *Innovations in Systems and Software Engineering - Springer*, <https://doi.org/10.1007/s11334-022-00489-9>. (<https://link.springer.com/article/10.1007/s11334-022-00489-9#citeas>)

International Conferences

1. Moumita Ghosh, Sourav Mondal, Anirban Roy, Kartick Chandra Mondal (2023), An Introduction to KDB: Knowledge Discovery in Biodiversity, *Computational Intelligence in Communications and Business Analytics. CICBA 2023. Communications in Computer and Information Science*, vol 1956. Springer, Cham. https://doi.org/10.1007/978-3-031-48879-5_24
2. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), Determining Dark Diversity of Different Faunal Groups in Indian Estuarine Ecosystem: A New Approach with Computational Biodiversity. In: Mandal, J.K., De, D. (eds) *Advanced Techniques for IoT Applications. EAIT 2021. Lecture Notes in Networks and Systems*, vol 292. Springer, Singapore. https://doi.org/10.1007/978-981-16-4435-1_16.
3. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), Analysis of Indian Estuarine Data of Flora & Fauna. In: Saraswat, M., Roy, S., Chowdhury, C., Gandomi, A.H. (eds) *Proceedings of International Conference on Data Science and Applications. Lecture Notes in Networks and Systems*, vol 287. Springer, Singapore. https://doi.org/10.1007/978-981-16-5348-3_31.

4. Moumita Ghosh, and Kartick Chandra Mondal (2022), Computational Biodiversity. In: Mandal, J.K., Buyya, R., De, D. (eds) Proceedings of International Conference on Advanced Computing Applications. Advances in Intelligent Systems and Computing, vol 1406. Springer, Singapore. https://doi.org/10.1007/978-981-16-5207-3_60.
5. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), FCA-Based Constant and Coherent-Signed Bicluster Identification and Its Application in Biodiversity Study. In: Mandal, J.K., Buyya, R., De, D. (eds) Proceedings of International Conference on Advanced Computing Applications. Advances in Intelligent Systems and Computing, vol 1406. Springer, Singapore. https://doi.org/10.1007/978-981-16-5207-3_57.

List of Patents

- Innovation Patent Number: 2021107061, Kartick Chandra Mondal, Somnath Mukhopadhyay, Sunita Sarkar, Samiran Chattopadhyay, Moumita Ghosh, Anindita Sarkar Mondal, Rohmatul Fajriyah, "Suffix Forest: A novel in-memory data structure for analyzing time-series data", Term of Patent: Eight years from 24 August 2021, Australian Govt. IP Australia.

List of Presentations in National/ International/ Conferences/ Workshops:

International Conference Presentation

1. Moumita Ghosh, Sourav Mondal, Anirban Roy, Kartick Chandra Mondal (2023), An Introduction to KDB: Knowledge Discovery in Biodiversity, Computational Intelligence in Communications and Business Analytics. CICBA 2023. Communications in Computer and Information Science, vol 1956. Springer, Cham. https://doi.org/10.1007/978-3-031-48879-5_24.
2. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), Determining Dark Diversity of Different Faunal Groups in Indian Estuarine Ecosystem: A New Approach with Computational Biodiversity. In: Mandal, J.K., De, D. (eds) Advanced Techniques for IoT Applications. EAIT 2021. Lecture Notes in Networks and Systems, vol 292. Springer, Singapore. https://doi.org/10.1007/978-981-16-4435-1_16.
3. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), Analysis of Indian Estuarine Data of Flora & Fauna. In: Saraswat, M., Roy, S., Chowdhury, C., Gandomi, A.H. (eds) Proceedings of International Conference on Data Science and Applications. Lecture Notes in Networks and Systems, vol 287. Springer, Singapore. https://doi.org/10.1007/978-981-16-5348-3_31.
4. Moumita Ghosh, and Kartick Chandra Mondal (2022), Computational Biodiversity. In: Mandal, J.K., Buyya, R., De, D. (eds) Proceedings of International Conference on Advanced Computing Applications. Advances in Intelligent Systems and Computing, vol 1406. Springer, Singapore. https://doi.org/10.1007/978-981-16-5207-3_60.
5. Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal (2022), FCA-Based Constant and Coherent-Signed Bicluster Identification and Its Application in Biodiversity Study. In: Mandal, J.K., Buyya, R., De, D. (eds) Proceedings of International Conference on Advanced Computing Applications. Advances in Intelligent Systems and Computing, vol 1406. Springer, Singapore. https://doi.org/10.1007/978-981-16-5207-3_57.

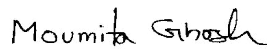
PROFORMA – 1
“Statement of Originality”

I, **Moumita Ghosh** registered on **18/06/2019**, do hereby declare that this thesis entitled **“Algorithms for Data Mining: Applications to Biodiversity”** contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.


I also declare that I have checked this thesis as per the “Policy on Anti Plagiarism, Jadavpur University, 2019”, and the level of similarity as checked by iThenticate software is **6 %**.

Signature of Candidate:

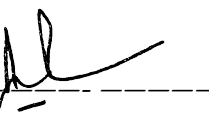


(Moumita Ghosh)

Certified by Supervisors:
(Signature with date, seal)

1. 

(Kartick Chandra Mondal)

2. 

(Anirban Roy)

Assistant Professor
Dept. of Information Technology
JADAVPUR UNIVERSITY
Block-LB, Plot-8, Sector-3
Salt Lake, Kolkata-700098, India

Research Officer
West Bengal Biodiversity Board
Dept. of Environment, Govt. of West Bengal
Prani Sampad Bhawan, 5th Floor
LB-2, Sector - III, Salt Lake City
Kolkata - 700106

PROFORMA – 2

“CERTIFICATE FROM THE SUPERVISORS”

This is to certify that the thesis entitled “**Algorithms for Data Mining: Applications to Biodiversity**” submitted by **Ms. Moumita Ghosh**, who got her name registered on **18/06/2019** for the award of Ph.D. (Engg.) degree of Jadavpur University is absolutely based upon her own work under the supervision of **Dr. Kartick Chandra Mondal, Department of Information Technology, Jadavpur University, Kolkata**, and **Dr. Anirban Roy, Department of Environment, West Bengal Biodiversity Board, Kolkata**, and that neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

1. Kartick Chandra Mondal

Signature of the Supervisor
and date with Office Seal

Assistant Professor
Dept. of Information Technology
JADAVPUR UNIVERSITY
Block-LB, Pic-8, Sector-3
Salt Lake, Kolkata-700098, India

2. [Signature]

Signature of the Supervisor
and date with Office Seal

Research Officer
West Bengal Biodiversity Board
Dept. of Environment, Govt. of West Bengal
Prant Sampad Bhawan, 5th Floor
LB-2, Sector-III, Salt Lake City
Kolkata - 700106

ACKNOWLEDGEMENTS

First of all, I wish to express my profound regards to my supervisors, **Dr. Kartick Chandra Mondal**, Assistant Professor, Department of Information Technology, Jadavpur University, and **Dr. Anirban Roy**, Research Officer, Department of Environment, West Bengal Biodiversity Board, for their valuable guidance, scholarly inputs, and consistent encouragement I received throughout the research work. No words would be sufficient to express my gratitude toward them. This feat was possible only because of the unconditional support provided by them. I consider it a great opportunity to do my doctoral programme under their guidance and to learn from their research expertise.

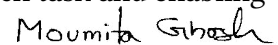
I want to express my deepest appreciation for the Head of the Department of Information Technology, members of the Research Advisory Committee (RAC), and members of the Research Committee (PRC) for their valuable inputs, helpful reviews, and suggestions. They were a great inspiration on the path of my research. Words are inadequate to express a deep sense of gratitude to members of the Doctoral Committee (DC), Jadavpur University, Kolkata.

I gratefully acknowledge the funding received for my Ph.D. research work from the Department of Science and Technology (DST) under the Scheme of Women Scientists Programs (DST WOS-A) for the duration of 2019 to 2022, by the Government of India.

I would like to thank all my co-researchers in the research lab at the Department of IT, Jadavpur University. All of them helped me directly or indirectly during this memorable journey.

Finally, I must express my sincere heartfelt gratitude to all the members of the Department of Information Technology of Jadavpur University who helped me directly or indirectly during this course of work.

The thesis would not be possible without the moral support of my husband, Mr. Pritam Das who has shown condense in me and supported me in completing my thesis work. My Parents and In-Laws have always been there to show their blessing and inspiration to complete this thesis work. Above all, I owe it all to Almighty God Lord Sri Krishna for granting me the wisdom, health, and strength to undertake this research task and enabling me to its completion.


Moumita Ghosh

ABSTRACT

One of the main areas of research in knowledge discovery from data (KDD) is pattern extraction. Association rule mining and bi-clustering are two significant data mining tasks that include pattern extraction as a primary component. In this thesis, I use data mining approaches to conduct case studies to extract pattern-based knowledge discovery from species biodiversity datasets. I create algorithmic frameworks that may be used by domain researchers to extract useful information from biodiversity datasets. I propose novel methods that make the most efficient use of resources to find frequent knowledge patterns. The proposed methods can build conceptually straightforward representations of frequent itemsets, bi-clusters, and association rules.

Biodiversity and ecosystem informatics is an emerging interdisciplinary field that mitigates the research gap among computer scientists, biologists, conservationists, and others related to various aspects of biological diversity. In this thesis, the study has spawned such initiatives and shown the support of algorithmic applications in the context of ecological improvements. From multiple directions, I have studied the efficacy of the existing, as well as the proposed algorithmic approaches to ecological data. I conduct a literature review on the current computational methods applied to biodiversity data.

A few case studies have been performed on Sundarban mangroves, Indian estuarine data of flora and fauna, and estuarine fish datasets. Also, exploratory data analysis has been performed on the Indian estuarine mangrove data using ecological indices. A case study using tri-clustering methodology is performed to study the year-wise changes in different categories (dense forest, medium dense, open forest) of forest cover across different states of India.

A novel framework has been proposed to compute the dark diversity of multiple faunal groups along the Indian estuarine ecosystem. Also, the incorporation of statistical methodology on top of the existing data mining approaches has been shown. Another novel data mining framework for studying salt marsh and mangrove species co-existence patterns has been proposed. All the experiments have been performed mainly on species presence/absence datasets, species occurrence datasets, and taxonomic datasets.

Two novel algorithms for extracting frequent patterns and frequent closed patterns have been proposed. Both use multiple FP-tree data structure and the concept of cellular learning automata. Detailed experimental analysis with respect to the leading algorithms on publicly available benchmark datasets show the better performance of the suggested approaches. Memory requirements and execution time requirements have been optimized for the proposed approaches. A novel concept has been formulated for extracting constant and coherent signed biclusters. Its domain-specific novel application in ecology has also been illustrated. Also, a framework along with a novel domain-specific rule filtering metric has been proposed and applied to IUCN red-listed Indian mangrove species.

CONTENTS

Table of Contents	1
I Introduction	1
1 Introduction	3
1.1 Introduction	4
1.2 Background and Motivation	5
1.2.1 Related to approaches	5
1.2.2 Related to applications	6
1.3 Research Objectives	7
1.3.1 Research Questions of the Thesis	8
1.4 Outline of the Thesis	12
2 State-of-the-art	15
2.1 Computational Biodiversity	16
2.2 Computational approaches in biodiversity	17
2.2.1 Biodiversity research groups	18
2.2.2 Comprehensive analysis on approaches	18
2.2.3 Comprehensive analysis on applications	18
2.3 Biodiversity Information Systems	20
2.4 Summary	22
3 Related terminology	23
3.1 Technological Background	24
II Case Study	31
4 Knowledge discovery on Sundarban mangrove	34
4.1 Introduction	35
4.2 Database specification	35
4.3 Database utility	36
4.4 Database analysis	38
4.4.1 Dataset 1	38
4.4.2 Dataset 2	42
4.5 Result and discussion	42
4.6 Summary	47
5 Exploratory data analysis on Indian mangroves	48
5.1 Introduction	49
5.2 Data Sources	49
5.3 Distribution of mangroves in India	49
5.3.1 Indian mangroves in east and west coasts of India	50
5.3.2 Indian mangroves in Sundarban	55

5.4	Diversity analysis	56
5.4.1	East and west coasts of Indian Mangroves	56
5.4.2	Indian Sundarban Mangroves	60
5.5	Summary	61
6	Analysing Indian estuarine flora and fauna	62
6.1	Introduction	63
6.2	Preparation of the dataset	64
6.3	Background Study and the Proposed methodology	68
6.4	Result and discussion	69
6.4.1	Discretized dataset of fauna	69
6.4.2	Presence-only dataset of fauna	71
6.4.3	Discretized dataset of flora	72
6.4.4	Presence-only dataset of flora	73
6.5	Summary	73
7	Knowledge discovery using tri-cluster	75
7.1	Introduction	76
7.2	Collection of data	76
7.3	Data preprocessing	76
7.3.1	Generating sorted-frequent dataset	78
7.3.2	Constructing frequent generalized itemset suffix-tree	82
7.3.3	Building frequent generalized suffix-forest	86
7.4	Tricluster generation and data interpretation	87
7.5	Extracting biclusters from suffix forest	88
7.6	Identifying rules from the clusters	89
7.7	Estimation of co-related parameters from triclusters	89
7.8	Summary	90
III	Proposed Framework	91
8	Data augmentation using dark diversity for finding species association	94
8.1	Introduction	95
8.2	Proposed framework	96
8.2.1	Dataset	96
8.2.2	Applying UNO function	97
8.2.3	Normalization and binarization of the dataset	98
8.2.4	Applying data mining methodology	98
8.3	Result and discussion	98
8.4	Summary	102
9	Mangrove regeneration framework using frequent co-existence pattern	103
9.1	Introduction	104
9.2	Materials	108
9.2.1	Study area	108
9.2.2	Data gathering and preprocessing	110

9.2.3	Dataset description	111
9.3	Methods	112
9.3.1	Proposed framework	112
9.3.2	Validation through statistical approach: multidimensional scaling (MDS)	112
9.4	Result	114
9.4.1	Finding FCI on <i>BSM</i> dataset	114
9.4.2	Association rule generation on <i>OEBM</i> , <i>MEBM</i> , <i>IEBM</i>	117
9.5	Discussion	121
9.5.1	Finding inferences from exact association rules	122
9.5.2	Predicting novel associations from approximate association rules	124
9.5.3	Implications on restoration practitioners	126
9.5.4	Comparisons with parallel studies in mangrove restoration in terms of used methodology, findings, and limitations	127
9.6	Summary	130
10 Knowledge Discovery in Biodiversity (KDB): Web application prototype in ecology		132
10.1	Introduction	133
10.2	Exploring KDB	134
10.3	Application	139
10.4	Datasets	139
10.5	Comparison with data mining tool	142
10.6	Comparative discussion on multiple previous case studies	142
10.7	Summary	143
IV Proposed Algorithm		144
11 FP tree-based frequent pattern mining using CLA		147
11.1	Introduction	148
11.1.1	Problem and motivation	148
11.1.2	Previous studies	149
11.1.3	Contribution	150
11.2	Approach	150
11.2.1	Algorithmic explanation	151
11.3	Completeness and correctness of the proposed algorithm	159
11.4	Result discussion and evaluation	161
11.4.1	Compared algorithms	161
11.4.2	Theoretical performance analysis	161
11.4.3	Test datasets	162
11.4.4	Running environment	163
11.4.5	Empirical performance comparison	163
11.5	Summary	167

12 FP-Forest based frequent closed pattern mining using CLA	168
12.1 Introduction	169
12.1.1 Problem and motivation	169
12.1.2 Previous studies	170
12.1.3 Contribution	170
12.2 Proposed approach	171
12.2.1 Algorithm CellBiClust	171
12.2.2 Analysis of Algorithmic Result:	181
12.3 Performance evaluation	183
12.4 Application on ecological forecasting	187
12.5 Summary	190
13 FCA-based constant and coherent-signed bicluster identification	191
13.1 Introduction	192
13.1.1 Background and motivation	192
13.1.2 Related work	192
13.1.3 Contribution	193
13.2 Constant and coherent signed biclusters	193
13.3 Pattern structure for the signed partition & bicluster generation	197
13.4 Summary	197
14 A data-driven rule filtering approach for forest restoration	199
14.1 Introduction	200
14.2 Previous studies and scope of the data mining approach	202
14.3 Materials and methods	204
14.3.1 Dataset description	204
14.3.2 Proposed framework	205
14.4 Results	208
14.4.1 Analysis on frequency of occurrence	208
14.4.2 Analysis on objective and subjective measures	208
14.5 Discussion	219
14.5.1 Comparison to the earlier findings	219
14.5.2 Potential effects on conservationists	220
14.5.3 Comparison to concurrent research initiatives	221
14.5.4 Current Situation of Mangroves in India: Emerging Threats, Policy, and Future Suggestions	221
14.6 Summary	223
V Conclusions	224
15 Conclusion and future scope	226
15.1 Conclusions	227
15.2 Future Scope	230
REFERENCES	231

Part I

Introduction

CHAPTER 1

INTRODUCTION

1.1	Introduction	4
1.2	Background and Motivation	5
1.2.1	Related to approaches	5
1.2.2	Related to applications	6
1.3	Research Objectives	7
1.3.1	Research Questions of the Thesis	8
1.4	Outline of the Thesis	12

1.1 Introduction

Biodiversity is essential for the economic and ecological security of human beings. The last few decades have been responsible for eroding biodiversity at an alarming rate due to rapid urbanization and alteration of habitats as well [1, 2, 3]. Thus, the conservation of ecosystems is of utmost need. Among the ecosystems, the forest ecosystem is very significant for being a very highly productive system and providing notable ecological services [4]. The components of a forest ecosystem may broadly be classified into two directions: abiotic components and biotic components. Abiotic components include soil, moisture, air, and sunlight, whereas biotic components mainly consist of producers, consumers, and decomposers. Here, we will concentrate mainly on biotic components, for example, plant species. It is obvious that without proper monitoring of these forest components and the changes in forest covers over time including the driving factors, the sustainable management of this ecosystem would not be possible [5]. Analysis of data from this domain may help to understand the scenario and to take necessary measures. Data mining tools and packages yield interesting analytical results [6] which we would like to apply to the biodiversity data of Indian forests.

Forests are the natural security forces that have immense ecological service in controlling a number of climate catastrophes, preventing soil erosion, inhibiting inward ingression of the sea in mangrove areas, and providing an ecological niche for animals and livelihood for humans. Despite providing excellent ecological and economic functions, there has been a noticeable loss in forest diversity [7]. Considering the mangrove ecosystem, it is estimated that up to 35% of the mangrove area on the Earth has been lost since the 1980s, mostly as a result of various developmental activities on the coast [8, 9, 10]. However, concern over the disappearance of coastal mangrove regions has recently grown significantly [9, 11, 12, 13, 14, 15, 16, 17]. These coastal habitats are said to be most fragile due to their susceptibility to climate change vis-a-vis sea level rise. The continuous loss and fragmentation of such habitats hinder species migration/ dispersal, obliterate local coastal resilience, and drive essential mangrove ecosystems into collapse [9]. For example, a case study reveals that between 1986 to 2012, 124.418 sq. km. of mangrove forest cover was lost. Article [18] highlights the loss in the mangrove forest in the Indian Sundarban is near about 5.5%. Different causes like overexploitation and illegal forest cutting, pollution, climate change, etc. are identified as the most dominant factors [18] for the degradation of a forest ecosystem. Therefore, applying cutting-edge techniques that include data mining could be of major importance to restoring and rehabilitating the precious ecosystem through proper management [19].

Mainly statistical analysis techniques are used in a few research articles but those are only confirmatory analysis techniques with respect to researchers' understanding. However, data mining tools perform exploratory analysis which is concerned with detecting and describing patterns within data, identifying predictor variables, and discovering the forms of relationships between predictors and responses. Thus, knowledge discovery in the data mining process is capable of identifying valid, potentially useful, and understandable patterns which is not a new application in the biodiversity domain. But, very few studies have been conducted in this area of research [20, 21, 22, 23]. The proposal of domain-specific

data mining algorithms along with statistical models in the regeneration and restoration of forest ecosystem will be a significant contribution towards the conservation of biodiversity [24].

1.2 Background and Motivation

1.2.1 Related to approaches

Classical data mining approaches in biodiversity study

There are very few studies that use data mining techniques to address problems with biodiversity. In a survey report [24] on biodiversity research, the problems and difficulties with biodiversity fieldwork are mentioned, and the useful application of data mining approaches is shown. For example, [25], and [26], both employed the linear regression technique in data mining. Logistic regression technique has been applied in species distribution model [27, 28]. In [29], the decision tree method is used to forecast water quality data and performs better in terms of accuracy, than statistical analysis techniques.

Pattern and rule mining approaches used in biodiversity study

For biodiversity study association rule generation technique is used by a few researchers to analyze species diversity [30], mining species distribution pattern [31, 32]. Two main traditional approaches in finding frequent closed patterns in association rule mining are, i) Breadth-first search (BFS) approach is used by Apriori-like algorithms for candidate generation and testing, and, ii) Depth-first search (DFS) approach is used by FP where the dataset is transferred to a compressed tree structure. But, Apriori-like algorithm [33] generates a large number of frequent patterns, and frequent scanning of the database is needed. The FP-Growth-like methods [34] are not suitable as they cannot be fitted in memory where the size of a tree is large. From the problem with these two types of algorithms, the bitwise frequent mining approach [35] is proposed but it faces inadequate memory space. In this context, a partition-based parallel mining algorithm is also proposed. In [35], the authors propose distributed cellular learning-based parallel mining algorithms and find an efficient and faster way for frequent itemset mining. Thus by generating efficient and faster association rules, the concept of cellular learning automata can be used fruitfully.

Bi-clustering and tri-clustering for biodiversity

Biclustering, a data mining technique, can be applied to biodiversity research in several ways. Biclustering allows for the simultaneous clustering of rows (samples) and columns (features) of a dataset, which can be particularly useful in analyzing high-dimensional biodiversity data. Although, its main application has been found in bioinformatics research [36, 37]. Article [38] successfully implements this technique for assessing birds' popula-

tion status. Biclustering can help identify associations between species and environmental variables. By simultaneously clustering species and environmental variables, biclustering methods can reveal patterns of species distribution that are correlated with specific environmental conditions. Biclustering can be used to analyze community structure by identifying subsets of species that co-occur in particular habitats or ecological niches. Biclustering can help identify groups of species with similar functional traits.

Triclustering [39], an extension of biclustering is a data mining technique that allows for the simultaneous clustering of rows, columns, and layers in a three-dimensional dataset. TriClustering is mainly used for analyzing gene expression data [40]. But it has the potential to provide additional insights into complex biodiversity data by identifying subsets of species, environmental variables, and spatial or temporal dimensions that exhibit coherent patterns.

1.2.2 Related to applications

Conservation and Management

The mangrove ecosystem is incredibly productive and offers millions of people a variety of goods and services. However, the quality of the mangrove environment is in a declining trend during the past few decades. Mangrove forests are the best carbon sinks and have the highest ecosystem service value(ESV). Due to severe natural and human-caused challenges, the total loss of ESVs in the Sundarban Biosphere Reserve (SBR) during the past 45 years has been assessed at 3310.79 million USD [41]. Unauthorized encroachment to the Sundarban forest is strictly prohibited through legislation and regulation to mitigate climate change disasters. It is essential to restore mangroves by data-driven planting, reforestation, and afforestation [42, 43]. Therefore, the data-driven framework might be seen as a step in the right direction for restoration professionals, stakeholders, and researchers to safeguard mangroves holistically.

Impacts of climate change and increasing salinity in coastal region

Climate-change-driven sea level rise causes an increase in salinity in coastal wetlands accelerating the alteration of the species composition. It triggers the gradual extinction of species, particularly the mangrove population which is intolerant of excessive salinity. Thus despite being crucial to a wide range of ecosystem services, mangroves have been identified as a vulnerable coastal biome. Hence restoration strategy of mangroves is undergoing rigorous research and experiments in literature at an interdisciplinary level [44, 45, 46, 47]. From a data-driven perspective, analysis of mangrove occurrence data could be the key to comprehending and predicting mangrove behavior along different environmental parameters, and it could be important in formulating a management strategy for mangrove rehabilitation and restoration. As salt marshes are natural salt-accumulating halophytes, mitigating excessive salinity could be achieved by incorporating salt marshes in mangrove restoration activities. Therefore it would be valuable to find a novel restoration strategy by assess-

ing the frequent co-existence status of salt marshes, with the mangroves, and mangrove associates in different zones of degraded mangrove patches for species-rich plantation. Article [21] explores the impact of climate change on the poor community of Bangladesh, as mangrove forests are highly affected by the changes in aquatic salinization.

Although community-based ecological mangrove restoration and environmental engineering have been studied previously, data-driven approach identifies previously unidentified, practically applicable, and easily accessible interpretations of knowledge required for decision-making. Examples of the application of data mining algorithms on mangroves in ecology are rare [48].

Novel pattern discoveries

Data mining offers the chance to make new discoveries and insights. Researchers can discover previously undiscovered patterns, species connections, or ecological processes that support forest biodiversity by analyzing huge and varied datasets. Such findings can improve our understanding of forest ecosystems and present opportunities for further study and conservation initiatives. It is hypothesized that closely associated plants experience more frequent coexistence patterns than animals [49]. The homogeneity of characters in the process of evolution, such as adaptability, migration, resource uptake, and reproductivity, are the primary influencers behind the co-existence pattern of different plant species for ecosystem resilience [50]. So, instead of using prevalent, widespread plant species for restoration purposes [51], exploration of co-occurrence data can suggest suitable species for system restoration. Hence, finding co-occurrence patterns can reduce the knowledge gap between data analyzers and restoration practitioners, and the advancement of knowledge sharing could assist domain researchers.

Mangroves and livelihood in India

According to the findings of the econometric research [52, 53], the region's limited agricultural and aquaculture production, as well as diminishing trends in the harvesting of prawn seeds, honey, and other non-timber forest items, pose the greatest risks to the lives of the locals in Sundarban. More than 70% of respondents cited livelihood hazards brought on primarily by the effects of climate change as the primary driver of such migration. Overall, the loss of mangroves in the Indian Sundarban region can have a major effect on poverty levels by lowering the availability of livelihood possibilities and exposing inhabitants nearby to greater risk of natural catastrophes and climate change effects.

1.3 Research Objectives

The present review of the literature shows that there are numerous challenges and problems related to forest ecology [54, 55, 18]. In spite of many statistical and algorithmic solutions to the challenges and problems of forest ecology, still, there is immense scope in applying

computational and algorithmic approaches, especially in the littoral forest ecosystem. In this backdrop, the present research entitled "Algorithms for Data Mining: Applications to Biodiversity" has been undertaken with the following objectives:

1. To comprehend the current research in the related field and determine the applicability of computational approaches in ecological research.
2. To examine the usage of data mining methods and technologies in ecological research.
3. To propose efficient domain-specific data mining frameworks to demonstrate how the suggested frameworks could potentially be used for ecosystem management and conservation.
4. To propose novel and resource-efficient data mining algorithms and show its adaptation in ecological research.
5. To compile a few distinctive datasets to examine the suggested frameworks and algorithms is also the primary requirement of this research study.

1.3.1 Research Questions of the Thesis

To effectively complete the study objectives, a few research questions have been put forth within the confines of this study. These research questions are considered as the building blocks of the present work and our target is to explore the answers to these questions. Figure 1.1 shows the mapping among the objectives, the various research questions, and their respective chapters along the published articles.

Research Question1: What is the motivation behind the study of Computational Biodiversity?

The motivation behind the study of Computational Biodiversity lies in the urgent need for understanding the conservation and sustainable use of biodiversity. Computational methods offer powerful tools to analyze and interpret vast amounts of biodiversity data, providing valuable insights into the distribution, composition, and dynamics of species and ecosystems.

The related study is addressed in **Chapter 2** and published in **article [56]**.

Research Question2: How do multiple mangrove patches differ in India?

The Jaccard dissimilarity index is a measure used in ecology research to quantify the dissimilarity between two ecological assemblages or communities. It is a commonly used index to compare the composition of species or other ecological entities in different sites or samples. Furthermore, species richness, species evenness, and Pielou's evenness (J') are indeed commonly used indices for alpha diversity in ecology research.

The related study is addressed in **Chapter 5**, and communicated through **article [57]**.

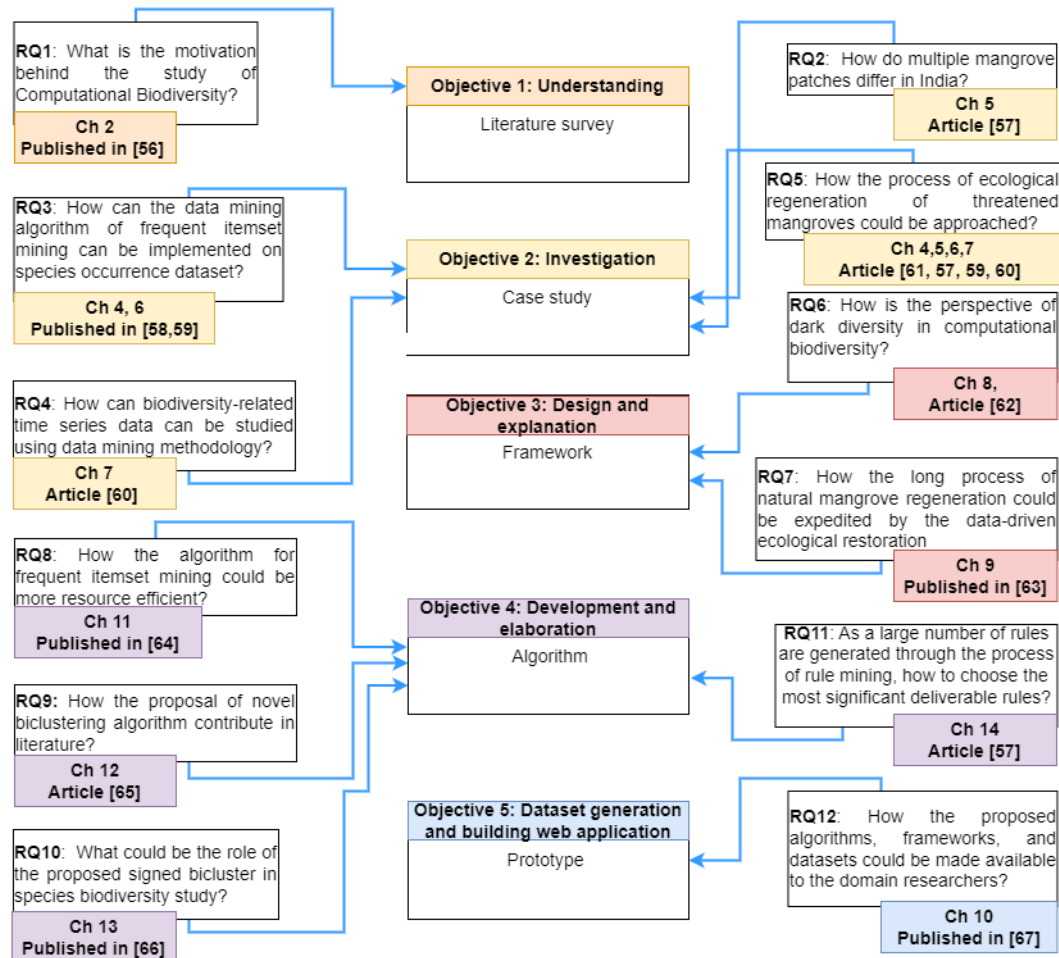


Figure 1.1: The mapping among the objectives, the various research questions, their respective chapters along with the published articles

Research Question3: How can the data mining algorithm of frequent itemset mining can be implemented on species occurrence dataset?

Species occurrence data is one of the most influential quantities in biodiversity/ ecological research. The occurrence data gives a view of the species distribution across the survey sites. These types of data may have a vision of the transaction dataset where traditional market basket analysis can be applied. For this, it should have a perspective of binary data representation. The mapping of species occurrence data on the transaction dataset would be alike all the sites would be along the rows and species along the columns. The matrix cells would contain the value corresponding to the occurrence of a species at a particular site.

The related study is addressed in **Chapters 4, and 6**, and published in **articles [58, 59]**.

Research Question4: How can biodiversity-related time series data can be studied using data mining methodology?

Time series data can be considered three-dimensional data. Three-dimensional data is

a collection of multiple two-dimensional data matrices that are of great importance for representing as well as understanding the relationship among data in various domains. In the case of species data along multiple locations and time dimensions, analytic queries will be like changes in species presence records along with different locations at varying times.

We have used the forest cover data in different states and in different years. The applied tri-clustering algorithm will reveal knowledge related to clusters of regions, along with time and forest cover type.

The related study is addressed in **Chapter 7** and published in **article [60]**.

Research Question5: How the process of ecological regeneration of threatened mangroves could be approached?

Instead of using prevalent, widespread plant species for restoration purposes, exploration of co-occurrence data can suggest suitable species for system restoration. Hence, finding co-occurrence patterns can reduce the knowledge gap between data analyzers and restoration practitioners, and the advancement of knowledge sharing could assist domain researchers. The use of the data mining algorithm is able to generate frequent closed itemset and biclusters that can be interpreted by the domain researchers for policy making in restoration practice.

The related studies as addressed in **Chapters 4, 5, 6, and 7** and published in **articles [61, 57, 59, 60]**.

Research Question6: How is the perspective of dark diversity in computational biodiversity counteract the biodiversity loss?

The proposition of applying dark diversity before processing the rule mining task reveals the information related to the absent part of the occurrence data. This helps in proper management in a survey or resurvey aiming at finding new sites for probable habitat for a particular species.

The related study is Addressed in **Chapter 8** and published in **article [62]**.

Research Question7: How the long process of natural mangrove regeneration could be expedited by the data-driven ecological restoration that entails rejuvenating native ecosystems in vulnerable areas while preserving the diversity of local flora and fauna through regeneration with a much shorter regeneration time?

The following proposals are made:

1. We made a proposal for an excessive salinity-affected mangrove community restoration approach where hyper-salinity could be neutralized by growing suitable salt marshes.
2. We perform a case study on Sundarban coastal area considering major environmental/ habitat factors, such as salinity, pH, soil texture, and tidal amplitude, along with the occurrence data of mangroves, mangrove associates, and salt marshes and compile 3 different datasets for inner, middle, and outer estuarine species records.
3. We establish salt mash-salt marsh co-existence pattern along the salinity gradient; salt marshes, mangroves, and mangrove associates co-existence patterns

with varying environmental factors; probable inter-species association from present co-existence data.

The related study is addressed in **Chapter 9** and published in **article [63]**.

Research Question 8: How the algorithm for frequent itemset mining could be more resource efficient?

The proposed algorithm is memory-efficient without compromising the time requirement as compared to the classical as well as recent algorithmic developments for mining frequent itemsets. The proposed algorithm for mining frequent itemset is efficient as,

- It builds a highly compact dense dataset (DDS) from a larger input transaction dataset, which is considerably smaller than the initial dataset and therefore saves the cost of larger dataset scans in the successive mining processes.
- It involves a partitioning-based method that dramatically reduces the number of transaction rows that have to be scanned to build a single tree.
- Several optimizations have been incorporated into the algorithm implementation steps, such as hashmap and hash table data structures. Also, parallelism has been incorporated in building multiple FP tree structures. These lower the execution time and are implemented in an optimized way to extract the frequent itemsets.

The related study is addressed in **Chapter 11** and published in **article [64]**.

Research Question9: How the proposal of novel biclustering algorithm contribute in literature?

Along with the proposal of a novel biclustering algorithm, predicting novel incidences or occurrences from the binary dataset is included to extend the use of biclustering. Experiments reveal that the proposed algorithm is efficient in terms of memory usage and execution time. It has a novel application on biodiversity in occurrence rule generation and prediction.

A species occurrence dataset could be mapped to a binary dataset where the species are across the rows, and the observation sites are across the columns. The frequent co-occurrence pattern of species has become a focused issue in ecology because it gives insight into the species distribution pattern. It needs substantial domain knowledge to identify those patterns. With the recent advancements of the data-science-based approach, examining frequently occurring species patterns could be trouble-free. We employ our proposed algorithm on a binary dataset of species occurrences. The analysis of the outcome is justified by the domain expert. The intuitive knowledge and understanding of the domain expertise in the ecological domain play an essential role in guiding us in our data mining application for ecological forecasting.

The related study is addressed in **Chapter 12** and communicated in **article [65]**.

Research Question10: What could be the role of the proposed signed bicluster in species biodiversity study?

The proposed constant and coherent sign-changing bicluster is novel as these are not

restricted to binary symbols. The change in symbolic direction has been addressed instead of the magnitude of the attribute value. This kind of multi-symbolic sign-changing bicluster identification and its domain-specific study is new in the literature.

As a case study, we have curated a dataset on the mangrove cover changes over the years to demonstrate the employment of signed bicluster.

We represent the data symbolically to express the changes in species count data that is given between 1986 to 2014. Each symbol conveys a specific meaning. We have used five such symbols, viz; 0, -, ~, +, and 1; where,

- 0: Historical absence (A species is absent in both 1986 and 2014)
- -: Range contraction (Decrease in species count)
- ~: Unchanged (Same abundance data in both 1984 and 2014)
- +: Range expansion (Increase in species count)
- 1: Introduced (Absent in 1986 but present in 2014)
- -1: Disappeared (Present in 1986 but absent in 2014)

The related study is addressed in **Chapter 13** and published in **article [66]**.

Research Question11: As a large number of rules are generated through the process of rule mining, how to choose the most significant deliverable rules?

Association rule mining extracts a large set of association rules. For obtaining the most significant rules, domain knowledge-based rule filtering measures can be used in addition to the conventional measures of association rule mining. Hence, a set of significant rules can be presented as useful for domain researchers in species-rich plantations/ ecosystem restoration.

The related study is addressed in **Chapter 14** and communicated in **article [57]**.

Research Question12: How the proposed algorithms, frameworks, and datasets could be made available to the domain researchers?

The most basic method of experimentation using data mining algorithms is the command prompt. A convenient approach of interactive graphical user interfaces(GUI) can be supplied for data exploration to build up complex studies. A GUI can be a platform for domain researchers to apply their datasets to domain-specific data mining algorithms for further analysis.

The related study is addressed in **Chapter 10** and communicated in **article [67]**.

1.4 Outline of the Thesis

Image 1.2 shows the outline of the thesis. The overall thesis is divided into multiple parts as follows.

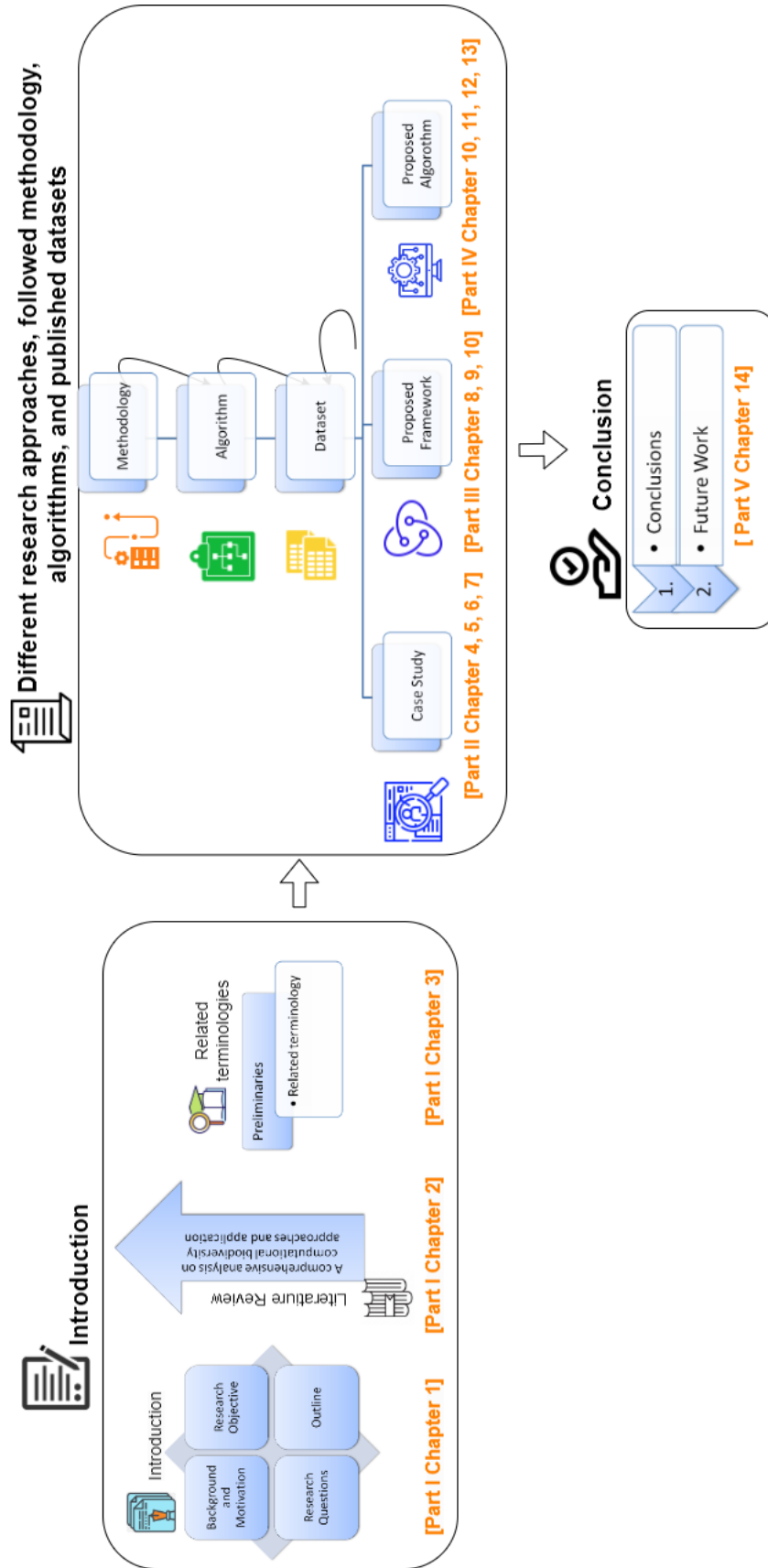


Figure 1.2: Outline of the Thesis

Part I provides a brief introduction in Chapter 1. A discussion related to the literature on computational approaches made to biodiversity data has been presented in Chapter 2. Chapter 3 lists the related terminologies used in this research work.

Part II focuses on the prevalent methodologies to perceive its usage in biodiversity data. It comprises four chapters (4 to 7) and describes the case studies. Chapter 4 explains the knowledge discovery on Sundarban Mangrove. Exploratory data analysis has been performed on the compiled dataset of Indian mangroves in Chapter 5. Estuarine data on flora and fauna has been studied in Chapter 6. A case study has been performed on the curated dataset of Indian state-wise forest cover data along the time dimension to extract meaningful information from 3-dimensional data and is shown in Chapter 7. We exploit the usage of tri-cluster algorithm for this.

Part III describes the proposed frameworks. It comprises 3 chapters (Chapter 8 to 10). By fusing statistical measures and data mining methods, Chapter 8 demonstrated a novel framework that was tailored for the domain research to compute dark diversity. Chapter 9 aims to identify a unique restoration approach of mangroves and mangrove associates in various zones of degraded mangrove patches. Chapter 10 presents an application prototype for domain researchers. The platform provides an integrated tool for datasets and algorithms that are specifically designed for working with species occurrence datasets, although they are relevant to other datasets as well.

Part IV focuses on the novel methodology proposals. It comprises four chapters (Chapter 11 to 14). The proposals include an efficient algorithm for frequent itemset mining (Chapter 11), frequent closed itemset mining (Chapter 12), novel signed biclustering (Chapter 13), association rule filtering (Chapter 14). The proposed algorithm for frequent itemset mining has been illustrated on mangrove occurrence data to identify the rare mangroves. The estuarine fish occurrence dataset has been studied using the proposed algorithm for frequent closed itemset mining. Mangrove cover changes over the years have been studied using the proposal of the constant and coherent signed biclusters. A domain-specific rule filtering methodology has been proposed for the restoration strategy. IUCN red-listed species data have been studied using the methodology.

The thesis concludes in Part V. This part provides a pictorial view of the whole contribution made in this research work. Chapter 15 presents the conclusion of all the works done and draws some perspectives on the future scope.

CHAPTER 2

STATE-OF-THE-ART

2.1	Computational Biodiversity	16
2.2	Computational approaches in biodiversity	17
2.2.1	Biodiversity research groups	18
2.2.2	Comprehensive analysis on approaches	18
2.2.3	Comprehensive analysis on applications	18
2.3	Biodiversity Information Systems	20
2.4	Summary	22

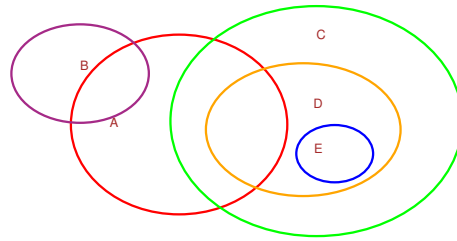


Figure 2.1: Illustration using Venn diagram: Regions A, B, C, D, E representing the domain of data mining, statistics, artificial intelligence, machine learning, deep learning respectively

2.1 Computational Biodiversity

The term computational biodiversity is a new phrase. Since the last few decades, biodiversity is declining globally and its shrinking rate poses a threat to many species. Therefore, there is an evolving need to recognize and evaluate complex ecological problems. Along with the statistical measures secured by ecologists, computer science researchers have perceived the prospect of algorithmic solutions in coming up against adverse environmental issues. Thus the application of various computational methodologies in biodiversity may coin the term computational biodiversity. In this work, we perform a comprehensive study on recent progress made toward the protection of biodiversity and thus highlight the importance of collaboration between ecologists and computer science researchers. We found that the recent computational approaches have broadened the data-driven modeling capability where algorithmic developments can extrapolate the behavior of the environmental variables and find relationships among them. Therefore, we may conclude that the computational approaches can infer holistic solutions toward ecological resilience.

Background: Biodiversity describes the whole range of the different varieties of living organisms. It is the single most important factor behind the equilibrium of the earth. Preservation of biodiversity is needed in order to maintain a stable ecosystem that consists of a biological community of living organisms and their nonliving components, together in a balanced form. Conservation of biodiversity offers a healthy, nutritious, and diverse ecosystem, feasible populations of species, genetic wealth, and sustainable use of biological resources. Presently, maintaining biodiversity is a crucial challenge, as it is difficult to properly monitor different biological components and their changes over time including driving factors. Data science can meet this necessity by affording a lot of computational approaches. Different computational approaches that we consider here are artificial intelligence (AI), machine learning (ML), data mining (DM), deep learning (DL), and statistical measures. All these domains are interrelated. AI is fairly recognizable from the rest. It behaves like an intelligent agent by controlling/ monitoring a circumstance. AI forms a superset for ML and DL. AI utilizes the models built by ML. A more powerful version of ML is DL. Both AI and ML use DM techniques and other learning algorithms for model development. DM exploits statistics to find patterns and phenomena. Figure 2.1 illustrates their relationships.

The motivation behind the study on computational biodiversity: A report on the case study [68] of the Ganges river basin has shown that more than 10% of the World's human population depends on the Ganges river basin. Despite this, the loss of biodiversity of the river is at an increasing rate. Some fish species like Hilsa (*Tenuous* is), Tiger prawns (*Macro brachium* Rosenberg), etc. are at the extinction edge. Having assessed the

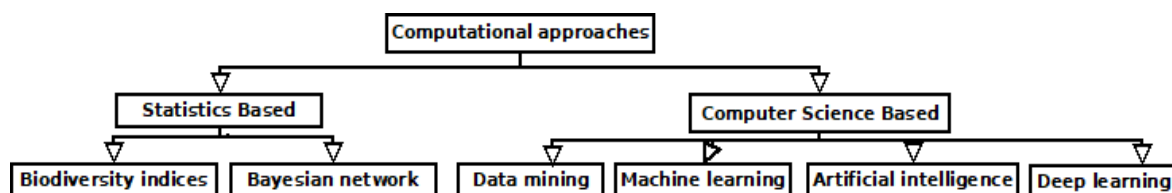


Figure 2.2: Different computational approaches

total economic worth of benefits, achieved by river Halda, Bangladesh, the authors recommend the urgent need to identify the endangered fish habitat, particular reasons for their steady disappearance, suitable environment for their growth, and conservation strategies [69]. In respect of the forest ecosystem, a case study [70] of Odisha, states that Odisha has covered 7.1% of the total forest of India (FSI, 2011), and it is severely affected as it loses its green cover gradually. Forest canopy closure data, fragmentation pattern, forest fire distribution, and impact of biological invasions help in measuring the degradation of the forest ecosystem. Conservation of Western Ghats freshwater biodiversity (A case study on decapod crustaceans) is studied in [71] where the distribution pattern analysis of freshwater invertebrates in the Western Ghats helps to find out the prior areas for conservation.

Biodiversity measure like socio-cultural measures basically is aiming at surveys and interviews for the ecosystem and environment management, economic evaluation quantifies the impact of biodiversity loss in monetary terms, ecological indicators are generally used to measure the species richness and identify endangered species. Though all of these help in measuring biodiversity from different perspectives, significant studies are still in need that use computational techniques.

Summarizing the discussion, we can say that the use of computational approaches in the biodiversity domain is highly appreciable and is in demand as it has the capability of treating heterogeneous and voluminous data. Also, it can handle a scalable dataset. Algorithm-based approaches give accurate, reliable prediction and the capability to handle a huge amount of data.

Methods and Objective: The aim of this study is to present a synopsis of the state-of-the-art computational approaches attempted in this field. We present a hierarchical classification of different computational approaches that have been used in this domain. We identify multiple research groups and their contributions. Both qualitative analysis (based on the problem and objective) and quantitative analysis (statistical study) of the collected articles are performed here. Our intention is to assist the researchers in gaining insight into the different computational approaches and their associated techniques to come up with new studies in biodiversity research for conservation purposes.

2.2 Computational approaches in biodiversity

Different computational approaches followed in the biodiversity study are briefly explained in this section (Table 2.1). Figure 2.2 segregates the methodologies that fall under different approaches.

We have categorized the commonly applied computational methods broadly in two categories - based on computer science and based on statistical analysis. Our main focus, here, is to analyze the application of different algorithms developed in computer science, specifi-

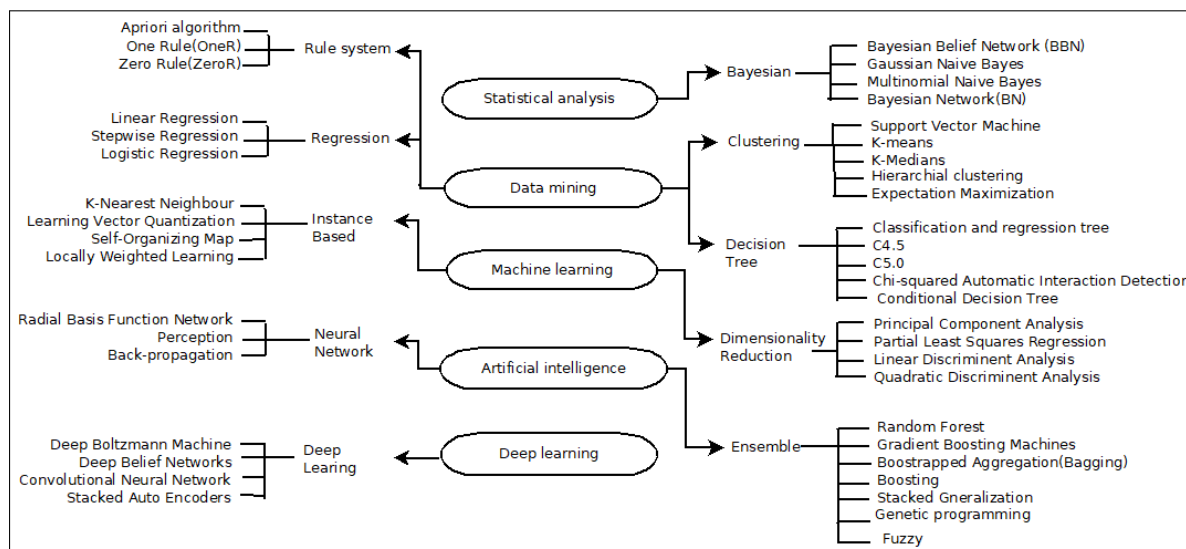


Figure 2.3: Different computational approaches used in biodiversity study

cally in the field of machine learning, data mining, artificial intelligence, and deep learning. In Figure 2.3, we have summarized the details of different approaches mainly used in the research on biodiversity.

2.2.1 Biodiversity research groups

This section highlights different groups of researchers working in the areas of various computational methodologies which are mentioned in section 2.2. A few research groups (Table 2.2) are mentioned here, primarily working with biodiversity data of different ecosystems:

2.2.2 Comprehensive analysis on approaches

Figure 2.4 depicts different algorithmic approaches that have taken for application in different directions. We classify the different research articles according to their main research interest. We find nine such directions like Rare species identification, Invasive alien species, Species Distribution model, Species diversity analysis, Species presence-absence analysis, Species classification, Effect of environmental factors, Flow regime for ecology, and Effect of human intervention. Figure 2.4 highlights the commonly used methods used in meeting different solutions. For example, it can be seen that the researchers have generally used linear regression and other statistical measures to identify rare species.

2.2.3 Comprehensive analysis on applications

The number of proposed works in different domains (i.e., aquatic, forest, mountain) using different computational approaches viz, artificial intelligence, machine learning, data mining, deep learning, and statistical analysis - is depicted by a graph in Figure 2.5. This figure shows that most of the research attempts to focus on aquatic biodiversity compared to forest and mountain biodiversity.

Table 2.1: Multiple approaches and their application in biodiversity data analysis

Statistical approaches	
Biodiversity indices	Shannon index, Simpson’s diversity index have been used for geographic distribution of micro-invertebrates species, tree species diversity analysis [72]
Bayesian Belief Network	Maximize the native fish outcome study of invasive alien species in the aquatic region [73], study on rare species in the forest region
Data Mining: Automated prediction and decision-making capability	
Decision tree	Classification trees, regression trees are used for classifying deciduous vegetation, spatial modeling of tree diversity [74], species presence/ absence analysis, flow regime in ecology
Clustering	Support vector machine (SVM), k-means, and hierarchical clustering are most well-known for species presence-absence analysis [75], analyzing the effect of environmental factors, study on alien invasive species
Rule system-based approach	[32] uses the database of ichthyoplankton and investigates the relationship present between the biotic and abiotic factors
Machine learning: Performing a specific task by the machine itself	
Instance-based	Self-organizing map (SOM), K-Nearest Neighbor (KNN), Discriminant Analysis (DA). [76] are used for species presence/ absence or distribution analysis
Principal Component Analysis	Used in high-dimensional data and reduces it in the simplest form for easier analysis[77]
Artificial intelligence: Enables a system to perform tasks like an intelligent agent	
Ensemble learning	genetic programming are highly used in species diversity analysis, and species distribution modeling [78]
Artificial neural networks	Perception and radial basis network (multilayered-perception) are used for species classification and prediction problems in [79]
Deep learning: Advanced neural network architecture with multiple hidden layers	
Deep convolution neural networks	Automatic semantic extraction of features from a massive volume of a large set of images [80], species classification from the data available in the form of image, audio, or video. Tree species classification from remote sensing data [81]

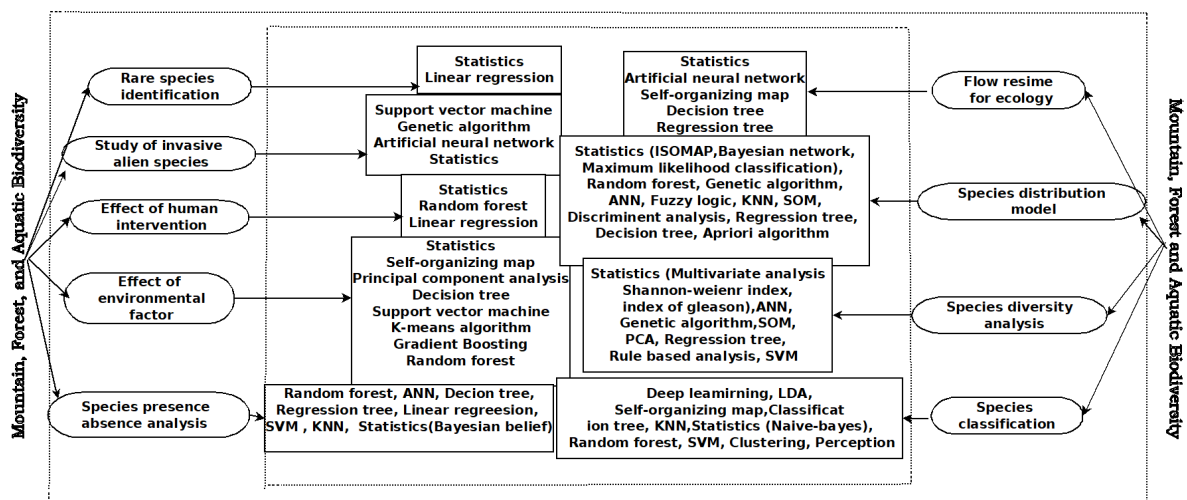


Figure 2.4: Listing of different algorithms used in different applications

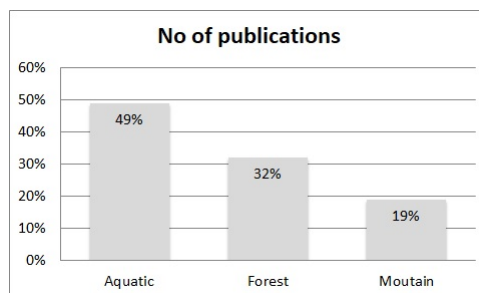


Figure 2.5: Percentage, of publications in different domains

Table 2.2: A few examples of contributions made towards biodiversity data analysis

Author	Methodology followed	Contribution	Data Used
Edwin E Herrick and group	Machine learning and data mining	Water quality, resource management, and environmental impact assessment on ecology [82]	Fish biodiversity
Peter L. M. Goethals and group	Artificial intelligence algorithms	Effect of micro-invertebrates on water quality in their presence and richness [78], establishing habitat suitability model	Micro-invertebrate datasets
Uttam K. Sarkar, and the group	Statistical analysis	Physiological biodiversity: length, weight, habitat, age, structure, growth pattern, etc. and fish stock identification, [83]	Fish species
Urs G. Kormann, and the group	Hierarchical modeling and statistical techniques	BIOFRAG database [84]: Specifically used for storing the complex results of biodiversity data in forest fragmentation studies, analyzing fragmented habitats	Forest biodiversity data
Nicolas Pasquier and group	Data mining	BioKET: [85] Biodiversity data warehouse, environmental effect on biodiversity	Plant species
Falk Huettmann and group	Data mining and machine learning	Use GMBA ¹ portal for analyzing the Himalayan plant database [86], species habitat modeling in Alaska, species distribution model	Ecological wildlife

Figure 2.6 shows the rate of use of different algorithms in aquatic biodiversity, mountain biodiversity, and forest biodiversity. In all cases, statistics and data mining are the leading methods followed by most of the researchers. The overall use of different algorithms in all domains is depicted in Figure 2.7.

2.3 Biodiversity Information Systems

Due to voluminous and heterogeneous biodiversity data, it is quite challenging to maintain a unique database that will keep records of all species close to a particular domain or region. A few initiatives by the government and others as well have already been made (Table 2.3).

Table 2.3: Digitalized biodiversity: Multiple initiatives for portal development

Portal	Contribution	Focused data
GFBI ²	Managing world's forest inventory database, policy-making for forest science, platforming forest study and research	Tree level forest inventory data and services
GBIF ³	Global network for providing research infrastructure and openly accessible biodiversity data.	Data on all types of live on earth.
Fishbase ⁴	Global database for species. It is an analytical and graphical tool for identifying, managing and restoring depleted fish stock and fish species data	
iDigBio ⁵	Aiming at digitization of biodiversity collection	Specimen data
BioGeo-Mancer ⁶	Maximize the quality and quantity of biodiversity data by integrating it with geospatial data. Thus support planning, conservation, and management in biodiversity data	Biodiversity data along with geographical location

Besides the highlighted portals (Table 2.3), a few more studies are there, e.g. [87] has worked on Western Ghat ecology, the biodiversity information system [88] on rare species is built on the European dataset. There may be scope for generating a unique interactive database having information regarding synonyms, common name, habitat, ecological status, distribution, habit, identification, description of different species of a particular region

²Global Forest Biodiversity Initiative: <https://www.gfbinitiative.org>

³Global Biodiversity Information Facility: <https://www.gbif.org/>

⁴<https://www.worldfishcenter.org/fishbase>

⁵Integrated Digitized Biocollections: <https://www.idigbio.org/portal>

⁶BG: <https://sites.google.com/site/biogeomancerworkbench/>

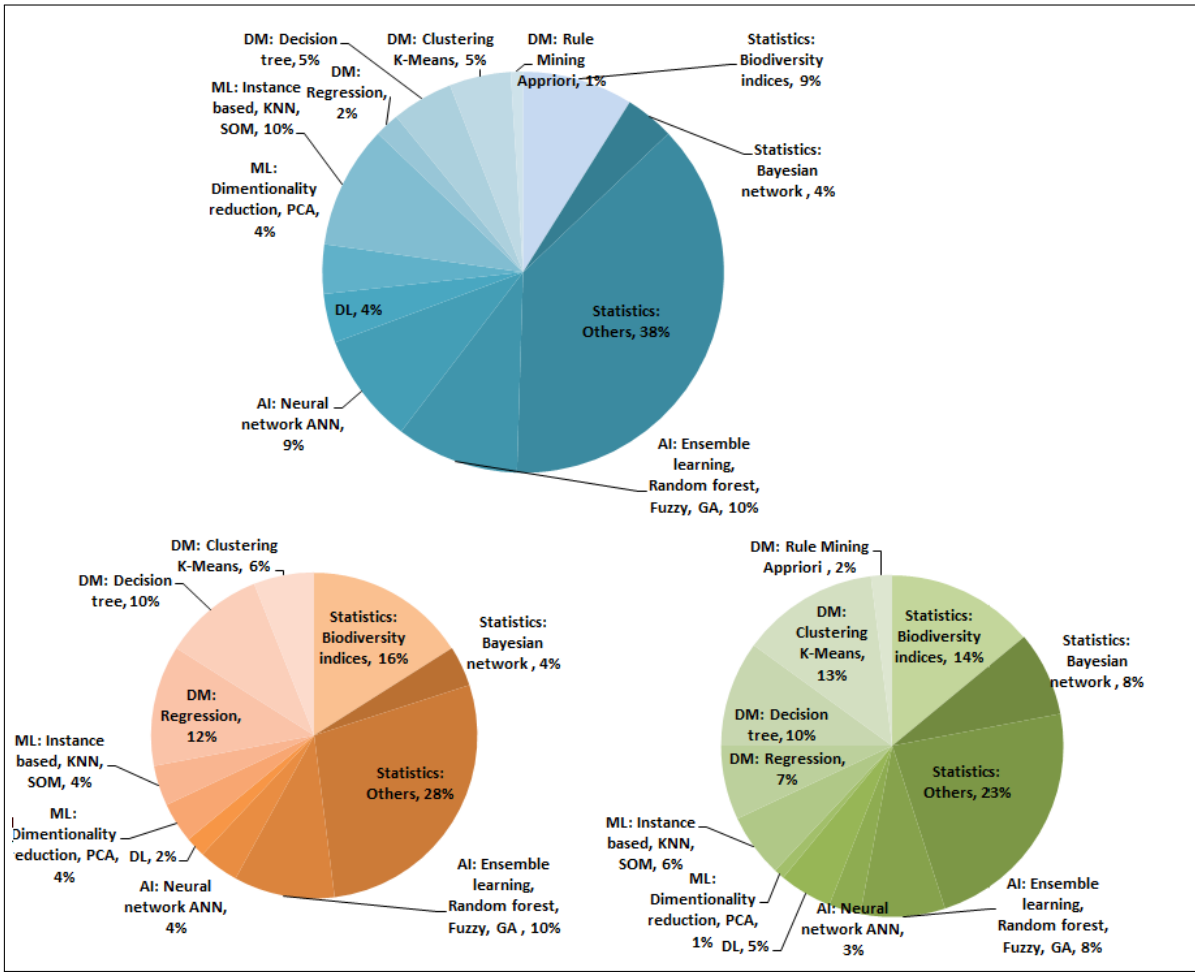


Figure 2.6: Comparative use of different algorithms in aquatic biodiversity (Top), mountain biodiversity (Bottom Left), and forest biodiversity (Bottom Right)

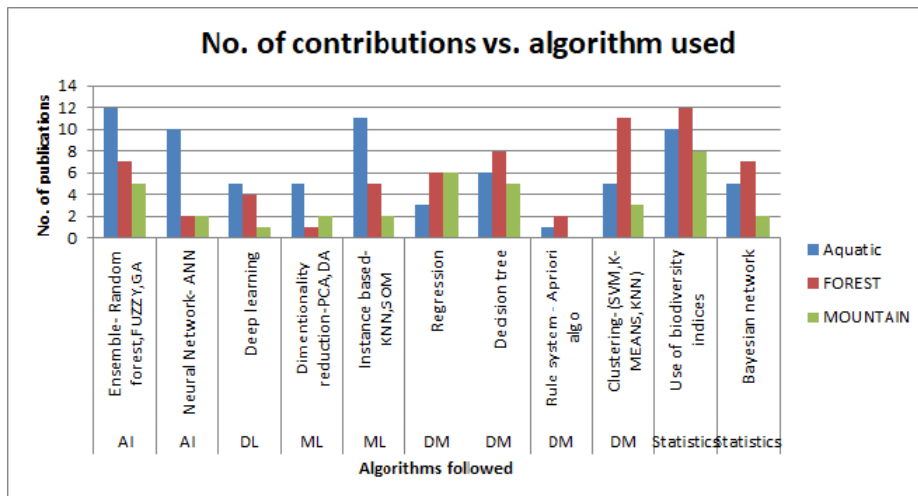


Figure 2.7: Percentage of the use of the algorithms

along the time dimension. It may find the reason for extinction or migration, etc. It also may help in knowing where some specific species arise. How many of them survive? How are they spreading? Such kind of biodiversity database has a high prospect of making the decision for the conservation purpose only when computational techniques will be incorporated at the back end to deal with the voluminous data.

2.4 Summary

This section provides a brief review of the computational approaches attempted in the biodiversity domain which is unavailable in the state-of-the-art. It would be helpful for the research community as a brief integrated scenario on approach and application is emphasized here. It has been noticed that most ecologists use statistical analytical tools where hypothesis tests have been performed to find relationships among the predictor and response variables. But, with the help of computer-science-based methodologies, exploratory analysis is possible instead of confirmatory analysis. Thus, exploring the data helps in building more accurate models in order to assist in future research. In the future, system modeling could be attempted using a computational framework for building holistic solutions for complex environmental and ecological issues, even incorporating big data. Henceforth, automation in a built-in the model would assist the ecologists in finding feasible solutions with minimum human intervention.

CHAPTER **3**

RELATED TERMINOLOGY

3.1	Technological Background	24
-----	------------------------------------	----

This chapter contains all the essential terminologies that have been used throughout the thesis.

3.1 Technological Background

1. **Frequent itemset (FI):** Say, a transaction database TDB consists of n number of rows, indicating a number of transactions. A set of items that appeared in an individual transaction is referred to as an itemset IS. So, TDB can be depicted as, $TDB \rightarrow \sum_{i=1}^n \langle T_i, IS_i \rangle$. T_i corresponds to a unique transaction id, and each IS_i can be viewed as $IS_i \rightarrow \sum_{j=1}^m item_j$. If the total number of unique items is k, then m must be $\leq k$. Now, support of an Itemset IS is the number of times it appears in TDB. If the support of IS achieves minimum user-defined threshold t, then we mention the itemset IS as a frequent itemset (FI). A frequent itemset is also known as a frequent pattern (FP) where patterns could be itemsets, subsequences, or substructures.

Now, frequent sequential itemset (FSI) refers to each item $\in FI_i$, is maintaining a specific sequence of appearance.

2. **Frequent closed itemsets (FCI):** A frequent itemset FI is considered to be a frequent closed itemset (FCI), if there exists no superset S of FI such that supporting objects of S is equal to the supporting objects of FI.
3. **Association rule mining:** Let $I = \{i_1, i_2, \dots, i_n\}$ denotes a set of items, $T = \{t_1, t_2, \dots, t_m\}$ denotes a set of transactions in the database. Each transaction consists of a subset of items from I. An association rule is in the form of $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. Here X is the antecedent and Y is the consequent. In the domain of market basket analysis, the rule $X \Rightarrow Y$ states if a person buys itemset X, then he is also likely to buy itemset Y. Hence, in other words, it can be said that consequent Y is an itemset that could be found in combination with itemset X.

The standard measures for assessing association rules are the **support** and **confidence** of a rule [33]. Two more measures, **lift** and **chi-squared** analysis are considered for co-relation analysis on association rule mining over the support-confidence framework.

Support quantifies how frequently an itemset appears in a dataset. So, mathematically, support of a rule ($X \Rightarrow Y$) is the probability of co-occurrence of the itemsets X and Y. Support can be measured as follows (Equation 3.1):

$$\text{Support of } X \Rightarrow Y = P(X \cup Y), \text{ and range : } [0, 1] \quad (3.1)$$

Confidence measures the conditional probability of occurrence of the itemset Y in the transaction dataset, given that it also contains X. It can be calculated as follows (Equation 3.2):

Residual probability	0.200	0.100	0.075	0.050	0.025	0.010	0.005	0.001	0.0005
min chi-squared	1.642	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116

Table 3.1: chiSquared distribution when the degree of freedom is 1

$$\begin{aligned}
 \text{Confidence of } X \Rightarrow Y &= \frac{\text{Support of } (X \Rightarrow Y)}{\text{Support of } (X)} \\
 &= \frac{P(X \cup Y)}{P(X)}, \text{ and range : } [0, 1]
 \end{aligned} \tag{3.2}$$

A rule with greater Support and confidence indicates stronger associations.

The generated rules can be classified mainly into two groups (exact and approximate associations). The exact association rule has a confidence value of 1 assuring the occurrence of the consequent part of the rule whenever the antecedent part is found to occur in the dataset. The approximate association rule has a confidence value that is less than 1. It leverages the key benefits of rule mining by identifying the probability of occurrence of the consequent part when the antecedent has already occurred. Hence, these approximate rules are useful in dealing with predictions. Using approximate rules, cases can be found to occur so that they meet the confidence value of 1. The approximate rules can be classified as the proper base and structural base. The structural base can be understood as a brief collection of the most informative and useful set of approximate rules. Whereas the proper base can be viewed as an abstract of all approximate rules that exist and are helpful in case the set of the structural base is large.

Lift is measured by $P(A \cap B) / P(A) * P(B)$ where the denominator says that A and B occur independently, they have no association, and is expressed by the multiplication of their probability value. The numerator says that they co-occur and is expressed by their probability of the intersection of A and B. Hence, a lift value equal to 1, or closer to 1 says that events A and B are independent. Whereas greater than one value (higher numerator) indicates that they have a positive correlation, and less than 1 value (higher denominator) identifies a negative correlation. Hence, the lift has the decision-making capability on the association between the antecedent and the consequent. Thus the rules with a 1/ closer value of 1 could be screened out. A histogram plot on the lift values could help in deciding the absolute threshold for the lift value. An example of such usage is found in [89] where filtering is done on association rules to be further analyzed. A lift value greater than 1 specifies that the occurrences between the two are dependent and suggests a strong co-occurrence relationship between A and B.

Chi-squared analysis prunes the long result set obtained by the rule mining algorithm and, hence, it keeps only the significant rules. The statistical significance of a rule identifies the level of dependence between the antecedent and the consequent. Chi-square value can easily be computed from the support, confidence, and lift value of a rule [90]. As we are concentrating on binary-valued attributes, the number of degrees of freedom will be 1. The chi-square distribution with 1 degree of freedom [91] is shown in Table 3.1 for the selected significance levels. It shows that a chi-square

value of 10 possesses a significance level better than 0.005. Here the chi-squared value of 10 can be interpreted and the underlying variables are actually independent where the residual probability is less than 0.001.

It is considered that a p-value less than 0.05 is statistically significant and less than 0.001 is statistically highly significant.

4. **Biclustering:** Biclustering [36] identifies a group of rows with a similar or coherent expression pattern under a specific subset of column values. Say, a data matrix D (M , N) is having M number of rows and N number of columns. We may rewrite this as, a set of all the rows $R = r_1, r_2, \dots, r_M$ and set of all the columns $C = c_1, c_2, \dots, c_N$. Any element of this data matrix is represented by d_{ij} where $1 \leq i \leq M$ and $1 \leq j \leq N$. Say, $D1(X,Y)$ is a submatrix of $D(M, N)$ where $X \subseteq M$ and $Y \subseteq N$. Bicluster may be formed based on constant values of each cell of the submatrix i.e. $\forall_{1 \leq i \leq X}$ and $\forall_{1 \leq j \leq Y}, d1_{ij} = \mu$ and μ is a constant value. Bicluster may also be formed with constant rows or constant columns, i.e. if $D2(x,y)$ is a submatrix where the values of this submatrix can be represented by $d2_{ij} = \mu + \alpha_i$ or $\mu * \alpha_i$ for constant rows and $d2_{ij} = \mu + \beta_j$ or $\mu * \beta_j$ for constant columns. Here, α_i and β_j are the adjustment factors for rows and columns respectively. It may happen that both row and column adjustment factors are contributing to create a bicluster at the same time which is called coherent bicluster and is of the form of $d3_{ij} = \mu + \alpha_i + \beta_j$ or $d3_{ij} = \mu * \alpha_i * \beta_j$.
5. FIST (A data mining tool): FIST [92] has been chosen primarily as it is an integrated approach of frequent closed itemset mining and rule generation.

Consider a dataset that has n number of items. Each of the items can have up to m number of attributes. A frequent closed itemset generated from that dataset is the set of a maximum number of items having the same set of attributes. Therefore, it could generate the non-redundant information list for a cluster of items that meet the cluster of attributes. Later, using the set of frequent closed itemsets, deduction of all the non-redundant set of association rules [93] is possible.

Both numerical and categorical datasets can be employed in FIST. It requires the input dataset along with minimum support and minimum confidence values to generate the output.

The output set contains all the frequent closed itemsets, exact association rules, and approximate association rules.

6. Formal context and formal concept: For every binary relation, a complete lattice can be formed [94] and this establishes the basis for the formal concept analysis. Let us consider a 2-dimensional matrix where the rows are represented by the set of objects M , and the columns are represented by the set of attributes N where $M = \{m_1, m_2, m_3, \dots, m_i\}$, and $N = \{n_1, n_2, n_3, \dots, n_j\}$ where i is the number of rows and j is the number of columns.

Let $R \subseteq M \times N$, where R is a set of symbols, then, for any pair (P, Q) , satisfying $P \subseteq M$ and $Q \subseteq N$, $P' = Q$ and $Q' = P$, is called a formal concept with respect to the formal context (M, N, R) , i.e. a formal context is a triplet where there is a set of objects, attributes, and a relation. They form a complete lattice named as concept lattice of

(M, N, R) . A formal concept basically reflects that an object $m \in M$ encodes an attribute value $r \in R$ for an attribute $n \in N$.

7. Pattern structure: It is a combination of a set of objects with their descriptions, where there is a semi-lattice among the descriptions with a similarity operation, and a mapping from the objects to the descriptions. Below, we describe a pattern structure in definition 17 denoted by $(M, (V, \sqcap), \delta)$.
8. Constant signed bicluster: An extracted constant-signed bicluster represents a subset of objects exhibiting similar signed values for a subset of attributes. If a subset of objects is denoted as $A \in M$, for any attribute $n \in N$, $n(A)$ denotes the column sub-matrix. In case of a constant-signed bicluster, for i^{th} attribute and j^{th} attribute, $n_i(A) = n_j(A)$, and they would form a bicluster having an identical sign for all the elements.
9. Coherent signed bicluster: Considering the dataset (M, N) , a coherent signed bicluster can be represented by (P, Q) where $(P \subseteq M)$ and $(Q \subseteq N)$. Now, if the bicluster is column coherent, then $\forall q_i, q_j \in Q, q_i(P)$ would be column coherent to $q_j(P)$. If the bicluster is row coherent, then $\forall p_i, p_j \in P, p_i(Q)$ would be row coherent to $p_j(Q)$.
10. Signed attribute: Let the set of attributes is denoted by N . $n \in N$, be an attribute, and $*$ $\in \{-1, 0, \sim, +, +1\}$, be a sign. So, n^* would be called a signed attribute having sign $*$. For example, $n1^+$ could be denoted as a signed attribute where the $+$ sign is assigned to $n1$.
11. Signed partition component: Let s be a signed partition component and s be a subset of N . Each attribute in s possesses the corresponding sign $*$. Therefore, the signed partition component or sp-component s can be represented as, $s = (n1^*, n3^*, \dots, nn^*)$.
12. Constant sp-component: All the attributes within s exhibit the same symbol. For example, $s1 = (n1^+, n3^+, n5^+)$ is a constant-sp component.
13. Coherent sp-component: It contains signed attributes where the signs exhibit coherency among themselves.
14. Equality of two sp-component: A sp-component contains attributes along with signs. Therefore, the equality of two sp-components can be recognized if they have an identical set of attributes with the same sign associated with them (constant signed cluster) or the same set of attributes coherently signed (coherent signed cluster). If $s1 = (n1^+, n3^+, n5^+)$ and $s2 = (n1^+, n3^+, n5^+)$, then only we could say that $s1$ and $s2$ are equal and they are constant signed bicluster. Types of coherency are discussed later. As coherency would be domain-specific, we illustrate it using an example in our case.
15. Signed partition: A signed partition P is formed by a collection of sp-components, i.e. $P = \sum_{i=1}^n s$. Thus P is the set of signed partitions where every attribute in N must be covered and should be present in exactly one component. Say, an object $m_j = (n1^+, n2^+, n3^+, n4^-, n5^+)$, a valid signed partition could be $\{\{n1^+, n2^+, n3^+\}, \{n4^-, n5^+\}\}$.
Let V represent the set of all signed partitions. Now, we have to create a signed partition mapping, $\delta : M \rightarrow V$, for assigning an object to a signed partition over N . Let us

say, that for any object m , $\delta(m)$ represents a signed partition. It may have only one sp-component. This sp-component should cover all attributes in N for that particular object m , where $m \in M$. For example, $\delta(m) = (n1^+, n2^+, n3^+, n4^-, n5^+)$; if in our addressed dataset ($M \times N$), a specific row holds $(n1^+, n2^+, n3^+, n4^-, n5^+)$ entry for object m .

16. Signed partition space: The relation between any two sp-components can extract the biclusters. To illustrate the concept of extracting bicluster, we need the notation $n(s)$ that denotes the sign for an attribute n in a sp-component s . For, $s = (n1^+, n2^+, n3^+, n4^-, n5^+)$; $n5(s) = +$. With this notion, the similarity between two sp-components can be specified with \cap .

$s1 \cap^* s2 = \{n_j^* \in s1 | n_j(s1) = n_j(s2)\}$ representing constant signed bicluster; and $s1 \cap^{\diamond} s2 = \{n_j^{*\diamond} \text{ where } n_j^* \in s1 \text{ and } n_j^{\diamond} \in s2 \text{ and } n_j(s1) \text{ coherent to } n_j(s2)\}$ representing coherent signed bicluster;

17. Signed partition pattern structure: For our considered dataset ($M \times N$), the lattice of signed partitions of N is (V, \sqcap) , where $\delta : M \rightarrow V$ corresponds to an object mapped to a signed partition. A signed partition pattern structure can be denoted by $(M, (V, \sqcap), \delta)$ where (A, B) is a signed partition pattern concept. Here, $A \subseteq M$, and $B \in V$; a signed bicluster can be formed from a signed partition pattern concept. For the pattern concept (A, B) , a signed bicluster would be (A, b) where $b \in B$.

18. Triclustering: Clustering is a widely accepted data, mining approach that performs grouping of related data but it considers only one dimension at a time. In Figure 3.1-A, we have a cluster where, in two-dimensional space, a subset of Y-axis objects is creating a subspace, keeping all the objects from the X-axis. We may term this a row-major cluster (C1, C2). Similarly, in Figure 3.1-B, a subset of objects from the X-axis and all the objects from the Y-axis may form a column-major cluster (C3, C4). [36] has proposed bi-clustering, a subspace clustering method. It considers both the dimensions of a 2-D matrix where a subset of rows has a coherent value under a subset of columns. Thus, clustering along both row and column generates a bi-cluster as shown in Figure 3.1-C (C5, C6, and C7). In the context of 3-dimensional data, a tri-clustering algorithm is proposed in [95], where a homogenous group of subspaces along three dimensions is extracted. If we consider multiple row-major clusters sharing the same subset of objects along the Y-axis, and all objects from the X-axis and Z-axis, that could give us a tri-cluster based on the row-major cluster as illustrated in 3.1-D (C8). Similarly, Figure 3.1-E (C9) gives a tri-cluster based on the column-major cluster. Figure 3.1-F (C10) shows the formation of a tri-cluster based on multiple copies of bi-clusters. Figure 3.1-G (C11) presents an irregular tri-cluster where each copy of the bicluster is not homogeneous with the others.

19. Cellular Learning Automata (CLA): CLA is typically used for systems with simple components whose behavior is defined and modified depending on that of their neighbors as well as previous states[96, 97]. A CLA model consists of learning automata (LA) and cellular automata (CA) and interacts with the environment. Figure 3.2 shows the relationship between learning automata and the environment. Each of

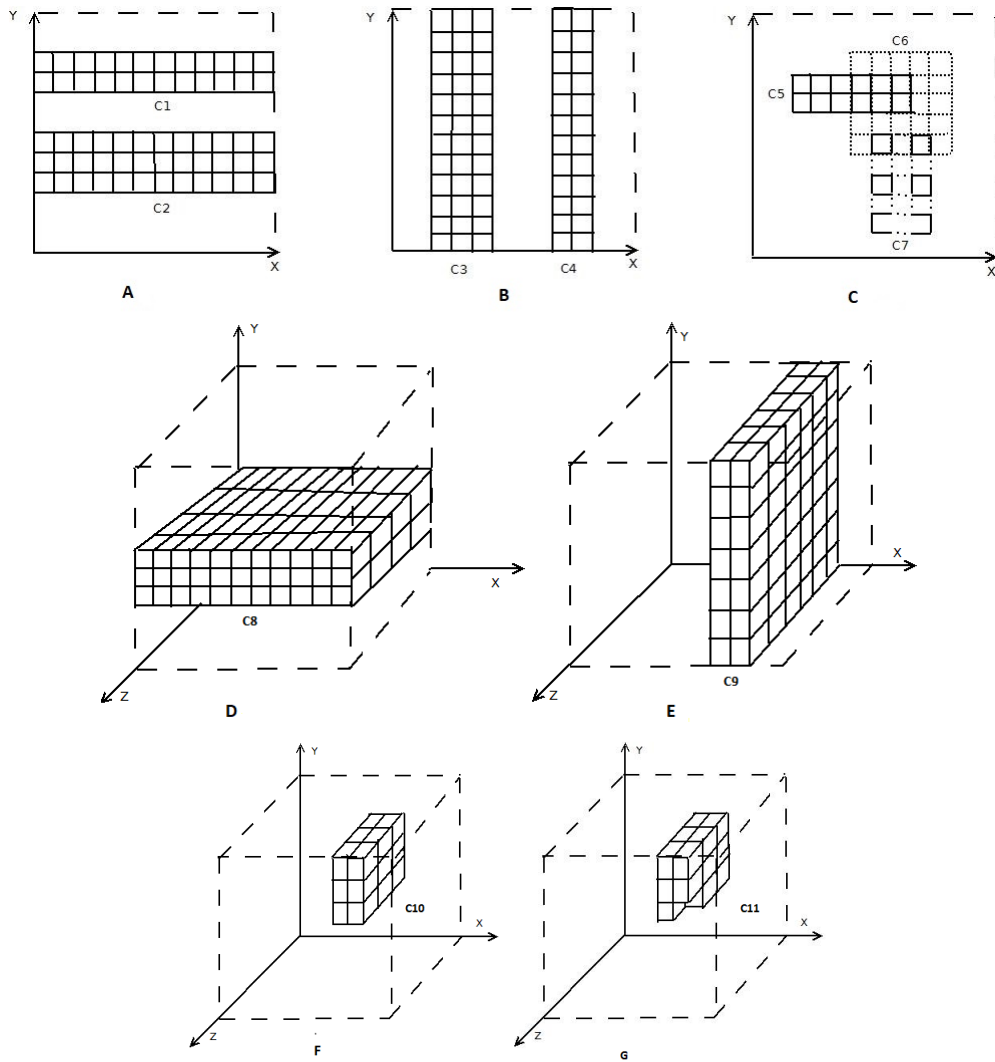


Figure 3.1: Concepts of clustering in 2D and 3D space where A, B: Clustering; C: Bi-clustering; D, E, F: Regular Tri-clustering, G: Irregular Triclustering

the components in a CLA model has a learning capability and the capability to act together. Learning capability is assured by the learning automaton, which is inhibited by each cell.

LA is an abstract model that can execute a finite set of actions.

Based upon the received response from the environment, the automata performs some action, updates its internal state and the process continues until no new response is sent to the automata. CA is a discrete mathematical model. It is called cellular as it consists of an array of cells. The Automata term is added simply because it follows a rule. The new state of a cell at time t depends on the previous states of a set of neighbor cells and the cell itself. So, CLA is more skillful than CA, as it has the power to learn, and more skillful than LA, as it is a collection of interactive learning automata. Figure 3.2 shows the relationship between learning automata and the environment. An environment may be taken as a combination of $\alpha(n)$ and $\beta(n)$ where $\alpha(n) \Rightarrow \alpha_i, \forall i = 1$ to n represents a set of actions and $\beta(n) \Rightarrow \beta_i, \forall i = 1$ to n represents a set of responses corresponding to the actions.

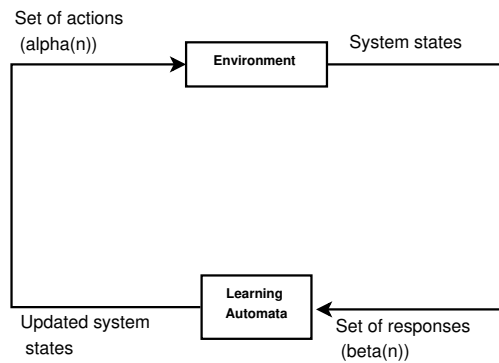


Figure 3.2: Interaction between Environment and Learning automata

Every new input will yield either a reward or a penalty depending on the learning rules in the CLA and neighbors' modes. Reward or penalty modifies the CLA's structure accordingly. In Cellular Automata, the following three modes are taken into account:

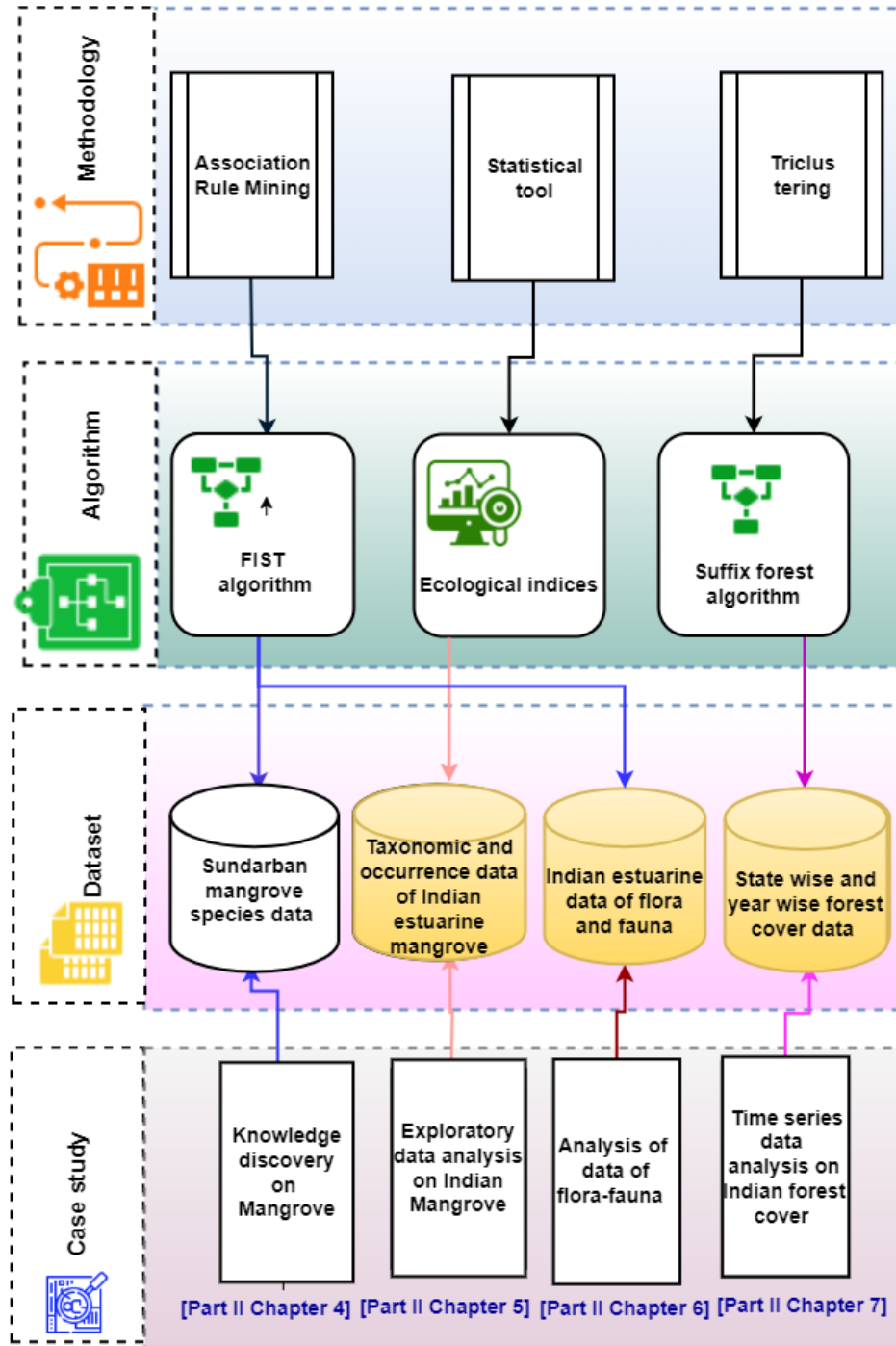
In Linear Reward Penalty, the rewards and penalties are both equal in size. In Linear Reward Epsilon Penalty, the reward is multiplied by the penalty. Linear Reward Inaction causes rewards to be granted with no penalties.

The cellular automata can be of regular or irregular depending on their characteristics. A regular CLA is described as a rectangular grid of cells and it has a finite number of possible states. Irregular CLA (ICLA) [98] is not restricted to a rectangular grid structure of CLA. It may follow a graph or tree data structure. We have used the tree data structure in our proposed methodology. Additionally, a CLA is referred to as uniform if all of its cells share the same neighborhood function, rules, and learning. The term non-uniform is used in its absence.

Part II

Case Study

The methodology, algorithm, and dataset utilized in each case study described in the chapters is shown below:



- Reusing existing algorithm / dataset
- Newly proposed framework/ algorithm / dataset

Figure 3.3: Outline for Part II

KNOWLEDGE DISCOVERY ON SUNDARBAN MANGROVE

4.1	Introduction	35
4.2	Database specification	35
4.3	Database utility	36
4.4	Database analysis	38
	4.4.1 Dataset 1	38
	4.4.2 Dataset 2	42
4.5	Result and discussion	42
4.6	Summary	47

4.1 Introduction

The mangrove ecosystem is continuously losing its dignity. A few studies have focused on understanding the changing behavior of Sundarban Mangrove Forest. But, there is a lack of database interpretation and useful pattern extraction, that could be more useful for standing against the degrading nature of the mangrove ecosystem. Understanding the present scenario, the main contribution of this study lies in the information retrieval task by assessing the natural growth of native mangrove species of Sundarban. This study generates rules showing the effect of soil pH, and water salinity on mangrove community structure, and individual mangrove species and finds a relation to biodiversity indices. This gives assistance towards the restoration of the mangrove ecosystem in terms of predicting probable occurrences.

Realizing the importance of analyzing the growing loss of the mangrove ecosystem [99, 100, 101] and the reasons behind it, we keep the focus on the advantageous use of data mining techniques over statistical analysis techniques [32]. The statistical techniques (e.g. study of diversity indices [102], community characteristics study [22]) are only confirmatory analysis techniques concerning the researchers' understanding. Whereas, the data mining approach is a tool for exploratory analysis techniques, that manages huge complex data in a programming environment and is responsible for data manipulation, querying, and visualization (used in [103, 32]).

Coming to the existing literature related to Sundarban Mangrove, in [21], a research team has found that climate change affects water salinity vis a vis mangrove species occurrence. The changes in salinity distribution due to the combined effect of sea-level rise and climate change along the coastal region of Sundarban Mangrove Forest have been addressed in [104]. It has been found that the present concentration is on quantifying information rather than revealing decision-making knowledge.

Hence, we are willing to use a data mining knowledge discovery procedure, and the application is made on the dataset provided in [105]. In [105], the authors use ordination or gradient analysis, data clustering, and the Shannon-Wiener diversity index, in a study of plant species diversity in Bangladesh mangrove forest. A total of 49 plant species that grow on forest floors, including small trees and shrubs are considered to have different functionalities [54].

4.2 Database specification

The datasets [105] we used to describe 29 sites in terms of different parameter values like pH, salinity, Shannon Diversity Index (SDI), N_2 Diversity, and traces 49 undergrowth species belonging to those sites. The datasets are given in Table 4.1 and Table 4.2.

Salinity itself plays an important role in the physiological characteristics of mangrove and their abundance [105]. Low soil pH lowers the amount of phosphorous and other nutrients. On the other hand, high pH acts as a barrier for plant essential nutrients as it hinders iron, zinc, manganese, etc. from being obtainable. Thus pH has a significant effect

Table 4.1: Dataset 1 obtained from [105] showing the details of 29 sites

S. No	Sites	Mangrove Community Type	pH	Salinity ppt.	SDI	N ₂ Div
1	JongraBeel	<i>Heritiera-Xylocarpus-Bruguiera</i>	6.8	4.6	2.59	12.35
2	Mirgamaria	<i>Heritiera-Bruguiera-Xylocarpus-Avicennia</i>	8.5	5	2.53	10.97
3	Sharonkhola Panirghat	<i>Excoecaria-Heritiera</i>	7.6	0	2	6.37
4	Shorankhola : South of Dhabribarani	<i>Excoecaria-Heritiera</i>	7.7	0	2.33	8.68
5	Sharonkhola Terabeck-aKhal	<i>Heritiera-Excoecaria</i>	7.5	0	2.15	6.95
6	Kotka Range Office	<i>Heritiera-Excoecaria-Sonneratia</i>	7.7	1.3	2.42	8.75
7	Kotka North Jamtala	<i>Sonneratia-Heritiera-Excoecaria</i>	7.6	1	2.19	7.07
8	Kotka South Jamtala	<i>Sonneratia-Heritiera-Excoecaria</i>	7.7	1.5	2.74	13.55
9	Deemyer char	<i>Sonneratia-Excoecaria-Heritiera</i>	7.9	3	3.09	19.51
10	Dhanshiddher Char	<i>Heritiera-Xylocarpus-Bruguiera</i>	7.6	17	1.9	5.53
11	KNM collection Centre	<i>Heritiera-Excoecaria</i>	7.5	16.5	1.41	3.68
12	Kewrabunia Char	<i>Sonneratia-Heritiera-Excoecaria</i>	7.5	16.5	7.98	5.64
13	KalagachiaDanokhal	<i>Excoecaria-Ceriops-Xylocarpus</i>	6.9	20.5	2	5.41
14	Patakata	<i>Heritiera-Excoecaria-Ceriops</i>	7.8	10	1.84	6.01
15	Tiarchar	<i>Sonneratia-Excoecaria-Ceriops</i>	7.6	15	1.97	6.33
16	Pakhirchar	<i>Sonneratia</i>	7.8	7	2.3	8.56
17	Dublar Char	<i>Excoecaria-Sonneratia</i>	7.5	20.5	2.51	11.63
18	Mandarbaria	<i>Excoecaria-Ceriops</i>	6.8	20	2.1	7.47
19	Kalir Char (north)	<i>Ceriops-Excoecaria-Sonneratia</i>	5.9	20.2	1.63	3.92
20	PusphaKathi	<i>Ceriops-Excoecaria</i>	5.1	22	1.95	6.2
21	Koikhali	<i>Bruguiera-Heritiera-Xylocarpus</i>	6.8	20.5	1.74	5.45
22	Kalogachia	<i>Excoecaria-Heritiera-Xylocarpus-Avicennia</i>	7	14.5	1.57	4.33
23	Kochikhali	<i>Heritiera-Sonneratia-Excoecaria</i>	7.7	4.5	3.11	18.89
24	Karamjal	<i>Heritiera-Sonneratia-Ceriops-Nypa</i>	7.9	4.4	2.86	15.33
25	Hoddo	<i>Bruguiera-Heritiera-Sonneratia</i>	7.6	16	2.46	10.79
26	Andharmanik	<i>Ceriops-Excoecaria-Avecennia-Xylocarpus</i>	7.7	18	2.36	9.65
27	Kobadak	<i>Sonneratia-Xylocarpus-Excoecaria</i>	7.5	21	1.65	4.78
28	Dobeki	<i>Sonneratia-Ecoecaria-Avecennia</i>	7.5	22	2.51	11.63
29	Notabeki	<i>Ceriops-Excoecaria-Xylocarpus</i>	7.4	23	1.41	3.68

on species composition. Simpson's dominance index is converted to diversity statistic by Hill [106] and is termed as N_2 diversity or N_2 Div [107]. Its usefulness lies behind its aspect of enumerating community diversity as it weighs all species equally, independent of their abundance. More weight is given to the abundance of common species and less weight to the rare species. Thus, it considers the negligible influence on the addition/ deletion of rare species. The higher the value of N_2 diversity, the greater the diversity is. Besides the dominance index, the information-statistic index is also considered which includes Shannon Diversity Index (SDI). SDI gives a measure of both species richness and evenness. A high value of an SDI would be a representation of a diverse and equally distributed community. A low value represents a less diverse community.

4.3 Database utility

These two datasets, together, give useful facts and can generate rules like which parameters are more influencing for the addressed population. This can also be treated as a base

KNOWLEDGE DISCOVERY ON SUNDARBAN MANGROVE

Table 4.2: Dataset 2 obtained from [105] showing species presence in 29 sites

Scientific name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
<i>Acanthus ilicifolius</i>	5	5	2	4	2	4	2	4	4	1	2	1	8	6	5	5	1	6	6	7		8	5	4	3	4	3	2	1
<i>Acrostichum aureum</i>	4			2	2	2	1	4	2		8	1	5					6	1	5	3		6	3	3	3	4	4	3
<i>Aegiceras corniculatum</i>								2				1	2			3	1					4	1	1					
<i>Brownlowia tersa</i>						1	1														3		3	1					
<i>Caesalpinia crista</i>								1										1		1									
<i>Clematis arborea</i>									2							1	1												
<i>Clerodendron inernermae</i>					1			2								1	1					3	1						
<i>Carallia bracheata</i>																1	1												
<i>Crotalaria saltiana</i>					1			2	3											1									
<i>Crinum asiaticum</i>									1																				
<i>Cuscuta reflexa</i>									1																				
<i>Cynometra ramiflora</i>	5	6	4	5	8		4	2				1	8									2	1	2	2				
<i>Cyperus exaltatus</i>		3		2													2		1			3	2						
<i>Cyperus japonicus</i>					1											3							2						
<i>Cyperus malaccensis</i>	2	1			1				1	1		1					1						1						
<i>Cyperus tagetiformis</i>	1	1	1		1					1							1						2						
<i>Dalbergia candellensis</i>					1											3							2						
<i>Dalbergia spinosa</i>	6								3						4		2	6											
<i>Derris trifoliata</i>	2	6	6	8	2	8	6	5	5	4	2	7	10	8	6	4	1	6	3	1	7	7	5	3	2	4	1	2	
<i>Dioscorea sp</i>									1																				
<i>Entada pursaetha</i>		1						2				1											1	1	1	1			
<i>Fimbristylis acuminata</i>	1		1			1	1								1	2		1	1				2		1				
<i>Flagelaria indica</i>	4	2																											
<i>Flueggia virosa</i>			1			4																							
<i>Hemarthria compressus</i>						1			4	1									1	1									
<i>Hibiscus tiliaceus</i>	5	5	4		4			2	4						1	1	1						1	1					
<i>Impereta cylindrica</i>					1				1	1													3						
<i>Ipomoea pescaprae</i>									1														2						
<i>Mikania cordata</i>			1	3					3						1														
<i>Myristachya wightiana</i>									3	1	2	2	1						1	1			1	2	2	3			
<i>Nypa fruticans</i>					4				3		6	1	6			2				4	4	1	2	1	2				
<i>Panicum repens</i>				1		1																	1		1				
<i>Pandanus foetidus</i>	4	4	2	2	4	2	6	2			1	1	4										1	2		2			
<i>Paspalum vaginatum</i>									2																				
<i>Phoenix paludosa</i>	4			6			1	2	1			1	6	3	1		5			6	5	1	2	1		1	4	4	5
<i>Phragmites karka</i>		1	2	1	1				1						1														
<i>Pongamia pinnata</i>		1					1		1				1				2						1	1					5
<i>Porterasia coarctata</i>								1	2			2	1										2	2	3	2		1	1
<i>Rhizophora mucronata</i>	2	1		2	1	2	1	1															1	3	1	2	2	1	
<i>Saccharum spontaneum</i>									1														1	1	1				
<i>Salacia chinensis</i>		1																					1						
<i>Sapium indicum</i>				2	1	1				2																			
<i>Sarcobolus globosus</i>				1	1		1	1	1		2	2					1	7			5	4	5	1	1	1	1	1	1
<i>Scirpus articulatus</i>									1																				
<i>Solanum xanthocarpum</i>																		2											
<i>Stenoclaena palustris</i>	7	4		7				3	3			1													1				
<i>Tamarix indica</i>	5	6						1							5	1							1	1		1		1	
<i>Typha elaphantica</i>				1		1																	1						
<i>Vitis trifolia</i>	2	2	2	5	5	1	2	2	3	3	2			4	1	1	1		2	5	9		3	1	1	2	1	1	

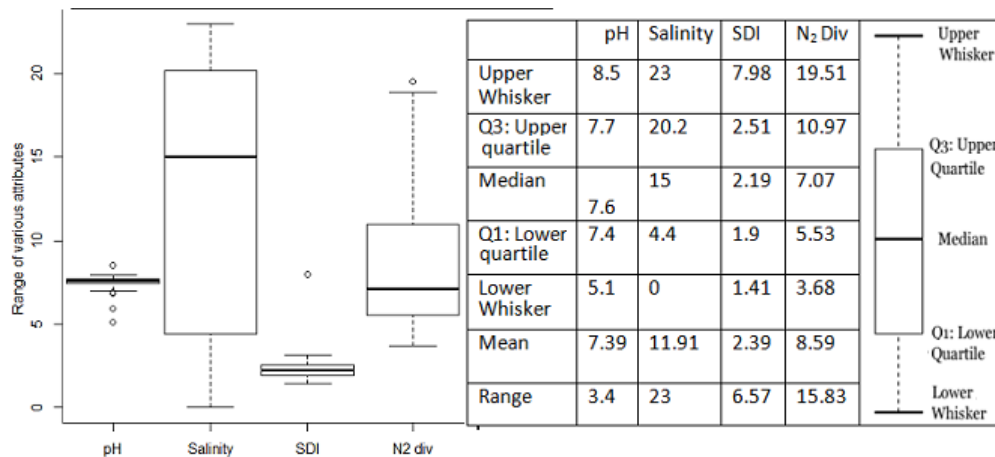


Figure 4.1: Visualization of features of attributes of dataset1

structure for the analysis of any other datasets. For this kind of utility, step by step process of knowledge discovery is followed. Such as,

Data Collection: The database specification part in section 4.2 illustrates the data source we have used.

Data Preprocessing: The exploration done by the mining algorithm is influenced by the various approaches of the preprocessing task (section 4.4).

Data Mining: In this study, an algorithm called FIST (Frequent Itemset Suffix Tree) [37] is used that executes and generates both bi-cluster and association rules in a single run on a dataset in both time and space-efficient manner. The generation of the output is based on frequent itemset generated by FIST [92].

Information Retrieval: The obtained result requires proper analysis by domain experts to extract meaningful facts. We have retrieved our facts by using FIST and done manual analysis as well on the output set. Later, significance checking is performed from the domain expert's end.

4.4 Database analysis

4.4.1 Dataset 1

The attributes of this dataset (Table 4.1) exhibit a different range of numerical values. Thus knowledge regarding the spread out of data is required before preprocessing.

Box-plot Visualization The graphical rendition of the statistical data through box plot in Figure 4.1 reveals that the attributes support a different range of values. As per Figure 4.1, the standard deviation for pH and SDI are lesser compared to salinity and N_2div as most of the values are very close to the median in the case of pH and SDI. It is clear from the picture that the median is closer to the upper quartile for salinity which indicates that the data consists of a large number of frequencies of high-valued scores. Similarly, N_2Div is having just the opposite scenario to Salinity distribution, i.e., the median is closer to the

KNOWLEDGE DISCOVERY ON SUNDARBAN MANGROVE

Table 4.3: Z-Score value corresponding to dataset 1 shown in Table 4.1.

Sl. No	pH	Salinity	SDI	N ₂ DIV
1	-0.907692308	-0.868171021	0.170940171	0.882629108
2	1.707692308	-0.820665083	0.11965812	0.558685446
3	0.323076923	-1.414489311	-0.333333333	-0.521126761
4	0.476923077	-1.414489311	-0.051282051	0.021126761
5	0.169230769	-1.414489311	-0.205128205	-0.384976526
6	0.476923077	-1.260095012	0.025641026	0.037558685
7	0.323076923	-1.295724466	-0.170940171	-0.356807512
8	0.476923077	-1.236342043	0.299145299	1.164319249
9	0.784615385	-1.058194774	0.598290598	2.563380282
10	0.323076923	0.604513064	-0.418803419	-0.718309859
11	0.169230769	0.545130641	-0.837606838	-1.15258216
12	0.169230769	0.545130641	4.777777778	-0.692488263
13	-0.753846154	1.020190024	-0.333333333	-0.746478873
14	0.630769231	-0.226840855	-0.47008547	-0.605633803
15	0.323076923	0.366983373	-0.358974359	-0.530516432
16	0.630769231	-0.583135392	-0.076923077	-0.007042254
17	0.169230769	1.020190024	0.102564103	0.713615023
18	-0.907692308	0.960807601	-0.247863248	-0.262910798
19	-2.292307692	0.98456057	-0.64957265	-1.09624413
20	-3.523076923	1.198337292	-0.376068376	-0.561032864
21	-0.907692308	1.020190024	-0.555555556	-0.737089202
22	-0.6	0.30760095	-0.700854701	-1
23	0.476923077	-0.880047506	0.615384615	2.417840376
24	0.784615385	-0.89192399	0.401709402	1.582159624
25	0.323076923	0.485748219	0.05982906	0.516431925
26	0.476923077	0.72327791	-0.025641026	0.248826291
27	0.169230769	1.079572447	-0.632478632	-0.894366197
28	0.169230769	1.198337292	0.102564103	0.713615023
29	0.015384615	1.317102138	-0.837606838	-1.15258216

lower quartile. Here, data constitutes a higher number of low-valued scores.

Z-score normalization As the attributes have different normal distributions, therefore, there is a requirement for converting all variables to a common scale. Thus zero-mean or z-score normalization technique would be appropriate here. Z-score normalization represents the distance of a raw value in terms of how many standard deviations below or above the mean, a data point is situated. A z-score value corresponding to the pH value of 5.8 can tell us where that value is located when it is compared to the average population’s mean pH value. Finally, we omit the decimal point and round off up to one significant digit after the decimal point in order to discretize the values.

It is a measure computed based on mean and Z-score normalization and represents the distance of a raw value in terms of how many standard deviations below or above the mean, a data point is situated. A z-score value corresponding to the pH value of 5.8 can tell us where that value is located when it is compared to the average population’s mean pH value. Finally, we omit the decimal point and round off up to one significant digit after the decimal point in order to discretize the values. In this regard, we can say that the z score is a standard score and we can place it on a normal distribution curve. Z-score ranges from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). A z-score value corresponding to the pH value of 5.8 can tell us where that value is located when it is compared to the average population’s mean pH value. Finally, we omit the decimal point and round off up to one significant digit after the decimal point in order

Table 4.4: Mean and Standard deviation for different attributes

	pH	Salinity	SDI	N_2DIV
Mean	7.386207	11.91379	2.389655	8.59
Standard Deviation	0.650123	8.426223	1.168731	4.263779

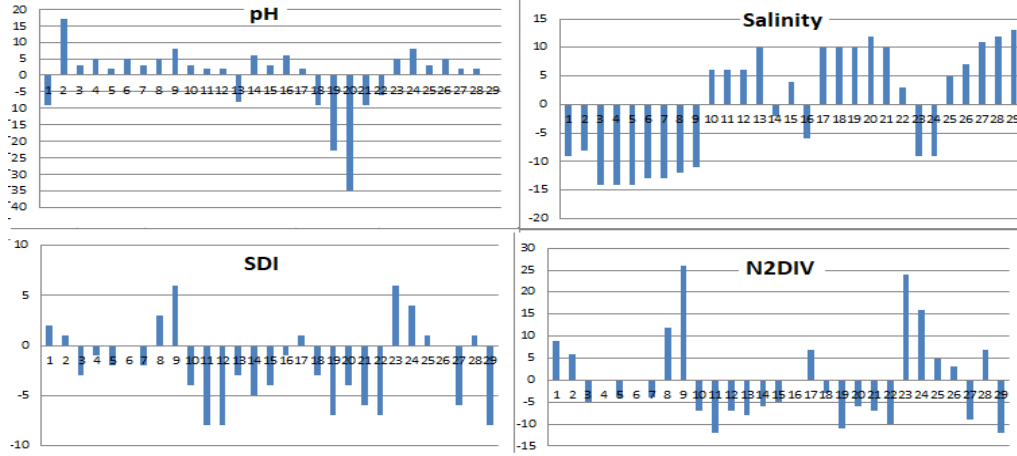


Figure 4.2: Histogram visualization, X-Axis: Site Nos.; Y-Axis: normalized attribute-values; Mapping of discretization range with label for pH, Salinity, SDI, and N_2 Div: $-20 < \text{Range} \leq -15$: Very Very Low; $-15 < \text{Range} \leq -10$: Very Low; $-10 < \text{Range} \leq -5$: Low; $-5 < \text{Range} \leq 5$: Moderate; $5 < \text{Range} \leq 10$: High; $10 < \text{Range} \leq 15$: Very High; $15 < \text{Range} \leq 20$: Very Very High

to discretize the values. Mathematically it can be explained as, say, we have a database having 5 columns A, B, C, D, E, and n number of rows. Here, we are showing the z-score calculation for column E. All other columns follow the same rule. The normalized value of e_i for column E in the i^{th} row is calculated as:

$$Normalized(e_i) = (e_i - \bar{E})/std(E) \quad (4.1)$$

where

$$std(E) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (e_i - \bar{E})^2} \quad (4.2)$$

$$and \quad \bar{E} = (1/n) \sum_{i=1}^n (e_i) \quad (4.3)$$

The standard deviation and mean for each column attribute are calculated using equation 4.2 and equation 4.3 and the values we get are shown in Table 4.4.

For Table 4.1, we calculate the z-score value for each attribute by using the equation 4.1. Table 4.3 shows the obtained z-score values of Table 4.1. Finally, we omit the decimal point and round off up to one significant digit after the decimal point in order to discretize the values as shown in Table 4.5.

Histogram Visualization: Discretization of the dataset A histogram gives a visual impression of the distribution of data. So, we divide the entire range of normalized values into buckets and create seven discrete levels identified as very very high, very high, high, moderate, low, very low, very very low. The histogram obtained for different attributes after

Table 4.5: Z-Score value by removing decimal point obtained from Table 4.3

SI No.	pH	Salinity	SDI	N ₂ DIV
1	-9	-9	2	9
2	17	-8	1	6
3	3	-14	-3	-5
4	5	-14	-1	0
5	2	-14	-2	-4
6	5	-13	0	0
7	3	-13	-2	-4
8	5	-12	3	12
9	8	-11	6	26
10	3	6	-4	-7
11	2	6	-8	-12
12	2	6	48	-7
13	-8	10	-3	-8
14	6	-2	-5	-6
15	3	4	-4	-5
16	6	-6	-1	0
17	2	10	1	7
18	-9	10	-3	-3
19	-23	10	-7	-11
20	-35	12	-4	-6
21	-9	10	-6	-7
22	-6	3	-7	-10
23	5	-9	6	24
24	8	-9	4	16
25	3	5	1	5
26	5	7	0	3
27	2	11	-6	-9
28	2	12	1	7
29	0	13	-8	-12

discretization is represented in Figure 4.2. The final dataset of the raw data after putting discrete levels on the normalized z-score value is shown in Table 4.6.

Table 4.6: Discretized textual values for dataset 1 obtained after z-score normalization and scaling

Sl. No	Discretized pH	Discretized Salinity	Discretized SDI	Discretized N ₂ DIV
1	Low	Low	Moderate	High
2	Very Very High	Low	Moderate	High
3	Moderate	Very Low	Moderate	Moderate
4	Moderate	Very Low	Moderate	Moderate
5	Moderate	Very Low	Moderate	Moderate
6	Moderate	Very Low	Moderate	Moderate
7	Moderate	Very Low	Moderate	Moderate
8	Moderate	Very Low	Moderate	Very High
9	High	Very Low	High	Very Very High
10	Moderate	High	Moderate	Low
11	Moderate	High	Low	Very Low
12	Moderate	High	Low	Low
13	Low	High	Moderate	Low
14	High	Moderate	Moderate	Low
15	Moderate	Moderate	Moderate	Moderate
16	High	Low	Moderate	Moderate
17	Moderate	High	Moderate	High
18	Low	High	Moderate	Moderate
19	Very Very Low	High	Low	Very Low
20	Very Very Low	Very High	Moderate	Low
21	Low	High	Low	Low
22	Low	Moderate	Low	Low
23	Moderate	Low	High	Very Very High
24	High	Low	Moderate	Very Very High
25	Moderate	Moderate	Moderate	Moderate
26	Moderate	High	Moderate	Moderate
27	Moderate	Very High	Low	Low
28	Moderate	Very High	Moderate	High
29	Moderate	Very High	Low	Very Low

4.4.2 Dataset 2

Data preprocessing is not required for the other dataset (Table 4.2) as this dataset is free from incompleteness and inconsistency. Each cell is either vacant representing the absence of a species for a particular site or containing frequency values from 1 to 10 where 1 corresponds to 10% frequency of presence and so on. As 1 to 10 represents discrete numerical values, no data transformation is needed here for our analysis.

4.5 Result and discussion

Figure 4.3 is presenting the complete framework for the following methodology.

This section analyses the application of association rule mining and biclustering in biodiversity data analysis is the first of its kind of analysis

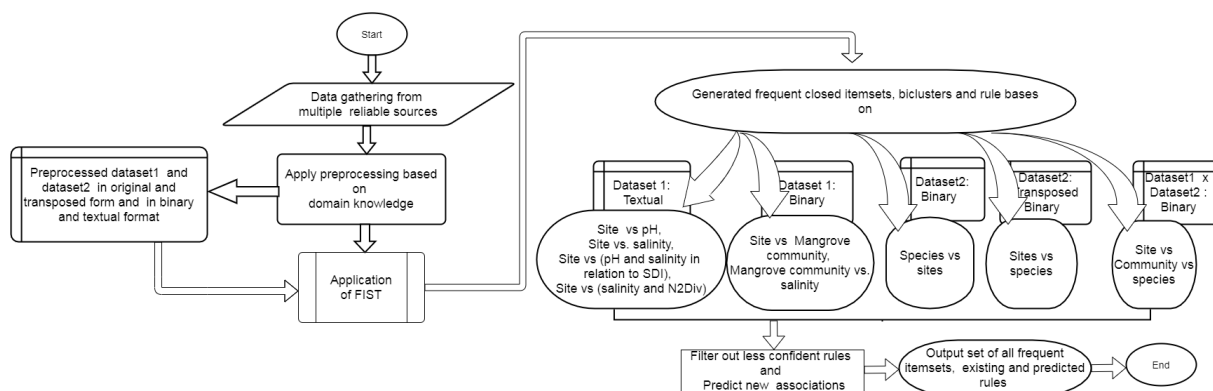


Figure 4.3: Flowchart for the followed methodology

Table 4.7: Effect of salinity on SDI found from association rules.

Antecedent (Salinity, pH)	Consequent SDI	Site list
High, Moderate	Moderate	10, 17, 26
High, Moderate	Low	11, 12
Very High, Moderate	Moderate	28
Very High, Moderate	Low	27, 29
Low, Moderate	High	23
Very Low, Moderate	Moderate	3, 4, 5, 6, 7, 8

Salinity is a more influential factor than pH for mangrove growth In [105], the authors conclude that SDI has a positive correlation with soil pH level and a negative correlation with salinity. Contradictorily, they state that a positive correlation exists among diversity indices and salinity gradient. We observe that (Table 4.7) if pH is kept at moderate, for high to very high salinity, the value of SDI varies from moderate to low. The later part of Table 4.7 shows that keeping pH at moderate and salinity at low to very low, the value of SDI varies from moderate to high. Table 4.7 shows the considerable effect of salinity on SDI compared to pH as it is kept the same for different saline conditions. Our conclusion justifies the reporting of previous study [108] that salinity has a major effect on mangrove distribution pattern and community zonation compared to the others.

Finding potential sites for plantation based on salinity condition Depending upon salinity and distribution of species composition Sundarban area can be categorized as Oligohaline, Mesohaline, and Polyhaline. Table 4.8 shows the salinity-wise frequent closed group of sites obtained from bicluster results. Similar to this, Table 4.9, is showing the frequently closed group of sites based on pH.

Table 4.8: Salinity-wise grouping of sites found from biclusters.

Salinity	Support	Site List	Zonation Pattern
Very Low	7	3, 4, 5, 6, 7, 8, 9	Oligohaline Zone
Low	5	1, 2, 16, 23, 24	Oligohaline Zone
Moderate	4	14, 15, 22, 25	Mesohaline Zone
High	9	10, 11, 12, 13, 17, 18, 19, 21, 26	Polyhaline Zone
Very High	4	20, 27, 28, 29	Polyhaline Zone

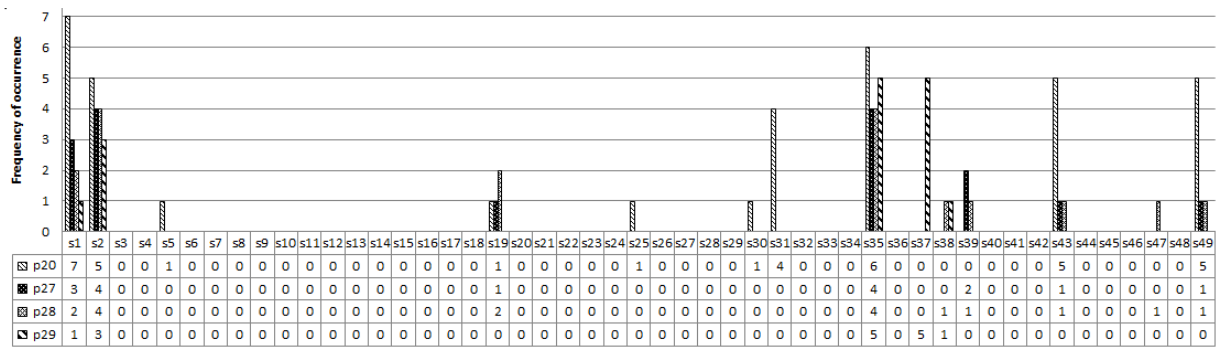


Figure 4.4: 49 plant species (s1 to s49) occurrence in 4 different sites (p20, p27, p28 and p29).

Table 4.9: pH-wise grouping of sites found from biclusters.

pH	Support	Site list
Very Very Low	2	19, 20
Low	4	1, 13, 21, 22
Moderate	16	3, 4, 5, 6, 7, 8, 10, 11, 12, 15, 17, 23, 25, 26, 27, 28, 29
High	4	14, 16, 24

Each bicluster may help in conservation planning. If, $\langle \text{Site } w, \text{Site } x, \text{Site } y, \text{Site } z \rangle$ is forming a closed group and a bicluster is obtained in the form of $\langle \text{Site } x, \text{Site } y, \text{Site } z \rangle \implies \langle \text{Species } a, \text{Species } b, \text{Species } c \rangle$, then remaining Site w should also imply the same. Similarly, the opposite scenario will also be true. Considering an example of the very high saline zone, FIST identifies sites numbered 20, 27, 28, and 29 as a closed group. As salinity plays a key role in vegetation growth, all these sites should have similarities in vegetation growth. So, if a bicluster would be like, $\langle \text{Site } 20, \text{Site } 27, \text{Site } 28, \text{Site } 29 \rangle \implies \langle \text{Species } 1, \text{Species } 2, \text{Species } 35 \rangle$, it represents that occurrence data of *Acanthus ilicifolius* (Species 1), *Acrostichum aureum* (Species 2), and *Phoenix paludosa* (Species 35) in site numbers 20, 27, 28 and 29 are almost similar (Figure 4.4).

Similarly, considering $\langle \text{Species } 19, \text{Species } 43, \text{Species } 49 \rangle \implies \langle \text{Site } 27, \text{Site } 28, \text{Site } 29 \rangle$, a close existence data regarding *Derris trifoliata* (Species 19), *Sarcolobus globosus* (Species 43), and *Vitis trifolia* (Species 49) can be found in site 27, 28 and 29. But, unlikely, in site 20, the frequency of appearance of *Derris trifoliata* (Species 19) is only 10% whereas *Sarcolobus globosus* (Species 43) and *Vitis trifolia* (Species 49), both have 50%. Thus, *Derris trifoliata* (Species 19) may be expected in a larger frequency in site 20. We get, $\langle \text{Species } 19, \text{Species } 43, \text{Species } 49 \rangle \implies \langle \text{Site } 27, \text{Site } 28, \text{Site } 20 \rangle$. So, Site 20 should have all these currently absent species. In a similar way, we can say that Sites 20, 27, and 28 should have the occurrence of *Pongamia pinnata* (Species 37) as it is surviving with a frequency value of 50% in site 29. In [105], based on gradient analysis, the dominated species names in the highest saline sites, and highest pH-containing sites are listed. In addition to this, our study reflects the potential sites of suitable salinity regimes for mangrove species occurrence.

Obtaining information related to community type Which community type has a preference for what kind of salinity zone and frequency of occurrence of each type of community is retrieved here. By expanding the Mangrove Community attribute and combining salinity details, biclusters are obtained (counted in Table 4.10).

Table 4.10: Frequent closed itemset count corresponding to each community type

Community Type	Very High Salinity	High Salinity	Moderate Salinity	Low Salinity	Very Low Salinity
<i>Ceriops</i>	2	4	3	1	
<i>Heritiera</i>		4	4	6	2
<i>Xylocarpus</i>	3	4	1	2	
<i>Avicennia</i>	1	1	1	1	
<i>Excoecaria</i>	7	8	4	1	2
<i>Sonneratia</i>	3	3	3	4	1
<i>Bruguiera</i>		1	1	2	
<i>Nypa</i>				1	

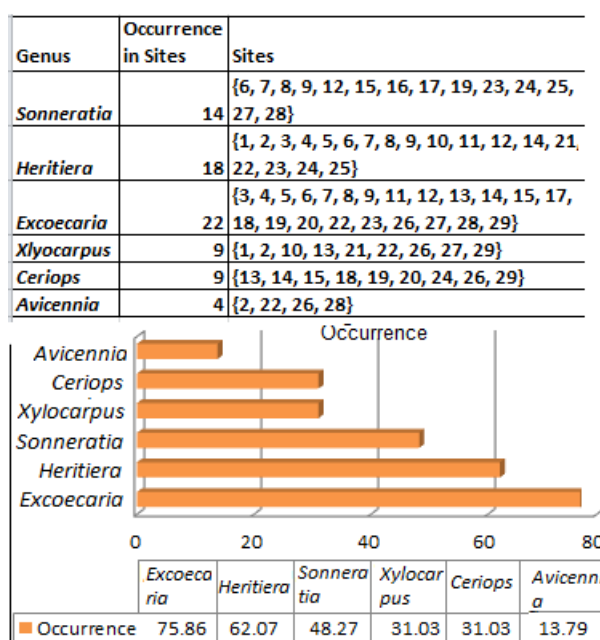


Figure 4.5: Percentage of the appearance of different community types

Knowledge acquired from the analysis of Table 4.10 are discussed below:

1. *Nypa* population is found in the less saline zone and is the least occurred type of the genus. So, afforestation of the *Nypa* population in a less saline zone may be facilitated.
2. *Excoecaria* is the most abundant genus that has adapted to a growing environment with all kinds of salinity conditions but is preferably found in an active saline zone.
3. Salinity zone-wise growth patterns of the different genera can be derived. Like, a mix-community of *Sonneratia-Excoecaria* can be found in the Oligohaline zone for 45% cases, in the Polyhaline zone for 45%, and for only 9% cases it is found in Meshohaline zone (Figure 4.6).
4. *Excoecaria*, *Heritiera*, and *Sonneratia* are three mostly found community types and are found in 76%, 62% and 48% respectively of sites (Figure 4.5) obtained from biclusters of community type data.
5. The combination of *Heritiera-Excoecaria* (41%), *Sonneratia-Excoecaria* (38%) are the most common types of community mix (Figure 4.6).

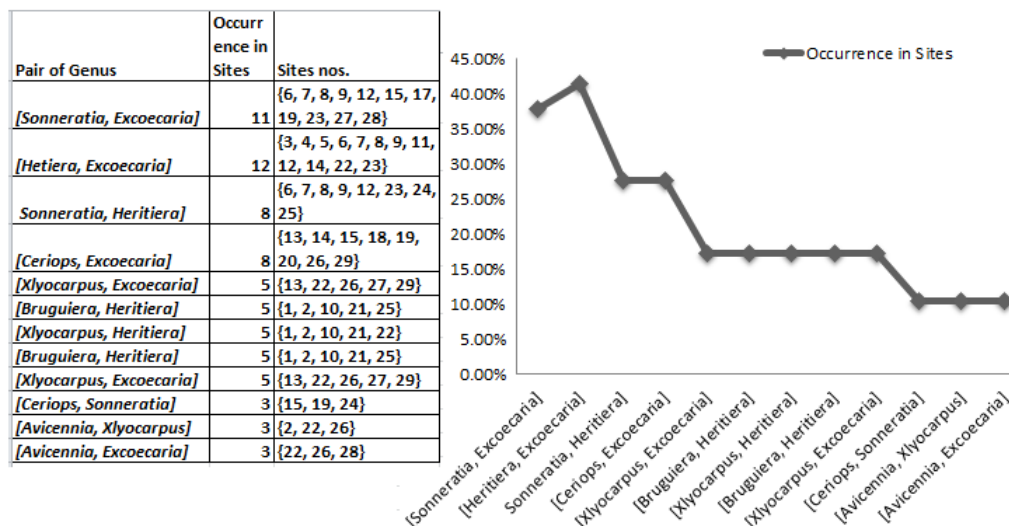


Figure 4.6: List of frequently occurred genus pair in different sites

Table 4.11: Species list and their occurrence sites

Closed Set of Species	Support	Site List
1. <i>Acanthus ilicifolius</i>	28	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29
2. <i>Derris trifoliata</i>	28	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28
3. <i>Vitis trifolia</i> , <i>Acanthus ilicifolius</i> , <i>Derris trifoliata</i>	23	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 15, 16, 17, 19, 20, 23, 24, 25, 26, 27, 28
4. <i>Acrostichum aureum</i>	21	1, 4, 5, 6, 7, 8, 9, 11, 12, 13, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29
5. <i>Phoenix paludosa</i>	19	1, 4, 7, 8, 9, 13, 14, 15, 16, 18, 20, 21, 22, 23, 24, 26, 27, 28, 29
6. <i>Cynometra ramiflora</i> , <i>Vitis trifolia</i> , <i>Acanthus ilicifolius</i> , <i>Derris trifoliata</i>	12	1, 2, 3, 4, 5, 7, 8, 14, 23, 24, 25, 26

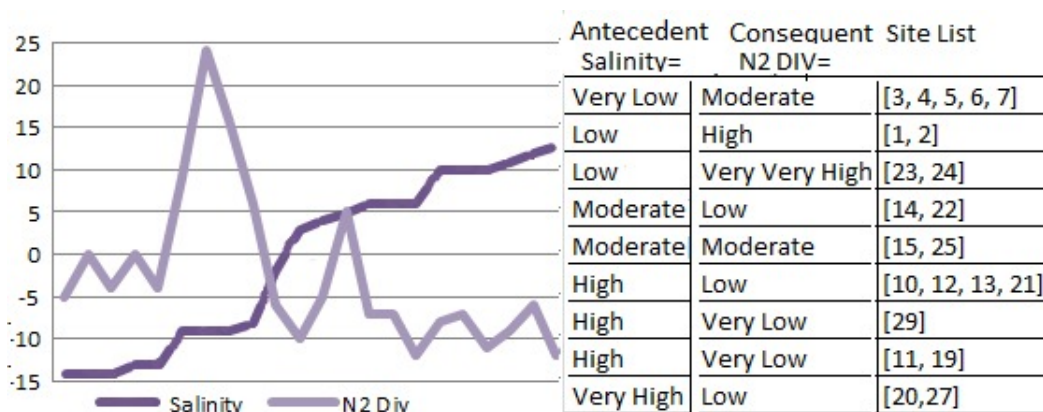


Figure 4.7: In the graph, X-Axis: Sites, Y-Axis: Discretization Range as referred in Figure 4.2

Diversity analysis among the species In [105], 6 undergrowth species having varying tolerance levels of salinity are mentioned. FIST is able to generate the site details for those species. Table 4.11 highlights the result generated by applying FIST on the presence/absence dataset. It shows that *Acanthus ilicifolius*, *Derris trifoliata*, *Vitis trifolia*, *Acrostichum aureum*, *Phoenix paludosa* are the most occurring species. Also, *Cynometra ramiflora* is quite compatible with *Vitis trifolia*, *Acanthus ilicifolius*, *Derris trifoliata* as 12 such sites have this vegetation combination (row 6). Thus, future probable growing sites for *Cynometra ramiflora* could be found by set difference operation of site-lists in between rows 3 and 6.

Variation of N_2Div in relation to salinity Our observation from the obtained rules exhibits that low salinity stimulates the growth of diverse mangrove species. Further increase/ decrease in salinity inevitably decreases the diversity index. Figure 4.7 shows that, for low salinity, N_2Div is high to very very high and for moderate to high salinity, N_2Div decreases, i.e., Salinity is inversely proportional to the N_2Div .

4.6 Summary

The current study emphasizes and intends to show the usefulness of data mining in biodiversity data analysis as interdisciplinary collaboration has already been embraced. A detailed understanding of raw data, data discretization through histogram visualization and thus converting the data in system applicable format is shown thoroughly. Our findings include potential sites for mangrove plantations, a grouping of sites based on salinity/pH, community-wise salinity preference, varying community occurrence at multiple sites, species diversity in relation to salinity, closed group of species lists, etc. This could be a way forward toward the management/ restoration of other related ecosystems as well.

EXPLORATORY DATA ANALYSIS ON INDIAN MANGROVES

5.1	Introduction	49
5.2	Data Sources	49
5.3	Distribution of mangroves in India	49
5.3.1	Indian mangroves in east and west coasts of India	50
5.3.2	Indian mangroves in Sundarban	55
5.4	Diversity analysis	56
5.4.1	East and west coasts of Indian Mangroves	56
5.4.2	Indian Sundarban Mangroves	60
5.5	Summary	61

5.1 Introduction

Systematic review and meta-analysis have a growing impact on ecology. These two, altogether aim at quantitatively encapsulating the study results of multiple experimental/observational published reports. To protect biodiversity and maintain the equilibrium of the ecosystem, relevant measures need to be taken from every aspect of ecology. Although several studies have focused on mangroves, all over the world, there exists adequate scope for improving and organizing the statistics of Indian mangroves for better assessment. This study is focused on enriching the knowledge in this field to accelerate the necessary steps for conservation. It summarizes the datasets of Indian mangroves, finds the statistics on the numerical data, and analyzes and explores the probable ways for management of mangroves playing an important role in the coastal ecosystem.

5.2 Data Sources

Multiple datasets, database platforms, and reports were found through a web search and literature review, adopting a solution of key terms (e.g., Indian mangrove, Indian estuarine mangrove biodiversity, Sundarban mangrove dataset, Indian mangrove dataset). A snowballing technique was applied to find other related research data.

Multiple leading websites working on biodiversity data were explored. Like Mangrove Reference Database and Herbarium (Dahdouh-Guebas F. (Ed.) (2021)), World Mangroves database (Accessed at <http://www.marinespecies.org/mangroveson2021-04-30.doi:10.14284/460>), Online database of Environmental Information System portraying mangrove cover of Indian states and territories (http://www.frienviis.nic.in/Database/Mangrove-Cover-in-India_2444.aspx). In addition to this, the latest versions of biodiversity reports on Indian Mangroves were mined for references to the mangrove-specific information, specifically, a unique report compiled by WWF on the State of the Art Report on Biodiversity in Indian Sundarbans [54], report from Forest Survey of India (http://www.frienviis.nic.in/Database/Mangrove-Cover-Assessment-2019_2489.aspx), and World Bank group report on Indian Sundarban [109]. Important book sources of our study are [55],[110].

5.3 Distribution of mangroves in India

We summarize our study in two broad sections as discussed below. Firstly, we tabulate the mangrove occurrences in 19 regions along the east and west coasts of India. Consequently, we concentrate and archive the mangrove occurrence detail at Hooghly-Matla estuary, i.e., in Sundarban. Description given in Table 5.1.

Table 5.1: Description for the compiled datasets

Sl no	Description	Rows	Columns
1	Indian mangroves occurrence data Table 5.2	34 Indian mangroves	19 estuaries along the east and west coasts
2	Taxonomic details for the mangroves in India Table 5.3	34 Indian mangroves	3 Unique identifiers (WoRMS ID and ITIS TSN and GBIF taxon ID), and 6 ranks in taxonomic hierarchy
3	Indian Sundarban mangroves occurrence data Table 5.4	102 mangroves specific to Indian Sundarban	Occurrence status and 22 regions across 5 main blocks
4	Taxonomic details for the mangroves in Indian Sundarban Table 5.5	102 mangroves specific to Indian Sundarban	3 Unique identifiers (WoRMS ID or ITIS TSN or GBIF taxon ID), and 6 ranks in taxonomic hierarchy

5.3.1 Indian mangroves in east and west coasts of India

Occurrence data: 19 estuaries along with Andaman are identified along the east and west coasts of India (Figure 5.1). The estuaries exhibit the most dynamic ecosystem because of the blending of riverine freshwater and salty seawater. Being an ecotone zone or transitional region, they are exceptionally fertile and home to uniquely adapted and precious mangrove ecosystems.

Hooghly-Matla estuary, in West Bengal has shaped the major part of Gangentia delta: Sundarban. It is the largest mangrove cover in the world and can be classified into very dense, moderately dense, and open mangrove forest areas. Almost 42% of the mangrove area of India has been nourished at Sundarban. Subarnarekha estuary is covered by Indian states of West Bengal, and Odisha. Brahmani-baitarani estuary, Mahanadi estuary are also situated in Odisha. Bhitarkanika is the second largest mangrove forest in India which is located in Odisha. Godavari estuary, Vamsadhara estuary, Kakinada bay, Krishna delta and Pennar estuary are located in Andhra Pradesh. Pichavaram mangrove is one of the largest mangrove forests located in Tamil Nadu. Besides, Cauvery, Vellar, and Ennore estuaries are also in Tamil Nadu. Andaman Island is another important home to mangroves where the mangrove community is scattered throughout the whole Island. A relatively low number of estuaries with mangrove cover are found that are situated along the west coast. Wandoor mangrove forest and Cochin estuary are situated in Kerala across the west coast of India. Zuari, Mandovi is located in Goa estuary. Tapi estuary located in Gujrat. Almost 93% of the total mangrove cover in India is reported from the above-mentioned states. A pictorial



Figure 5.1: Significant Mangrove Forest locations in India

view of the state-wise mangrove cover data is given in Figure 5.2. West Bengal has the major occurrence of mangrove and Sundarban delta is the mainland for this ecosystem. The occurrence data of 34 Indian mangroves is given in Table 5.2.

Taxonomic information: The importance of taxonomy in species conservation activity is already recognized [111]. Taxonomic data is required to set the rules to standardize the required species unit for in situ conservation. To identify each species globally, an authoritative tool utilized by experts in the field of taxonomy and biodiversity is the Integrated Taxonomic Information System (ITIS) [112]. We provide taxonomic serial number (TSN) for each of the species under study. We also use WoRMS-ID [113] and GBIF Taxon ID [114], for more clarification, as no particular database ID was found for all the species under study. The World Register of Marine Species, sometimes known as WoRMS, is a comprehensive database that offers details on the taxonomy and nomenclature of marine organisms. Whereas, the Global Biodiversity Information Facility, or GBIF, is a global database that makes biodiversity data from multiple sources available. The GBIF taxon ID is a special number that is assigned to each taxonomic entry within the database.

The taxonomic information for all 34 mangroves is listed in Table 5.3. They are from 3 classes, 11 orders, and 15 families. The data is assembled mainly from World Register of Marine Species (<http://www.marinespecies.org/aphia.php?p=taxdetails&id=211508>), Integrated Taxonomic Information System (<https://www.itis.gov/>), National Technology for Biotechnology Information (<https://www.ncbi.nlm.nih.gov/taxonomy>), and Global Biodiversity Information Facility (<https://www.gbif.org/>), Canadian Biodiversity Information Fa-

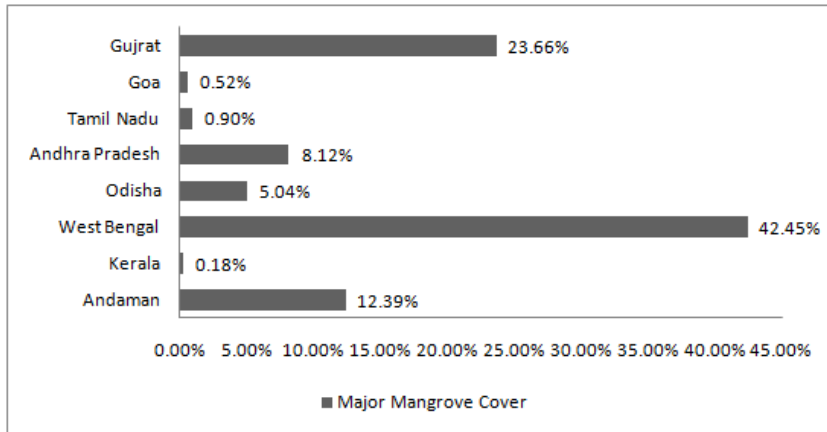


Figure 5.2: Major Mangrove covers in percentage in India (Source: 2019 Report by Forest Survey of India)

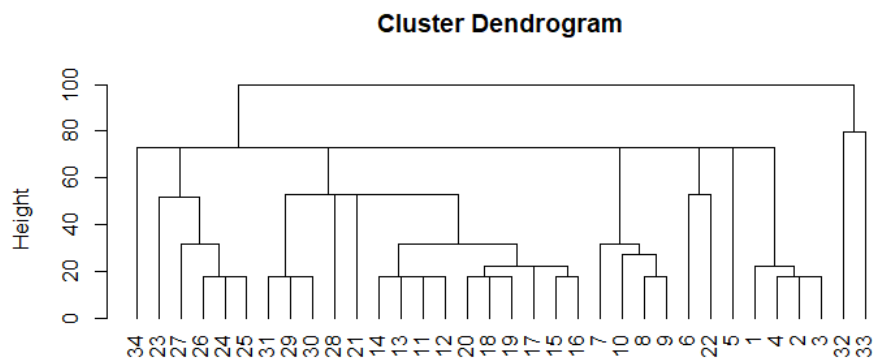


Figure 5.3: Hierarchical cluster showing mangrove species similarity in India

Table 5.2: Mangrove occurrence data at 19 identified regions in India: E1. Andaman Island E2. Wandoor mangroves E3. Hooghly-Matla estuary E4. Subarnarekha estuary E5. Brahmani-Baitarani estuary E6. Bhitarkanika Conservation Area E7. Mahanadi Mangroves E8. Vamsadhara estuary E9. Godavari estuary E10. Kakinada bay E11. Krishna delta E12. Pennar estuary E13. Ennore estuary E14. Cauvery estuary E15. Pichavaram Mangroves E16. Cochin estuary E17. Zuari estuary E18. Mandovi estuary E19. Tapi estuary

No	Species	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19
1	<i>Acanthus ilicifolius</i> Linnaeus	0	0	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1
2	<i>Avicennia alba</i> Blume	0	0	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0
3	<i>Avicennia marina</i> (Forsk.) Vierh	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	<i>Avicennia officinalis</i> L.	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0
5	<i>Aegialitis rotundifolia</i> Roxb.	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	<i>Aegiceras corniculatum</i> (L.) Blanco	0	0	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0
7	<i>Brownlowia tersa</i> (L.) Kosterm	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	<i>Heritiera fomes</i> Buch.-Ham	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
9	<i>Heritiera littoralis</i> Dryand	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	<i>Heritiera globosa</i> Kostermans	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	<i>Bruguiera cylindrica</i> (L.) Blume	0	0	1	0	1	0	0	1	1	1	1	1	0	0	1	0	1	1	0
12	<i>Bruguiera gymnorhiza</i> (L.) Savigny	0	0	1	0	1	0	1	0	1	1	1	1	0	0	0	1	1	1	0
13	<i>Bruguiera sexangula</i> (Lour.) Poir.	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14	<i>Bruguiera parviflora</i>	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
15	<i>Ceriops decandra</i> (Griff) Ding Hou	0	0	1	0	1	0	1	1	1	1	1	1	0	1	1	0	0	0	0
16	<i>Cerriops tagal</i> (Perr. C. B. Roby)	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
17	<i>Kandelia candel</i> Druce	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	0
18	<i>Rhizophora apiculata</i> Blume	0	1	1	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0
19	<i>Rhizophora mucronata</i> Lamk	0	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0
20	<i>Rhizophora annamalyana</i> Kathiresan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
21	<i>Excoecaria agallocha</i> L.	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
22	<i>Scyphiphora hydrophyllacea</i> C. F. Gaertn	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
23	<i>Sonneratia alba</i> J. Smith	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0
24	<i>Sonneratia caesularis</i> (L.) Engl.	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0
25	<i>Sonneratia apetala</i> Buch. - Ham	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	1
26	<i>Sonneratia griffithi</i> Kurz	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
27	<i>Lumnitzera racemosa</i> Willd.	0	0	1	0	1	0	1	0	1	1	1	1	0	1	1	0	1	1	0
28	<i>Cynometra ramiflora</i> L.	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	<i>Xylocarpus granatum</i> J. Konig	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
30	<i>Xylocarpus mekongensis</i> Pierre.	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
31	<i>Xylocarpus moluccensis</i> (Lamk.)	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
32	<i>Acrostichum aureum</i> L.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	<i>Phoenix paludosa</i> Roxb.	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	<i>Nypa fruticans</i> (Thunb.) Wurmb.	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.3: Taxonomic details for the mangroves in India

No	Species name	ITIS TSN	Unique identifier WoRMS-ID	GBIF taxon ID	Phylum	Class	Superorder	Order	Family	Genus
1	<i>Acanthus ilicifolius</i> Linnaeus	NA	344740	6359819	Tracheophyta	Magnoliopsida	Asteranae	Lamiales	Acanthaceae	Acanthus
2	<i>Avicennia alba</i> Blume	NA	235034	6413459	Tracheophyta	Magnoliopsida	Asteranae	Lamiales	Acanthaceae	Avicennia
3	<i>Avicennia marina</i> (Forsk.) Vierh	506840	235040	2925403	Tracheophyta	Magnoliopsida	Asteranae	Lamiales	Acanthaceae	Avicennia
4	<i>Avicennia officinalis</i> L.	NA	235041	6413451	Tracheophyta	Magnoliopsida	Asteranae	Lamiales	Acanthaceae	Avicennia
5	<i>Agelais rotundifolia</i> Roxb.	NA	235079	5668108	Tracheophyta	Magnoliopsida	Asteranae	Caryophyllales	Plumbaginaceae	Agelais
6	<i>Agelaeas coriandatum</i> (L.) Blanco	NA	235069	3721362	Tracheophyta	Magnoliopsida	Asteranae	Ericales	Primulaceae	Agelaeas
7	<i>Brownlowia tosa</i> (L.) Kosterm	NA	NA	4259431	Tracheophyta	Magnoliopsida	Rosanae	Malvales	Tiliaceae	Brownlowia
8	<i>Heritiera fomes</i>	NA	235117	3669142	Tracheophyta	Magnoliopsida	Rosanae	Malvales	Malvaceae	Heritiera
9	<i>Heritiera littoralis</i> Dryand	507332	235119	3152150	Tracheophyta	Magnoliopsida	Rosanae	Malvales	Malvaceae	Heritiera
10	<i>Heritiera globosa</i> Kostermans	NA	235118	3669117	Tracheophyta	Magnoliopsida	Rosanae	Malvales	Malvaceae	Heritiera
11	<i>Bunguetra cylindrica</i> (L.) Blume	NA	234496	5602619	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Bunguetra
12	<i>Bunguetra gymnorhiza</i> (L.) Savigny	501077	235082	NA	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Bunguetra
13	<i>Bunguetra saxangula</i> (Lour.) Boer	847567	235085	6370882	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Bunguetra
14	<i>Bunguetra parviflora</i>	506853	235084	5384934	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Bunguetra
15	<i>Cerriops decandata</i> (Griff) Ding Hou	NA	235087	3873955	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Cerriops
16	<i>Cerriops tagal</i> (Perr. C. B. Roby)	507282	235088	3086518	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Cerriops
17	<i>Kandelia candel</i> Druce	NA	235090	3874179	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Kandelia
18	<i>Rhizophora apiculata</i> Blume	507389	235093	3086526	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Rhizophora
19	<i>Rhizophora mucronata</i> Lamk	507389	235095	3086526	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Rhizophora
20	<i>Rhizophora amanalayaana</i> Kathiresan	NA	344747	3874103	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Rhizophoraceae	Rhizophora
21	<i>Excoecaria agallocha</i> L.	507312	235057	3071702	Tracheophyta	Magnoliopsida	Rosanae	Malpighiales	Euphorbiaceae	Excoecaria
22	<i>Scyphiphora hypophyllacea</i> C. F. Gaertn	507411	235104	NA	Tracheophyta	Magnoliopsida	Asteranae	Gentianales	Rubiaceae	Scyphiphora
23	<i>Sonneratia alba</i> J. Smith	507416	235107	5406998	Tracheophyta	Magnoliopsida	Asteranae	Myrtales	Lythraceae	Sonneratia
24	<i>Sonneratia caescolaris</i> (L.) Engl.	NA	235109	5635606	Tracheophyta	Magnoliopsida	Rosanae	Myrtales	Lythraceae	Sonneratia
25	<i>Sonneratia apetala</i> Buch. - Ham	NA	235108	5635615	Tracheophyta	Magnoliopsida	Rosanae	Myrtales	Lythraceae	Sonneratia
26	<i>Sonneratia griffithii</i> Kurz	NA	235110	6435171	Tracheophyta	Magnoliopsida	Rosanae	Myrtales	Lythraceae	Sonneratia
27	<i>Lumnitzera racemosa</i> Willd.	NA	235053	2966890	Tracheophyta	Magnoliopsida	Rosanae	Fabales	Combretaceae	Lumnitzera
28	<i>Cyrometra ramiflora</i> L.	507290	NA	2966890	Tracheophyta	Magnoliopsida	Rosanae	Fabales	Fabaceae	Cyrometra
29	<i>Xylocarpus granatum</i> J. König	507454	235064	3190517	Tracheophyta	Magnoliopsida	Rosanae	Sapindales	Melastomaceae	Xylocarpus
30	<i>Xylocarpus melanocephalus</i> Pierre	NA	235065	3851045	Tracheophyta	Magnoliopsida	Rosanae	Sapindales	Melastomaceae	Xylocarpus
31	<i>Xylocarpus moluccensis</i> (Lamk.)	507433	235066	3190516	Tracheophyta	Magnoliopsida	Rosanae	Sapindales	Melastomaceae	Xylocarpus
32	<i>Acrostichum aureum</i> L.	17305	235080	2651706	Tracheophyta	Polypodiopsida	Lilianae	Polypodiales	Peridaceae	Acrostichum
33	<i>Phoenix paludosa</i> Roxb.	NA	NA	5293182	Tracheophyta	Lilopsida	Lilianae	Areccales	Palmae	Phoenix
34	<i>Nypa frutescens</i>	507457	234451	2738422	Tracheophyta	Magnoliopsida	Lilianae	Areccales	Arecaceae	Nypa

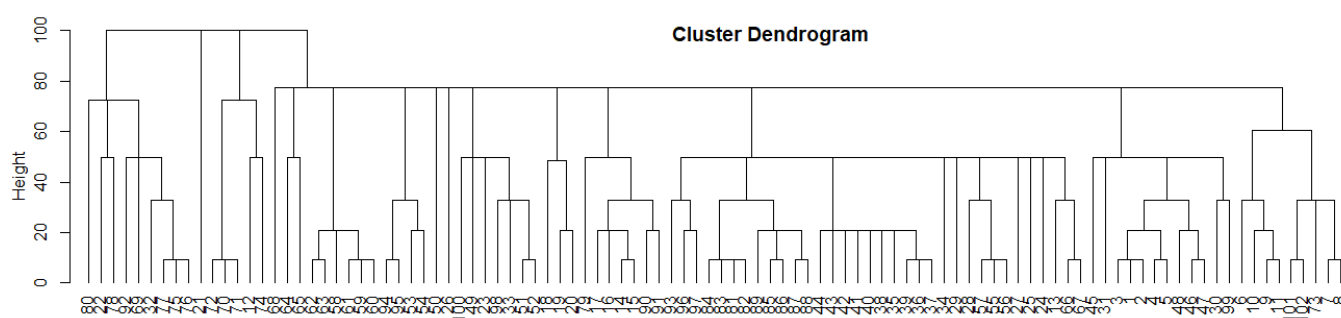


Figure 5.4: Hierarchical cluster showing mangrove species similarity in Sundarban

cility (<https://www.cbif.gc.ca/eng/>). For visualization and classification of the taxonomic relationship among the species, dendrogram offers an efficient way. It represents the visualization of the attribute distance between each pair of sequentially merged entities. It is basically the hierarchical clustering representation. With the help of the function *taxa2dist* of *vegan* package in R [115], we find the taxonomic similarity among the addressed 34 species. Figure 5.3 represents the corresponding dendrogram based on the taxonomic detail presented in Table 5.3. From the arrangement of the clades in the dendrogram, it is found that *Acrostichum aureum* L. and *Phoenix paludosa* Roxb (32 and 33 as per Figure 5.3) are forming a bifolious (two leaved clades) and having greater height indicating greater dissimilarity from the others. For recognizing the most similar species from a list, we need to follow a bottom-up approach. For example, species numbered 11, 12, 13, and 14, all are from the *Bruguiera* genus and under the same cluster.

5.3.2 Indian mangroves in Sundarban

Mangrove occurrence data in Sundarban Sundarban delta part of India is consisting of several forest blocks [54, 116]. The southern part is consisting of Bagmara, Gona, Mayadwip, and Ajmalmari. Jhilla, Pirkhali, and Panchmukhani are forming the northern blocks. The eastern blocks are Arbesi, Khatuajhuri, and Harinbhanga. The western blocks are Matla, Netidhopani, and Chottohardi. The central blocks are Chamta, Chandkhali, and Goasaba. The blocks of 24 Parganas (South) Forest Division are Herobhanga, Ajmalmari, Dhulibhasani, Chulkati, Thakuran, Saptamukhi, and Muriganga. The presence/ absence data of the mangroves and their associates in different forest blocks are given in Table 5.4. The occurrence distribution status is mentioned in the second column. It is discretized into three main groups: *Abundant*, if the occurrence is found in more than 90% of regions, *Frequent*, if the occurrence is reported from more than 45% of regions, Otherwise, it is denoted as a *Rare* species.

Taxonomic information Taxonomically Sundarban mangroves nourish a higher range of diversity (Table 5.5). 102 species, in our consideration, are from 3 classes, 23 orders, and 42 families. The data related to mangrove taxonomy has been gathered as stated before (section 5.3.1). For Sundarban mangroves also, the taxonomic relationship among the species has been studied and visualized through a dendrogram. The corresponding hierarchical

cluster is shown in Figure 5.4.

5.4 Diversity analysis

5.4.1 East and west coasts of Indian Mangroves

Exploratory analysis on the diversity of mangroves in India At the widest stage, the comprehensive and complete diversification of the ecosystem is attributed to its gamma diversity [117]. The diversity of a local community on a single site is referred to as alpha diversity in community ecology [117]. Alpha and gamma diversity measures are related to the third measure of diversity: beta diversity, which measures the distinctions between multiple sub-communities, also known as between-community diversity. Sub-communities differ in their formation, and the average dissimilarity of composition is measured by beta diversity. The measure of beta diversity is a more subtle approach to measurement than alpha and gamma diversity[117]

We utilize the metric for measuring beta diversity to examine the differences in the composition of the mangrove communities between each pair of the 19 locations we are considering in India. Here, we are particularly interested in the Jaccard dissimilarity measure [118]. Given that we have data on the binary presence/absence of species, this metric is suitable in this regard. The overlap in species from the presence/ absence data between two communities is examined in the Jaccard metric. If the two communities are exhibiting homogeneous data of presence, they would have a Jaccard score of zero. It does not consider the abundance of the individual community. Now, to visualize the relationship among the sites for the species presence data in a low dimensional space (instead of considering all the sites and species), we utilize the R package *vegan*. It can compute and plot the principal component analysis (PCA) for the Jaccard metric. Figure 5.5 shows the resultant plot we obtained for the mangrove presence data (Table 5.4). In PCA, the correlations among all the sites are plotted on this 2-dimensional graph. The highly correlated sites are clustered down. Andaman Island has completely different kinds of mangrove growth (as per the plot, it has a large distance from others). Whereas, Ennore estuary, Krishna delta, Kakinada Bay, Pennar estuary, Pichavaram Mangroves, and Vamsadhara estuary are forming a closed group of clusters.

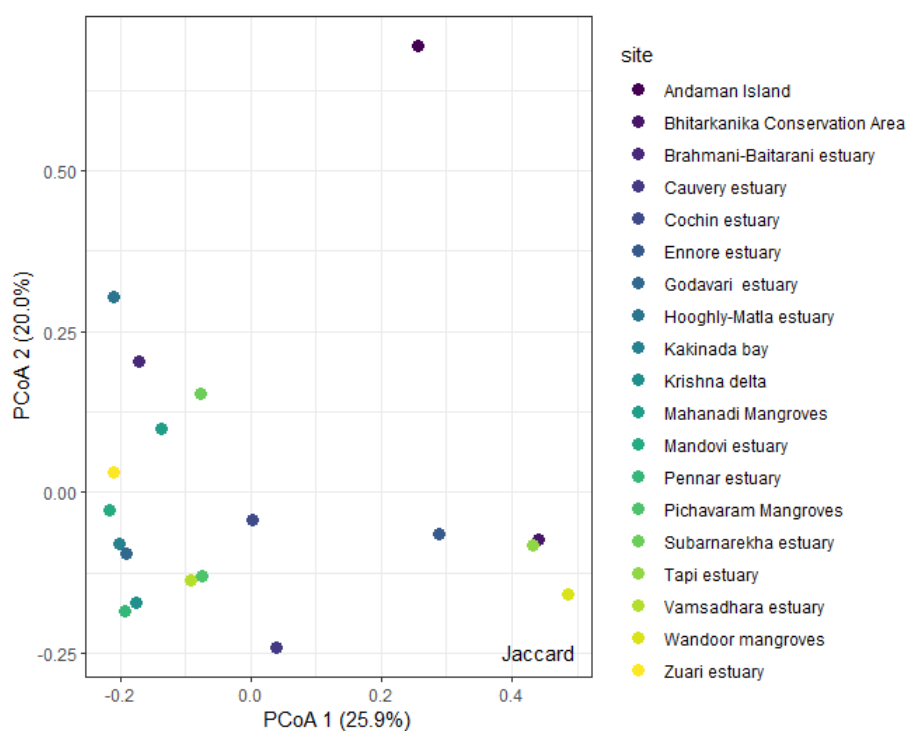


Figure 5.5: PCA analysis on Jaccard dissimilarity metric

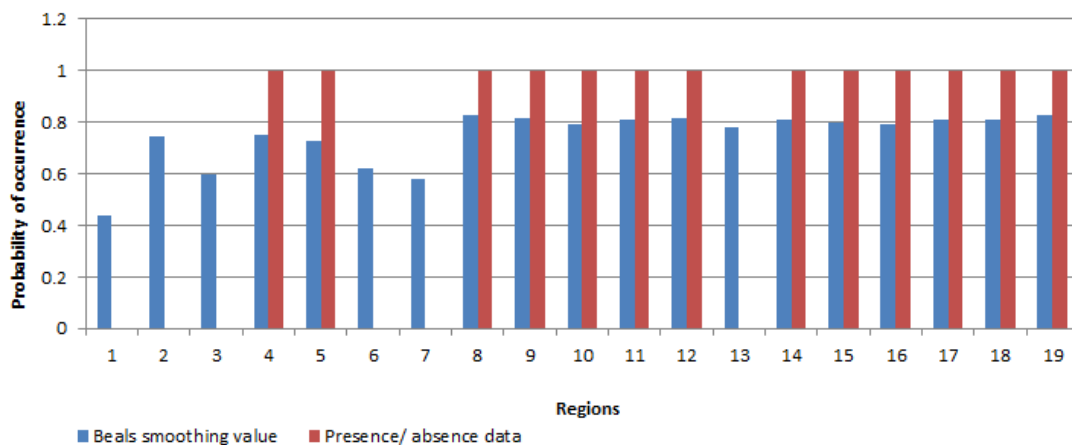


Figure 5.6: probability of occurrence based upon Beals smoothing index

Identifying the possibility of occurrence in the region of non-occurrences taking *Acanthus ilicifolius* for example Several research scientists have identified species-rich afforestation and reforestation as a significant concern [119]. Here we use the *Beals* function from *vegan* package for estimating the probability of the presence of a species (for example, *Acanthus ilicifolius* L.) in a local site where it has an absent record (Following Table 5.2). Beals smoothing replaces each entry in the community data with a likelihood of a target species occurring in that particular site based on the co-occurrences of the target species and the species that are actually present at the site. In Figure 5.6, red bars with values of 1 or 0 indicate the presence or absence of *Acanthus ilicifolius*. Therefore, it can be

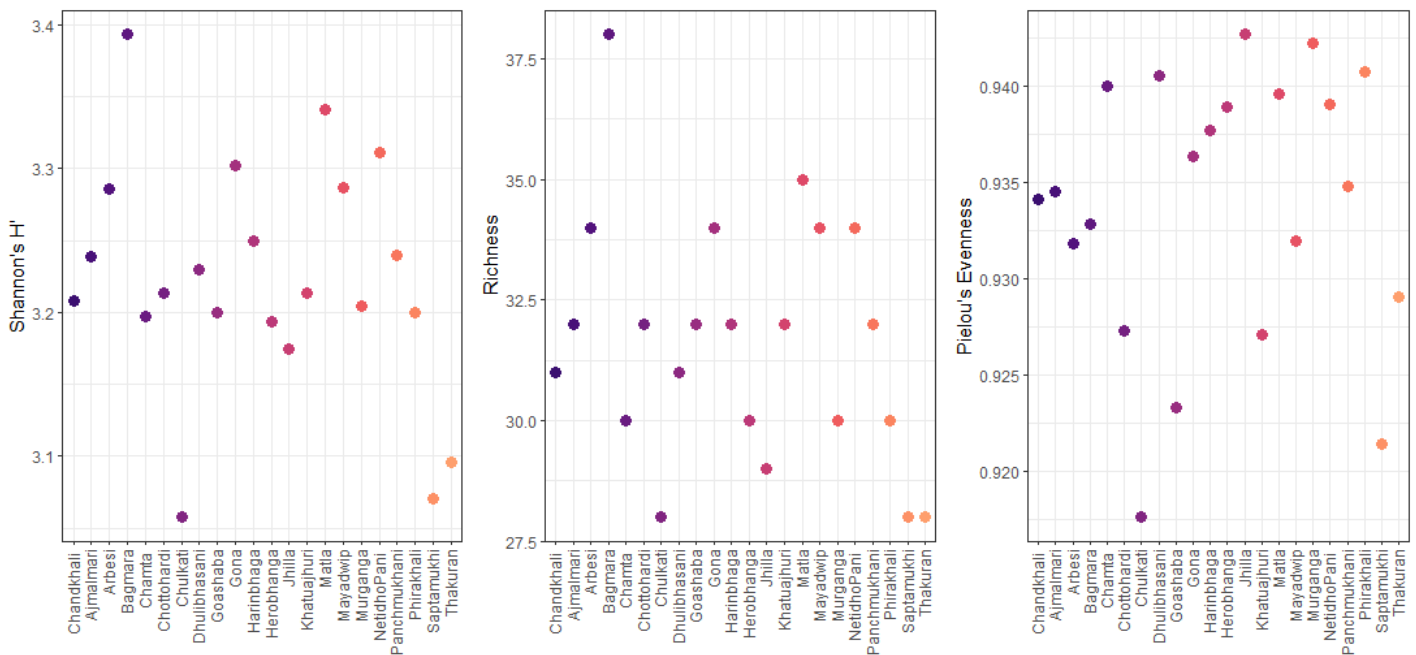


Figure 5.7: Alpha diversity that exists for the mangroves in different blocks in Sundarban

seen that *Acanthus ilicifolius* has been found to have absent records at the site numbered 1, 2, 3, 6, 7, and 13 (site numbers as mentioned in Table 5.2). Blue lines represent the probability of appearance after considering Beals smoothing index. It could be impactful for the conservationist/ policy maker/ forester in making decisions regarding the plantation or comprehending its most probable location of occurrence.

5.4.2 Indian Sundarban Mangroves

To focus on the diversity that exists among the different mangrove communities, we need to study the distribution of them across different blocks in Sundarban, i.e. the alpha diversity that exists in Sundarban for mangrove communities. The three most commonly used indices for alpha diversity are species richness, species evenness diversity by Pielou's Evenness J' [120], and the overall species diversity by Shannon's H' [121]. Exploiting the *vegan* package, we could derive the following Figure 5.7. The graphs generated in Figure 5.7 explain the relationship among the alpha diversity indices. Due to the lack of species count data, we compute the alpha diversity at the family level. In our study, richness simply represents the number of families occurring in a site, whereas, Pielou's Evenness describes how evenly families of species are distributed at a particular site. Shannon's H' considers both richness and evenness. If we focus on Jhilla, and Murganga, these have experienced maximum evenness compared to the others. But, they have moderate Shannon's H' which results from the low richness of these two sites. It suggests that although the family-level distribution of species is highly even, the low value of richness lowers Shannon's H' to a moderate value. It basically says that it is moderately difficult to recognize a randomly chosen individual from the community of these two sites. Similarly, Arbesi, Gona, Mayaowip,

and Netidhopani are found to have almost similar values of richness (family occurrence data are almost homogeneous). Shannon's H' varies depending upon the value of evenness. Again, a high value of Shannon's H' of Bagmara is the result of high richness but a relatively lower uneven value of the distribution. Hence, it can be said that along with the value of Shannon's H' , it may not be possible to identify the reasons behind the value. Richness and evenness both give rise to the value of Shannon's H' .

5.5 Summary

In the current research, a summary of the published and unpublished information on the presence of mangroves in India is provided. Before constructing the conservation strategies, they may be envisioned as a database that would hold the relevant data. In the Sundarban area, it was discovered that 45% of the species are considered rare. Therefore, it can be concluded that there is a need for concentrated study in biodiversity restoration. However, there is a sufficient taxonomic separation between the mangrove species that are found throughout India. Consequently, it can be inferred that there is plenty of room for biodiversity restoration.

ANALYSING INDIAN ESTUARINE FLORA AND FAUNA

6.1	Introduction	63
6.2	Preparation of the dataset	64
6.3	Background Study and the Proposed methodology	68
6.4	Result and discussion	69
6.4.1	Discretized dataset of fauna	69
6.4.2	Presence-only dataset of fauna	71
6.4.3	Discretized dataset of flora	72
6.4.4	Presence-only dataset of flora	73
6.5	Summary	73

6.1 Introduction

Background and motivation: Estuaries represent the transitional ecosystem between freshwater and marine environments. Being dominated by both kinds of aquatic realms, it offers one of the most diverse ecosystems. However, Indian estuaries need a more exhaustive survey for the proper management of the wetlands as the estuarine ecological niche of flora and fauna is at risk.

Mainly anthropogenic movements including trading, industrial and recreational activities, are the underlying reasons behind the deteriorating estuarine ecosystem and biodiversity. Comprehending the importance of the estuarine ecosystem, we tried to concentrate on knowledge discovery from Indian estuarine data of flora & fauna. Here, we show the efficient use of the combining approach for biclustering and association rule mining on a manually curated real dataset. We came up with a set of rules, presentable to the ecologists as it can summarize closely occurred member lists, predicted lists of sites for member expansion, etc. Hence our study would assist in reinforcing the estuarine diversity that could pioneer region-based further studies.

Different physicochemical parameters for example temperature, pH, salinity, electrical conductivity, total dissolved solids, total suspended solids, turbidity, etc. along with various biological parameters are the main influencing factors for the unique supporting ecosystem in estuaries. Though estuaries are dynamic systems favoring the proliferation of diverse biota, estuary health is losing its dignity [122, 123]. Pollution, overfishing, and nutrient run-off are the effects of different human activities, which are the major identified reasons for this perpetuating loss of the estuarine ecosystem.

To prevent the depletion of this self-sustaining habitat, global concern, as well as the implementation of policies and mechanisms, are obvious. A focused study for exploring useful data and revealing knowledge can help in generating eco-awareness and thus preserving biodiversity at estuaries. A study has shown that NCCOS (National Centers For Coastal Ocean Science) has completed a research project [124] aiming at developing a database, namely the Estuarine Living Marine Resources (ELMR) program database. This publicly available repository contains information on 5 regions in total: West Coast, Gulf of Mexico, Southeast, Mid-Atlantic, and North Atlantic. For each class, the database includes habitat type (tidal/ freshwater/ mixing zone), class presence (monthly distribution), and relative abundance (e.g. not present to highly abundant). The dataset is updated regularly and is approachable for an analytical job.

Leading by such kind of practice, here, we would like to investigate Indian estuarine biodiversity database containing the information related to diversity in presence along with the frequency of presence data. Government of India website for The Environment and Information System (ENVIS) Center on Wildlife and Protected Areas has summarized multiple resources for Indian flora, fauna, and data regarding multiple estuaries at the State level. We feel the necessity for gathering and summing up data in compliance with the research level. Currently, no such study describing Indian estuary database in terms of flora-fauna class diversity is available, to the best of our knowledge.

This drives us to introduce an estuarine floral-faunal diversity database. We have col-

lected the data from a book published by the Zoological Survey of India [110]. This book contains information related to integrated flora and fauna diversity, along with 20 major estuaries of India. We explore the database by exploiting data mining methodology. This kind of knowledge exploration would help ecologists in taking suitable measures for estuarine ecosystem preservation.

Contribution: Recapitulating our studies, the main contributions are:

- Introduce a real dataset of presence/absence status for the flora and fauna of Indian estuary.
- Showing up the detailed methodology on data preprocessing for making the dataset usable for employing data mining methodology.
- Discussion on the domain-specific significance of frequent closed itemsets and rule mining in investigating ecological data.

6.2 Preparation of the dataset

India has a long coastal area along the east and west. Our database contains data on 20 major estuaries from both coastal regions. The following 15 estuaries are from the east coast, Hooghly-Matla, Subarnarekha, Baitarani-Brahmani, Mahanadi, Rushikulya, Bahuda, Vamsadhara, Nagavali, Godavari, Krishna, Penner, Ennore, Adyar, Veller, and Cauveri (i.e., E1 to E15 in Table 6.1). Cochin, Zuari, Mandovi, Tapi, and Narmada are situated on the west coast of India (i.e., E16 to E20 in Table 6.1). All of these estuaries are pointed on Indian Map on page 2, Figure 1 of the book "Indian Estuarine Biodiversity" [110]. 23 rows for faunal groups and 3 rows for floral groups are manually curated from the book and it stores the estuary-wise number of classes found for each group of flora and fauna. This actual dataset is shown in Table 6.1. Our main aim, here, is to discover useful information from the estuarine dataset that could assist in preserving the diversity of the estuarine ecosystem. For this purpose, we would use a data mining approach that demands a preprocessing step on this numerical dataset for a better understanding of the resultset.

Data preprocessing task consists of 5 elementary sub-tasks, namely data cleaning, integration, transformation, reduction, and discretization. Here, the database contains no missing, noisy, or inconsistent data. Also, we are not performing operations on multiple databases. Our dataset contains count data for different classes. Again, the data volume is small enough and no redundancy is occurring here. Thus, data cleaning, integrating, transformation, and reduction are not relevant in this case. We have performed only data discretization that allows us to categorize the column values into a set of discrete levels. Binning is one such discretization technique that is simple and widely used. It works in two ways, as mentioned below:

- Equal-width partitioning: Partition the range into N equal intervals. According to the data values in the dataset, the highest and lowest data values are 377 and 1, respectively. For $N = 7$, the width is 54. A high value of N guarantees less information

Table 6.1: Floral and faunal data at Indian Estuaries

Taxonomic group	Major estuaries of India																			
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Floral Groups																				
<i>Phytoplankton</i>	45	40			162				61	68	12	53		41	39		67		53	47
<i>Mangroves</i>	23	9	26	10			8		16	13	12	3			8	11	15	15	3	
<i>Other Flora</i>			19				9		19	13		1			3		10	10		
Faunal Groups																				
<i>Protozoa</i>					25	26					20	3		23					4	3
<i>Foraminifera</i>					5				47	11					73		14			
<i>Porifera</i>	1								2								2	2		
<i>Cnidaria</i>	24	12		11	20	5			13	3	10				34	3	3			
<i>Ctenophora</i>	1			1	2						1	1			1					
<i>Rotifera</i>	5								14		2	16		13						
<i>Nematoda</i>	2				11												20			
<i>Acanthocephala</i>									1											
<i>Sipuncula</i>	1	1																		
<i>Mollusca</i>	83	49	19	152	47		28	43	73	103	82	10		11	51	26	40	41	30	32
<i>Annelida</i>	91	37	11	34	19		13	4	70	45				24	48	47	70	3		
<i>Arthropoda</i>	377	53	88	45	159	99	24	17	88	118	125	56	58	35	55	167	72	21	25	60
<i>Bryozoa</i>	4																			
<i>Brachiopoda</i>	1	2								1										
<i>Chaetognatha</i>	4			3	6	2				1	2	3				4	6	6		3
<i>Echinodermata</i>	22	6	1						7	2	1									
<i>Hemichordata</i>	1																			
<i>Urochordata</i>				3	6	4					3					1				
<i>Class Pisces</i>	314	146	157	177	45	91	64	71	307	268	63	17	135	82	135	126	73	44	64	49
<i>Class Amphibia</i>	13	3	14							4		3				27				
<i>Class Reptilia</i>	57	5	45	6	2	1	1	1	1	10					4	7	7			
<i>Class Aves</i>	156	108	269	46	1			52	75	17						45	43	150	23	23
<i>Class Mammalia</i>	41	2	27		4				1	11		8			2	5				
Subtotal	1198	424	631	478	352	228	130	188	652	630	320	118	193	188	291	560	336	291	146	170
Total	1266	473	676	488	514	228	147	188	748	724	344	175	193	229	341	571	428	316	198	217

loss, and a low value of N gives low categorization of data that in turn makes result interpretation too difficult. 76% of the data fall in the category of very very low which takes the data range of 1 to 54. It is almost straightforward but we can see that it has skewed the data towards the left (belong to Very Very Low). Next, we have tried for Equal-depth partitioning.

- Equal-depth partitioning :

- Partition the data range into N intervals containing approximately Here, N = 7 is taken for the fauna and N = 3 is for flora datasets. Dataset of flora contains 185 numbers of samples. If we divide it into 7 intervals, as in the previous case, it can be justified in a much better way. Histogram distribution of the data samples for both Flora and Fauna are shown in Figure 6.1.
- It gives an almost homogeneous distribution of data at each bin.

We rename the data attribute values according to the above binning techniques as shown in Figure 6.1. Discretized datasets of fauna and flora are shown below in Table 6.2 and Table 6.3, respectively. Corresponding to the raw dataset of fauna (Table 6.1), their presence only data is shown in Table 6.4.

A statistical representation of the dataset can be depicted below in Figure 6.2. Fauna presence data based on their frequency of class occurrence and the total number of estuaries having a similar kind of frequency of class are identified, and shown in graphical format.

It is evident from Figure 6.2 that the presence of *Pisces* and *Arthropoda* are Moderate to Very Very High, and they cover all 20 estuaries. *Pisces*, *Arthropoda*, *Aves*, and *Mollusca*

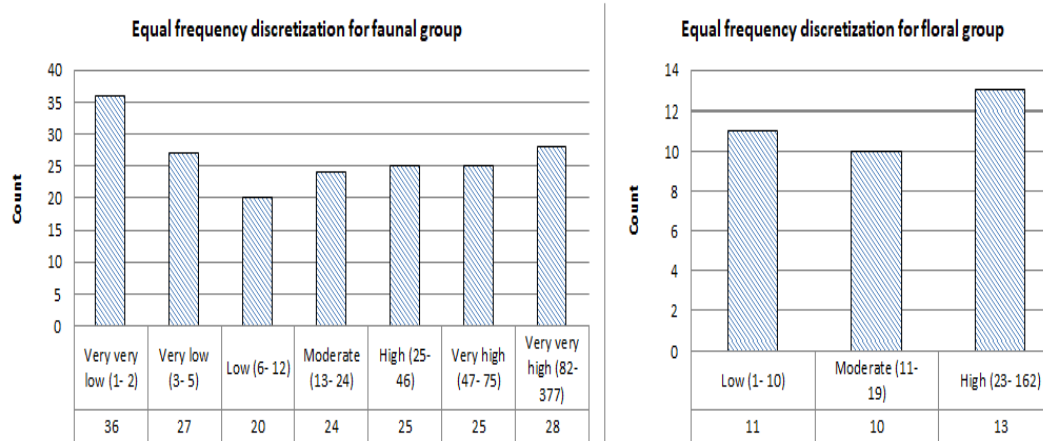


Figure 6.1: Histogram distribution of flora and fauna for data discretization

Table 6.2: Discretized Faunal data at Indian Estuaries(E1 to E20): VVH→ Very Very High, VVL→ Very Very Low, VH→ Very High, VL→ Very Low, H→ High, L→ Low, M→ Moderate; and ? refers to the unavailability of the information

Taxonomic group	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Protozoa	?	?	?	?	H	H	?	?	?	?	M	VL	?	M	?	?	?	?	VL	VL
Foraminifera	?	?	?	?	VL	?	?	?	?	VH	L	?	?	?	?	VH	?	M	?	?
Porifera	VVL	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	VVL	VVL	?	?
Cnidaria	M	L	?	L	M	VL	?	?	M	VL	L	?	?	?	?	H	VL	VL	?	?
Ctenophora	VVL	?	?	VVL	VVL	?	?	?	?	?	VVL	VVL	?	?	?	VVL	?	?	?	?
Rotofera	VL	?	?	?	?	?	?	?	M	?	VVL	M	?	M	?	?	?	?	?	?
Nematoda	VVL	?	?	?	L	?	?	?	?	?	?	?	?	?	?	?	M	?	?	?
Acanthocephala	?	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	?	?	?	?
Sipuncula	VVL	VVL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Mollusca	VVH	VH	M	VVH	VH	?	H	H	VH	VVH	VVH	L	?	L	VH	H	H	H	H	H
Annelida	VVH	H	L	H	M	?	M	VL	VH	H	?	?	?	?	M	VH	VH	VH	VL	?
Arthropoda	VVH	VH	VVH	H	VVH	VVH	M	M	VVH	VVH	VVH	VH	VH	H	VH	VVH	VH	M	H	VH
Bryozoa	VL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Brachiopoda	VVL	VVL	?	?	?	?	?	?	?	VVL	?	?	?	?	?	?	?	?	?	?
Chaetognatha	VL	?	?	VL	L	VVL	?	?	VVL	VVL	VL	?	?	?	?	VL	L	L	?	VL
Echinodermata	M	L	VVL	?	?	?	?	?	L	VVL	VVL	?	?	?	?	?	?	?	?	?
Hemichordata	VVL	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Urochordata	?	?	?	VL	L	VL	?	?	?	?	VL	?	?	?	?	VVL	?	?	?	?
Class Pisces	VVH	VVH	VVH	VVH	H	VVH	VH	VH	VVH	VVH	VH	M	VVH	VVH	VVH	VVH	VH	H	VH	VH
Class Amphibia	M	VL	M	?	?	?	?	?	?	VL	?	VL	?	?	?	H	?	?	?	?
Class Reptilia	VH	VL	H	L	VVL	VVL	VVL	VVL	VVL	L	?	VVL	?	?	?	VL	L	L	?	?
Class Aves	VVH	VVH	VVH	H	VVL	?	?	VH	VH	M	?	?	?	?	?	H	H	VVH	M	M
Class Mammalia	H	VVL	H	?	VL	?	?	?	VVL	L	?	L	?	?	VVL	VL	?	?	?	?

Table 6.3: Discretized Floral Data of Indian Estuaries: VVH→ Very Very High, VVL→ Very Very Low, VH→ Very High, VL→ Very Low, H→ High, L→ Low, M→ Moderate; and ? refers to the unavailability of the information

Taxonomic group	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Phytoplankton	H	H	?	?	VVH	?	?	?	VH	VH	L	VH	?	H	H	?	VH	?	VH	VH
Mangroves	M	L	H	L	?	?	L	?	M	M	L	VL	?	?	L	L	M	M	VL	?
Other Flora	?	?	M	?	?	?	L	?	M	M	?	VVL	?	?	VL	?	L	L	?	?

Table 6.4: Presence only dataset corresponding to table 6.1 where 1 represents the presence, and ? refers to the unavailability of the information

Taxonomic group	Major estuaries of India																			
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Floral Groups																				
Phytoplankton	1	1	?	?	1	?	?	?	1	1	1	1	?	1	1	?	1	?	1	1
Mangroves	1	1	1	1	?	?	?	1	?	1	1	1	?	?	1	1	1	1	1	?
OtherFlora	?	?	1	?	?	?	1	?	1	1	?	1	?	?	1	?	1	1	?	?
Faunal Groups																				
Protozoa	?	?	?	?	1	1	?	?	?	?	1	1	?	1	?	?	?	?	1	1
Foraminifera	?	?	?	?	1	?	?	?	?	1	1	?	?	?	?	1	?	1	?	?
Porifera	1	?	?	?	?	?	?	?	1	?	?	?	?	?	?	?	1	1	?	?
Cnidaria	1	1	?	1	1	1	?	?	1	1	1	?	?	?	1	1	1	1	?	?
Ctenophora	1	?	?	1	1	?	?	?	?	1	1	?	?	?	1	?	?	?	?	?
Rotofera	1	?	?	?	?	?	?	?	1	?	1	1	?	1	?	?	?	?	?	?
Nematoda	1	?	?	?	1	?	?	?	?	?	?	?	?	?	?	1	?	?	?	?
Acanthocephala	?	?	?	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?
Sipuncula	1	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Mollusca	1	1	1	1	1	?	1	1	1	1	1	?	?	1	1	1	1	1	1	1
Annelida	1	1	1	1	1	?	1	1	1	1	?	?	?	1	1	1	1	1	?	?
Arthropoda	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bryozoa	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Brachiopoda	1	1	?	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?
Chaetognatha	1	?	?	?	1	1	?	?	?	1	1	1	?	?	?	1	1	1	?	1
Echinodermata	1	1	1	?	?	?	?	?	?	1	1	1	?	?	?	?	?	?	?	?
Hemichordata	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Urochordata	?	?	?	?	1	1	?	?	?	?	1	?	?	?	?	1	?	?	?	?
Class Pisces	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Class Amphibia	1	1	1	?	?	?	?	?	?	1	?	1	?	?	?	1	?	?	?	?
Class Reptilia	1	1	1	1	1	1	1	1	1	1	?	1	?	?	?	1	1	1	?	?
Class Aves	1	1	1	1	1	?	?	?	1	1	1	?	?	?	?	1	1	1	1	1
Class Mammalia	1	1	1	?	1	?	?	?	1	1	?	1	?	?	1	1	?	?	?	?

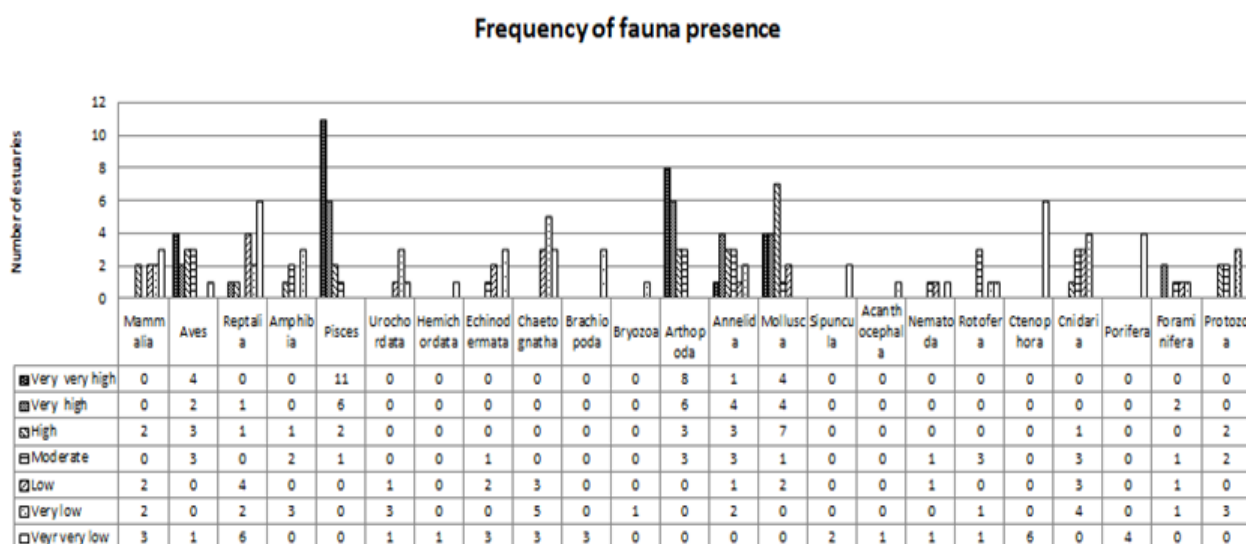


Figure 6.2: Frequency of fauna presence at Indian Estuaries

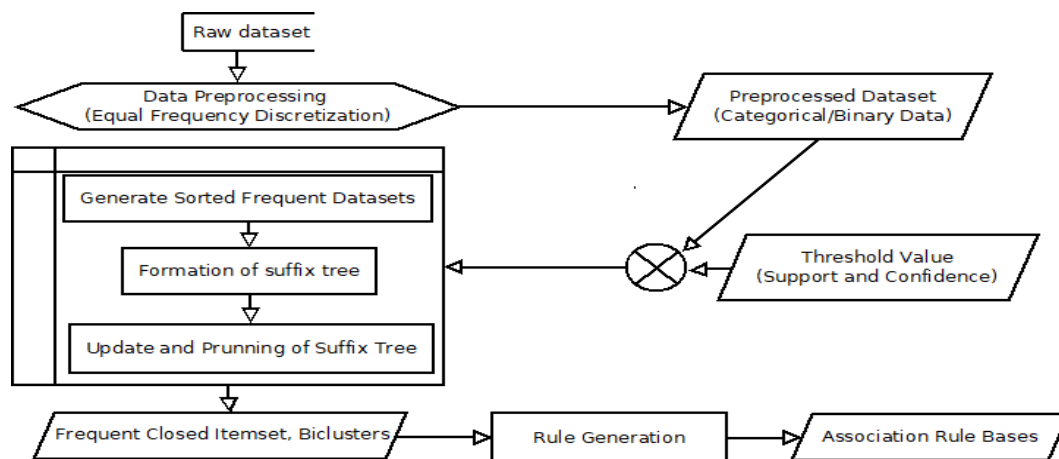


Figure 6.3: Block diagram of the used approach.

are also found with Very Very High frequency with 11, 8, 4, and 4 numbers of estuaries, respectively. Most of the classes are not reported from many of the estuaries. The tool will efficiently be able to extract information related to these without any manual intervention. Thus, the information regarding the chances of class occurrence at the estuaries where they have not been found could establish a proper measure to build the diversity richer. Section 6.4 would deal with this issue.

6.3 Background Study and the Proposed methodology

The advantageous usage of data mining in biodiversity data analysis is not unfamiliar to the research community. Previously, a few studies have identified the beneficial use of it in ecological data analysis [125]. Clustering, classification, rule mining, etc. are the major tasks that can be performed using data mining algorithms. Clustering identifies a similar group of data items from a huge number of data based on the predefined similarity threshold. It can be in 1 or more dimensions. Two-dimensional clustering is mainly gaining popularity in bioinformatics for gene data analysis. Exceptionally, [38] has used it in assessing migratory bird population data. Classification is another major task of data mining and it has been used for decades for species classification in multiple research works [126, 127, 128]. Classification of huge data where data is available in image, video, or audio format rather than simple text, is a challenging task, and deep learning-based approaches [129, 130] are proven to be useful here. Rule mining for future prediction is followed by a few authors where the biodiversity domain is taken into account [32, 30]. Relation among data items and their dependency can be found using rule-based approaches.

Background study directs that association rule mining and bi-clustering are the two promising approaches for analyzing ecological data where the dataset can be clustered down first in both row and column, then based on frequently occurred data, useful rules can be mined. We would like to use a tool proposed in [92, 37] where the computational approach of both the above-mentioned tasks is found in one algorithm. The whole approach has appeared in Figure 6.3.

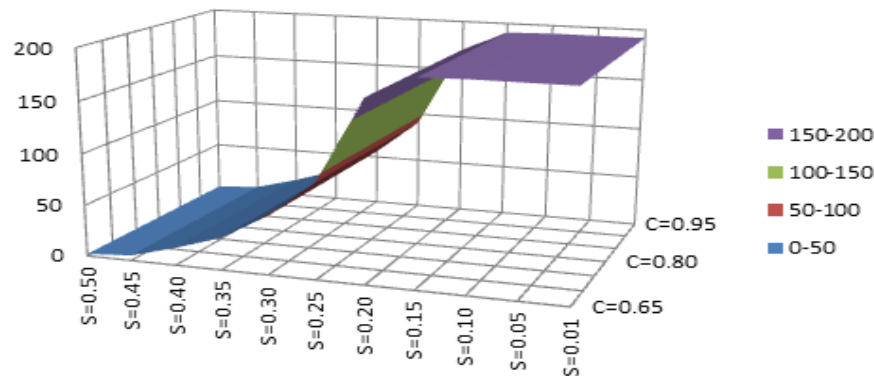


Figure 6.4: Number of rules generated by applying tool: x-axis: min-support; y-axis: min confidence; z-axis: rule count

The algorithm [92] extracts the itemsets that are occurring frequently in the dataset. These are identified as closed frequent itemsets. Then it generates the underlying facts in the form of antecedent and consequent. These facts are known as association rules. Along with the generated rules, it shows the object lists satisfying the facts which will make the algorithm more reliable. In addition to this, the data structure (suffix-tree) used behind the tool makes it better both in terms of time and space complexity [37].

6.4 Result and discussion

We have followed the operations as shown in Figure 6.3. The information retrieval task is performed on the discretized datasets shown in Table 6.2 and Table 6.3 and the presence-only dataset as shown in Table 6.4.

A huge set of rules can be generated by lowering the constraints (support and confidence). By changing the values of minimum support and minimum confidence (step size 5%), we have listed down the statistics of rules generated with the presence only dataset of fauna for all combinations of min-support and min-confidence values and shown the variation in the number of rules generated in Figure 6.4. It can be seen that the number of rules is larger for the lower support values as the constraints are relaxing.

Below, we are going to highlight our observations from the generated set of rules and show the way the tool-generated resultant file can easily be interpreted and can be sorted and searched according to the needs of the user.

6.4.1 Discretized dataset of fauna

The result obtained from the discretized dataset of fauna presented in Table 6.2 is given below:

- Consider the example of class Pisces. Extracted rows from the result file (Table 6.5, the rule set 1) show that the frequency of occurrence of species counts for this class is Very High to Very Very High. Class Pisces belongs to one of the most diverse classes. Mollusca, Arthropoda, and Aves form a closed group along with Pisces for a class frequency level of Very Very High (shown in Table 6.5 of the rule set 2.)

Table 6.5: Sample rule set 1(upper) and rule set 2(lower) for the class Pisces extracted from discretized fauna dataset

Closed set	Support	Estuary List
[Class Pisces= VVH]	11	1, 2, 3, 4, 6, 9, 10, 13, 14, 15, 16
[Class Pisces= VH]	6	7, 8, 11, 17, 19, 20
[Arthropoda= VVH, Class Pisces= VVH]	6	1, 3, 6, 9, 10, 16
[Mollusca= VVH, Class Pisces= VVH]	3	1, 4, 10
[Mollusca= VVH, Arthropoda= VVH]	3	1, 10, 11
[Class Aves= VVH, Class Pisces= VVH]	3	1, 2, 3

- This states that co-occurrence among these classes is quite familiar, suggesting the probability of more availability at not-found estuaries. Like, both Mollusca and Pisces are found in very very high frequency at estuary E1, E4, and E10. As Pisces is very very high at 11 estuaries in total (E1, E2, E3, E4, E6, E9, E10, E13, E14, E15, E16), remaining 8 estuaries are probable sites for Mollusca with very very high frequency.
- Table 6.6 says, class numbered C10 (*Mollusca*), C11 (*Annelida*), and C22 (*Class Aves*), in 60% cases, can be found with very high frequency at estuary 9 when at estuary 1 they occur with very very high frequency. Hence, it can be inferred that for the remaining 40% cases, estuary 9 may have a very high frequency of occurrence for all the classes having very very high frequency at estuary 1. We obtain [E1=VVH] for classes 10, 11, 12, 19, 22 from frequent closed itemsets, So, it can be concluded that C12 and C19 have the probability of occurring with very high frequency at E9 and the confidence of this inference is 60%.

Table 6.6: Generated rules having the same antecedent extracted from discretized fauna dataset

Rule	Antecedent	Consequent	Support	Confidence	Class List
R1	[E1=VVH]	[E9=VH]	3	0.6	[10, 11, 22]
R2	[E1=VVH]	[E10=VVH]	3	0.6	[10, 12, 19]
R3	[E1=VVH]	[E15=VH]	3	0.6	[10, 11, 12]
R4	[E1=VVH]	[E4=H]	3	0.6	[11, 12, 22]
R5	[E1=VVH]	[E17=VH]	3	0.6	[11, 12, 19]
R6	[E1=VVH]	[E3=VVH]	3	0.6	[12, 19, 22]

- Again, as the antecedents for all the rules are the same, R1 - R2 = 11, 22; indicates that C11 (*Annelida*) and C22 (*Class Aves*) should occur at estuary 10 with 60% confidence. Similarly, R1 - R3 = 22; indicates that C22 (*Class Aves*) should occur at estuary 15 with very high frequency.

6.4.2 Presence-only dataset of fauna

The result obtained from the presence-only dataset of fauna presented in Table 6.4 is given below:

- We get the following frequent closed set without thoroughly investigating the dataset which specifies that *Arthropoda* and *Pisces* are found in all the estuaries
 $[Arthropoda = 1, Class Pisces = 1]$ Support = 20 Estuary list = 1, 2, 3, 4, 5, 6, 7,8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
- *Mollusca* is the 3rd highest found class after *Arthropoda* and *Pisces*. The following frequent closed itemset highlights that *Mollusca* is found at 18 different sites listed in the second position.
 $[Mollusca = 1, Arthropoda = 1, ClassPisces = 1]$ Support = 18 Estuary list = 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20
- Hooghly-Matla is the most diversified estuary. Followed by, Rushikulya, Krishna, and Cochin. 19 different types of faunal classes are found in Hooghly-Matla estuary as shown in Table 6.7.

Table 6.7: Top 4 diversity-rich estuaries extracted from presence only fauna dataset

Closed item	Support	Classes list
[E1=1]	19	3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23
[E5=1]	14	1, 2, 4, 5, 7, 10, 11, 12, 15, 18, 19, 21, 22, 23
[E10=1]	13	2, 4, 10, 11, 12, 14, 15, 16, 19, 20, 21, 22, 23
[E16=1]	13	2, 4, 5, 10, 11, 12, 15, 18, 19, 20, 21, 22, 23

- From the generated rule set, we obtain the rule shown in Table 6.8. It can be interpreted as a class will occur at E16 with 83.33% probability when that class has found at both the estuaries E10 and E1. Alternatively, it can be explained as 83.33% of the classes found at E10 and E1, are also found at E16. For the remaining classes, they have a probability of occurrence with 83.33% confidence.

Table 6.8: Rule generated from the presence only dataset of fauna

Antecedent	Consequent	Support	Confidence	Class List
E10=1, E1=1	E16=1	10	0.8333333	4, 10, 11, 12, 15, 19, 20, 21, 22, 23

- Future probable habitat for a class can be identified from the closed set, alternatively, a future probable class for an estuary can be identified from the closed set.
 - Here, Hooghly-Matla, Subarnarekha, and Krishna estuary are forming a closed set as 11 types of classes shown in Table 6.9 are common for all of them.

$[E2 = 1, E10 = 1, E1 = 1]$ Support= 11 Class= 4, 10, 11, 12, 14, 16, 19, 20, 21, 22, 23

It could be concluded that C9 and C15 class has the probability of occurrence at estuary E10, E2 respectively.

Table 6.9: Presence data for all fauna classes (C1 to C23) at estuaries E1, E2 and E10 where ? refers to the unavailability of the information

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23
E1	?	?	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1
E2	?	?	?	1	?	?	?	?	1	1	1	1	?	1	?	1	?	?	1	1	1	1	1
E10	?	1	?	1	?	?	?	?	?	1	1	1	?	1	1	1	?	?	1	1	1	1	1

- A few closed sets are summarized below in Table 6.10. Using transition law on them, we can say that E1, E2, E5, E10, and E16 are forming a closed group. Thus, these estuaries should have homogeneous presence data for classes.

Presence data of these estuaries are as follows in Table 6.11 where C4, C10, C11, C12, C19, C21, C22, and C23 have a presence in all 5 estuaries. For other classes, probable estuarine habitats can be predicted.

Table 6.10: Closed itemsets extracted from the presence dataset of fauna

Closed item	Support	Estuary list
[E2=1, E1=1]	12	4, 9, 10, 11, 12, 14, 16, 19, 20, 21, 22, 23
[E10=1, E1=1]	12	4, 10, 11, 12, 14, 15, 16, 19, 20, 21, 22, 23
[E16=1, E5=1]	12	2, 4, 5, 10, 11, 12, 15, 18, 19, 21, 22, 23
[E16=1, E1=1]	11	4, 5, 10, 11, 12, 15, 19, 20, 21, 22, 23
[E5=1, E1=1]	11	4, 5, 7, 10, 11, 12, 15, 19, 21, 22, 23

Table 6.11: Presence data for all fauna class (C1 to C23) at estuaries E1, E2, E5, E10 and E16; where ? refers to the unavailability of the information

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23
E1	?	?	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1
E2	?	?	?	1	?	?	?	?	1	1	1	1	?	1	?	1	?	?	1	1	1	1	1
E5	1	1	?	1	1	?	1	?	?	1	1	1	?	?	1	?	?	1	1	?	1	1	1
E10	?	1	?	1	?	?	?	?	?	1	1	1	?	1	1	1	?	?	1	1	1	1	1
E16	?	1	?	1	1	?	?	?	?	1	1	1	?	?	1	?	?	1	1	1	1	1	1

- Rules generated from the fauna dataset in Table 6.12 have a similar object list consists of C6, C10, C12, and C19. Also, E14, E12, E11, E9, and E1 are forming a closed group, and the occurrence of any class is associated with all others belonging to the same group with a minimum of 67%,

Table 6.12: Rules generated from presence only fauna dataset

Antecedent	Consequent	Support	Confidence	Class List
E14=1	E12=1, E11=1, E9=1, E1=1	4	0.6666667	6, 10, 12, 19
E14=1, E12=1, E11=1	E9=1, E1=1	4	0.8	6, 10, 12, 19
E14=1, E9=1, E1=1	E12=1, E11=1	4	0.8	6, 10, 12, 19
E12=1, E11=1, E1=1	E14=1, E9=1	4	0.6666667	6, 10, 12, 19
E12=1, E9=1, E1=1	E14=1, E11=1	4	0.6666667	6, 10, 12, 19
E11=1, E9=1, E1=1	E14=1, E12=1	4	0.6666667	6, 10, 12, 19

6.4.3 Discretized dataset of flora

The result obtained from the discretized dataset of flora presented in Table 6.3 is given below:

- Table 6.13 shows that E7, E9, and E10 are forming the closed group based on mangroves and other flora.

Table 6.13: Rules generated from the discretized flora dataset

Antecedent	Consequent	Classes List
[E7= Low]	[E9= Moderate, E10= Moderate]	2, 3
[E9= Moderate]	[E7= Low, E10= Moderate]	2, 3
[E10= Moderate]	[E7= Low, E9= Moderate]	2, 3

- Mangroves with medium density are associated with Phytoplankton with very high density and are found at 3 estuaries E9, E10, and E17 (observed from frequent closed itemset). However, Table 6.14 extracts a rule showing medium dense mangrove and very high dense phytoplankton accompanying other flora with medium dense with 66.6% confidence at estuary E9 and E10. This indicates that E17 may be expected to have other flora with medium or higher density.

Table 6.14: Rules generated from the flora dataset

Antecedent	Consequent	Support	Confidence	Estuary List
[Mangroves=M, Phytoplankton=VH]	[Other Flora=M]	2	0.6666667	[9, 10]

Table 6.15: Rules generated from the presence only dataset of flora

Antecedent	Consequent	Support	Confidence	Estuary List
Phytoplankton=1	Mangroves=1	9	0.75	[1, 2, 9, 10, 11, 12, 15, 17, 19]

6.4.4 Presence-only dataset of flora

The result obtained from the presence-only dataset of flora as presented in Table 6.4 is discussed below: Table 6.15 highlights a rule with a support value of 9 and a confidence value of 0.75. Table 6.16 says that phytoplankton appears in 12 estuaries which are E1, E2, E5, E9, E10, E11, E12, E14, E15, E17, E19, E20 (generated from frequent closed itemset file) out of which consequent (for the rule in Table 6.15) has occurred in 9 estuaries (E1, E2, E9, E10, E11, E12, E15, E17, E19). Therefore, it can be thought of as reasonably high to predict that mangrove class may also be present in estuaries E5, E14, and E20 with 75% confidence.

Table 6.16: Closed itemsets extracted from the presence only dataset of flora

Closed item	Support	Estuary list
Phytoplankton=1	12	1, 2, 5, 9, 10, 11, 12, 14, 15, 17, 19, 20

6.5 Summary

The aim of this study is to assist ecologists in taking suitable steps for ecosystem conservation via an algorithmic approach. We gather Indian estuarine data and illustrate the way

of knowledge extraction using data mining-oriented methodology. Our followed approach internally utilizes suffix trees. This data structure enables the efficient storage of data and computation of relevant patterns. It requires only a unique scan of the dataset to extract all valid patterns. It can discover the minute details for each class regarding its level of appearances, and co-occurrences and for each estuary regarding its homogeneity in diversity. It is also able to find out the future probable occurrence of a particular class for a particular estuary. Undoubtedly this kind of information could help the ecologists in managing estuarine biodiversity by establishing policy and suitable measures.

CHAPTER **7**

KNOWLEDGE DISCOVERY USING TRI-CLUSTER

7.1	Introduction	76
7.2	Collection of data	76
7.3	Data preprocessing	76
7.3.1	Generating sorted-frequent dataset	78
7.3.2	Constructing frequent generalized itemset suffix-tree . . .	82
7.3.3	Building frequent generalized suffix-forest	86
7.4	Tricluster generation and data interpretation	87
7.5	Extracting biclusters from suffix forest	88
7.6	Identifying rules from the clusters	89
7.7	Estimation of co-related parameters from triclusters	89
7.8	Summary	90

7.1 Introduction

Three-dimensional data, or tridiac, is a collection of multiple two-dimensional data matrices which are of great importance for representing as well as understanding the relationship among data in various domains, like business analysis (product category-sales measure-time), medical data analysis (patient-record-time), bioinformatics (gene-sample-time), biodiversity (species-location-time), etc. The collection of observed data in the varying period of time, i.e. time-series data can solve the analytics query like progress or deterioration of a product sale with time, patient response over time for some particular drug, gene expression changing under certain conditions over time, or changes in species presence record along with different locations in varying time, etc. Thus, modeling an unsupervised learning algorithm on these kinds of data for attributing the underlying structure or distribution of data is a demanding task and is an emerging research topic.

7.2 Collection of data

We have applied the proposed algorithm to Indian forest data available on the website of the Forest Survey of India (FSI). FSI biennially prepares a report named the Indian State of Forest Report (ISFR) under the Ministry of Environment, Forest & Climate Change Government of India. The report publishes different forest parameters in tabular form, which highlights the state/district-wise forest cover and changes in area coverage across time. The remote sensing data and satellite imagery are the major sources for the report generation along with the field data from National Forest Inventory. We retrieve the records starting from 2003 and followed by 2005, 2009, 2011, 2013, 2015, and 2017 (Table 7.1 to 7.7). 9 States and 3 Union Territories have been considered- Andhra Pradesh (AP), Goa, Gujrat (GU), Karnataka (KA), Maharashtra (MA), Kerala (KE), Odisha (OD), Tamil Nadu (TN), West Bengal (WB), A & N Islands (ANI), Daman & Diu (DD) and Puducherry (PU). 7 monitoring parameters are considered here. These are Very Dense Mangrove (VDM), Medium Dense Mangrove (MDM), Open Mangrove (OM), Total Mangrove (TM), Geographic area (GA), Very Dense Forest (VDF), Medium Dense Forest (MDF), Open Forest (OF) and Total Forest (TF). VDF, MDF, and OF are distinct categories of Forest Cover, which are important indicators for forest status. Forest cover is measured in terms of Forest Canopy Density, which reflects the proportion of forest floor covered by the vertical projection of the tree crowns. VDF refers to all lands with tree canopy density of 70% and above. MDF refers to all lands with tree canopy density from 40% to 70%. OF refers to all lands with tree canopy density from 10% to 40%. VDM, MDM, and OM fall in Mangrove Cover Assessment. The density measures are pertaining to forest cover density. The datasets we collected are depicted here (The area is in sq km).

7.3 Data preprocessing

Following the algorithm proposed in [60] we have applied the tri-clustering algorithm to the Indian State-wise forest cover data.

Table 7.1: Statewise forest cover and mangrove cover assessment in sqkm in 2003

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	15	314	329	275069	23	24356	20040	44419
GOA	0	10	0	10	3702	0	1255	901	2156
GU	0	198	762	960	196022	114	6231	8601	14946
KA	0	3	0	3	191791	431	22030	13988	36449
MA	0	3	5	8	307713	8070	20317	18478	46865
KE	8	44	64	116	38863	334	9294	5949	15577
OD	0	160	47	207	155707	288	27882	20196	48366
TN	0	18	17	35	130058	2440	9567	10636	22643
WB	892	894	334	2120	88752	2303	3742	6298	12343
ANI	262	312	97	671	8249	3475	2809	680	6964
DD	0	0	1	1	112	0	2	6	8
PU	0	0	1	1	480	0	17	23	40

Table 7.2: Statewise forest cover and mangrove cover assessment in sqkm in 2005

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	15	314	329	275069	130	24199	20043	44372
GOA	0	14	2	16	3702	55	1095	1014	2164
GU	0	195	741	936	196022	114	6024	8577	14715
KA	0	3	0	3	191791	464	21634	13153	35251
MA	0	3	5	8	301713	8191	20193	19092	47476
KE	0	58	100	158	38863	1024	8636	5935	15595
OD	0	156	47	203	155707	538	27656	20180	48374
TN	0	18	17	35	130058	2650	9790	10604	23044
WB	892	895	331	2118	88752	2302	3777	6334	12413
ANI	255	272	110	637	8249	3359	2646	624	6629
DD	0	0	1	1	112	0	2	6	800
PU	0	0	1	1	480	0	17	25	42

Table 7.3: Statewise forest cover and mangrove cover assessment in sqkm in 2009

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	126	227	353	275069	820	24757	19525	45102
GOA	0	14	3	17	3702	511	624	1016	2151
GU	0	188	858	1046	196022	376	5249	8995	14620
KA	0	3	0	3	191791	1777	20181	14232	36190
MA	0	3	2	5	301713	8739	20834	21077	50650
KE	0	69	117	186	38863	1443	9410	6471	17324
OD	82	97	42	221	155707	7073	21394	20388	48855
TN	0	16	23	39	130058	2926	10216	10196	23338
WB	1038	881	233	2152	88752	2987	4644	5363	12994
ANI	285	262	68	615	8249	3762	2405	495	6662
DD	0	0	1	1	112	0	1	5	6
PU	0	0	1	1	480	0	13	31	44

Table 7.4: Statewise forest cover and mangrove cover assessment in sqkm in 2011

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	126	226	352	275069	1957	14051	12139	46389
GOA	0	20	2	22	3702	538	576	1115	2219
GU	0	182	876	1068	196022	378	5200	9179	14757
KA	0	3	0	3	191791	4502	20444	12604	36194
MA	0	3	3	6	301713	8736	20652	21294	50682
KE	0	69	117	186	38863	1663	9407	9251	17300
OD	82	97	43	222	155707	6967	21370	23008	48903
TN	0	16	23	39	130058	3672	10979	11630	23625
WB	1038	881	236	2155	88752	2994	4147	9706	12995
ANI	283	261	73	377	8249	5678	684	380	6724
DD	0	0.12	1.44	1.56	112	1.4	5.82	13.27	6
PU	0	0	1	1	480	0	17.6	36.07	50.00

Table 7.5: Statewise forest cover and mangrove cover assessment in sqkm in 2013

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	126	226	352	275069	850	26079	19187	46116
GOA	0	20	2	22	3702	543	585	1091	2219
GU	0	175	928	1103	196022	376	5220	9057	14653
KA	0	3	0	3	191791	1777	20179	14176	36132
MA	0	69	117	183	307713	8720	20770	21142	50632
KE	0	3	3	6	38863	1529	9401	6992	17922
OD	82	88	43	213	155707	7042	21298	22007	50347
TN	0	16	23	39	155707	7042	21298	22007	50347
WB	993	699	405	2097	88752	2971	4146	9688	16805
ANI	276	258	70	604	8249	3754	2413	544	6711
DD	0	0	1	1	12	0	1.87	7.4	9.27
PU	0	0	1	1	480	0	35.23	14.83	50.06

Table 7.6: Statewise forest cover and mangrove cover assessment in sqkm in 2015

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	161	191	352	160204	375	13093	10956	24424
GOA	0	16	6	22	3702	542	580	1102	2224
GU	0	135	968	1103	196022	376	5220	9064	14660
KA	0	1	2	3	191791	1781	20063	14577	36421
MA	0	2	4	6	307713	8712	20747	21169	50628
KE	0	88	98	186	38863	1523	9301	8415	19239
OD	59	87	67	213	155707	7023	21470	21861	50354
TN	1	15	23	39	130058	2993	10469	12883	26345
WB	982	692	423	2097	88752	2948	4172	9708	16828
ANI	386	169	49	604	8249	5686	685	380	6751
DD	0	0	3	3	112	1.4	5.82	12.39	19.61
PU	0	0	2	2	480	0	30	25.7	55.38

Table 7.7: Statewise forest cover and mangrove cover assessment in sqkm in 2017

State	VDM	MDM	OM	TM	GA	VDF	MDF	OF	TF
AP	0	213	191	404	162968	1957	14051	12139	28147
GOA	0	20	6	26	3702	538	576	1115	2229
GU	0	172	968	1140	196244	378	5200	9179	14757
KA	0	2	8	10	191791	4502	20444	12604	37550
MA	0	5	4	9	307713	8736	20652	21294	50682
KE	0	88	216	304	38852	1663	9407	9251	20321
OD	82	94	67	243	155707	6967	21370	23008	51345
TN	1	25	23	49	130060	3672	10979	11630	26281
WB	999	692	423	2114	88752	2994	4147	9706	16847
ANI	399	169	49	617	8249	5678	684	380	6742
DD	0	0	3	3	111	1.4	5.82	13.27	20.49
PU	0	0	2	2	490	0	17.6	36.07	53.67

This step makes the subsequent phases faster as preprocessed data can be used repeatedly with varying support values as required by the users. Different tasks that are performed during the preprocessing step are described in this subsection.

Dataset normalization and discretization This step makes the dataset suitable for applying the algorithm. Normalization, in data mining, is a mapping or scaling technique that linearly transforms an existing range into a new range. Given the dataset, the type of normalization to be applied will be decided with the help of domain knowledge. As the algorithm is applicable to the discretized datasets after obtaining normalized values discretization is performed. It reduces the number of values a continuous variable is supposed to have by grouping them into several bins.

Generating sorted frequent dataset (SFD) The first step is to create a sorted frequent dataset to reduce the processing time of suffix-tree generation as well as the memory usage of the algorithm. This is achieved by eliminating infrequent items from the dataset. User-defined threshold value decides whether the item is frequent or infrequent. The generation of a sorted frequent dataset comprises of item table creation and sorted frequent number table creation.

7.3.1 Generating sorted-frequent dataset

Min-max normalization As discussed in subsection 7.3, we have used min-max normalization to feature all attributes on the same scale. Here, for each individual attribute, the minimum value is transformed into 0 and the maximum value is transformed into 1. Every

Table 7.8: Min-max normalized table for VDM

State	2003	2005	2009	2011	2013	2015	2017
AP	0	0	0	0	0	0	0
GOA	0	0	0	0	0	0	0
GU	0	0	0	0	0	0	0
KA	0	0	0	0	0	0	0
MA	0	0	0	0	0	0	0
KE	0.008	0	0	0	0	0	0
OD	0	0	0.078	0.078	0.078	0.056	0.079
TN	0	0	0	0	0	0.0	0.0
WB	0.860	0.860	1	1	0.957	0.946	0.962
ANI	0.252	0.246	0.274	0.273	0.266	0.372	0.385
DD	0	0	0	0	0	0	0
PU	0	0	0	0	0	0	0

Table 7.9: Table 7.8 after removing decimal point followed by multiplying with 10

State	2003	2005	2009	2011	2013	2015	2017
AP	0	0	0	0	0	0	0
GOA	0	0	0	0	0	0	0
GU	0	0	0	0	0	0	0
KA	0	0	0	0	0	0	0
MA	0	0	0	0	0	0	0
KE	0	0	0	0	0	0	0
OD	0	0	0	0	0	0	0
TN	0	0	0	0	0	0	0
WB	8	8	10	10	10	9	10
ANI	2	2	3	3	3	4	4
DD	0	0	0	0	0	0	0
PU	0	0	0	0	0	0	0

other value is transformed into decimal values between 0 and 1. These will be calculated as :

$$Normalized\ value = \frac{(value - min)}{(max - min)} \quad (7.1)$$

We prepare 9 datasets corresponding to 9 attributes, viz., VDM, MDM, OM, TM, GA, VDF, MDF, OF, and TF. So, now our datasets contain variations in areas covered in 12 states from 2003 to 2017. We will illustrate the preprocessing of the dataset for VDM. Applying min-max normalization, we obtain Table 7.8. Table 7.9 shows the corresponding values of Table 7.8 after removing the decimal point by multiplying it by 10 and rounding it off to an integer value.

Equal-width discretization This step follows finding the minimum and maximum value for the attribute and then dividing the range into a number of equal-width intervals as discussed in subsection 7.3. Following the domain knowledge, Table 7.9 contains 10 discrete values from 1 to 10. So, in straightforward binning, we use 5 different levels: A, B, C, D, and E, which correspond to 1-2, 3-4, 5-6, 7-8, and 9-10. So, Table 7.10 represents the final discretized input dataset for VDM that we use for the processing of our algorithm. Similarly, Tables 7.11 to 7.18 are representing the discretized datasets for MDM, OM, TM, GA, VDF, MDF, OF, and TF, respectively.

Item table creation Here, the first task is to scan the dataset and count the support values for each attribute-value pair. Now, the second task is to eliminate all the items having a support value less than the user-given threshold, thus reducing the dataset to be handled later on. Table 7.19 to 7.27 present the item tables for VDM, MDM, OM, TM, GA, VDF, MDF, OF, and TF, respectively. All of these have listed the attribute-value pairs and their respective support values. The right-end column of each table lists the frequent items only. As it is known that the supersets of all the infrequent items will be infrequent [131], earlier elimination of infrequent items will make the algorithm faster.

Table 7.10: Preprocessed dataset of VDM

State	2003	2005	2009	2011	2013	2015	2017
AP	?	?	?	?	?	?	?
GOA	?	?	?	?	?	?	?
GU	?	?	?	?	?	?	?
KA	?	?	?	?	?	?	?
MA	?	?	?	?	?	?	?
KE	?	?	?	?	?	?	?
OD	?	?	?	?	?	?	?
TN	?	?	?	?	?	?	?
WB	D	D	E	E	E	E	E
ANI	A	A	B	B	B	B	B
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.12: Preprocessed dataset of OM

State	2003	2005	2009	2011	2013	2015	2017
AP	B	B	A	A	A	A	A
GOA	?	?	?	?	?	?	?
GU	D	D	E	E	E	E	E
KA	?	?	?	?	?	?	?
MA	?	?	?	?	A	?	?
KE	?	A	A	A	?	A	A
OD	?	?	?	?	?	?	?
TN	?	?	?	?	?	?	?
WB	B	B	A	A	B	B	B
ANI	A	A	?	?	?	?	?
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.14: Preprocessed dataset of GA

State	2003	2005	2009	2011	2013	2015	2017
AP	E	E	E	E	E	C	C
GOA	?	?	?	?	?	?	?
GU	C	C	C	C	C	C	C
KA	C	C	C	C	C	C	C
MA	E	E	E	E	E	E	E
KE	A	A	A	A	A	A	A
OD	C	C	C	C	C	C	C
TN	B	B	B	B	B	B	B
WB	B	B	B	B	B	B	B
ANI	?	?	?	?	?	?	?
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.16: Preprocessed dataset of MDF

State	2003	2005	2009	2011	2013	2015	2017
AP	E	E	E	C	E	B	C
GOA	?	?	?	?	?	?	?
GU	A	A	A	A	A	A	A
KA	D	D	D	D	D	D	D
MA	D	D	D	D	D	D	D
KE	B	B	B	B	B	B	B
OD	E	E	D	D	D	D	D
TN	B	B	B	B	D	B	B
WB	A	A	A	A	A	A	A
ANI	A	?	?	?	?	?	?
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.18: Preprocessed dataset of TF

State	2003	2005	2009	2011	2013	2015	2017
AP	E	E	E	E	E	C	C
GOA	?	?	?	?	?	?	?
GU	B	B	B	B	B	B	B
KA	D	D	D	D	D	D	D
MA	E	E	E	E	E	E	E
KE	B	B	B	B	B	B	B
OD	E	E	E	E	E	E	E
TN	B	B	B	B	E	C	C
WB	A	A	A	A	B	B	B
ANI	A	A	A	A	A	A	A
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.11: Preprocessed dataset of MDM

State	2003	2005	2009	2011	2013	2015	2017
AP	?	?	A	A	A	A	A
GOA	?	?	?	?	?	?	?
GU	A	A	A	A	A	A	A
KA	?	?	?	?	?	?	?
MA	?	?	?	?	?	?	?
KE	?	?	?	?	?	?	?
OD	A	A	A	A	?	?	A
TN	?	?	?	?	?	?	?
WB	E	E	E	E	E	E	E
ANI	B	B	B	B	B	A	A
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.13: Preprocessed dataset of TM

State	2003	2005	2009	2011	2013	2015	2017
AP	A	A	A	A	A	A	A
GOA	?	?	?	?	?	?	?
GU	B	B	B	B	C	C	C
KA	?	?	?	?	?	?	?
MA	?	?	?	?	?	?	?
KE	?	?	?	?	?	?	A
OD	?	?	A	A	?	?	A
TN	?	?	?	?	?	?	?
WB	E	E	E	E	E	E	E
ANI	B	B	B	A	B	B	B
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.15: Preprocessed dataset of VDF

State	2003	2005	2009	2011	2013	2015	2017
AP	?	?	?	A	?	?	A
GOA	?	?	?	?	?	?	?
GU	?	?	?	?	?	?	?
KA	?	?	A	C	A	A	C
MA	E	E	E	E	E	E	E
KE	?	A	A	A	A	A	A
OD	?	?	D	D	D	D	D
TN	B	B	B	B	B	B	B
WB	B	B	B	B	B	B	B
ANI	B	B	B	C	B	C	C
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.17: Preprocessed dataset of OF

State	2003	2005	2009	2011	2013	2015	2017
AP	E	E	D	C	D	C	C
GOA	?	?	?	?	?	?	?
GU	B	B	B	B	B	B	B
KA	C	C	C	C	C	C	C
MA	D	D	E	E	E	E	E
KE	A	A	A	B	B	B	B
OD	E	E	E	E	E	E	E
TN	B	B	B	C	E	C	C
WB	A	A	A	B	B	B	B
ANI	?	?	?	?	?	?	?
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Table 7.19: Item table for VDM and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:D	1	2003:A	1	2005:D	1	No frequent item list is found
2005:A	1	2009:E	1	2009:B	1	
2011:E	1	2011:B	1	2013:E	1	
2013:B	1	2015:E	1	2015:B	1	
2017:E	1	2017:B	1			

Table 7.20: Item table for MDM and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:A	2	2003:E	1	2003:B	1	AP: 2009A, 2011A, 2013A, 2015A, 2017A
2005:A	2	2005:E	1	2005:B	1	GU: 2003A, 2005A, 2009A, 2011A, 2013A, 2015A, 2017A
2009:A	3	2009:E	1	2009:B	1	OD: 2003A, 2005A, 2009A, 2011A, 2017A
2013:A	2	2013:E	1	2013:B	1	ANI: 2015A, 2017A
2015:A	3	2015:E	1	2017:A	4	
2017:E	1					

Table 7.21: Item table for OM and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:A	1	2003:B	2	2003:D	1	AP: 2003B, 2005B, 2009A, 2011A, 2013A, 2015A, 2017A MA: 2013A KE: 2005A, 2009A, 2011A, 2015A, 2017A WB: 2003B, 2005B, 2009A, 2011A ANI: 2005A
2005:A	2	2005:D	1	2005:B	2	
2009:A	3	2009:E	1	2011:E	1	
2011:A	3	2013:A	2	2013:E	1	
2013:B	1	2015:A	2	2015:E	1	
2015:B	1	2017:E	1	2017:A	2	
2017:B	1					

Table 7.22: Item table for TM and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:A	1	2003:B	2	2003:E	1	AP: 2009A, 2011A, 2017A GU: 2003B, 2005B, 2009B KE: 2017A OD: 2009A, 2011A, 2017A ANI: 2003B, 2005B, 2009B, 2011A
2005:A	1	2005:E	1	2005:B	2	
2009:A	2	2009:E	1	2009:B	2	
2011:A	3	2011:B	1	2011:E	1	
2013:A	1	2013:E	1	2013:B	1	
2013:C	1	2015:A	1	2015:E	1	
2015:B	1	2015:C	1	2017:E	1	
2017:A	3	2017:B	1	2017:C	1	
2017:C	1					

Table 7.23: Item table for GA and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:A	1	2003:B	2	2003:C	3	AP: 2003E, 2005E, 2009E, 2011E, 2013E, 2015C, 2017C GU: 2003C, 2005C, 2009C, 2011C, 2013C, 2015C, 2017C KA: 2003C, 2005C, 2009C, 2011C, 2013C, 2015C, 2017C MA: 2003E, 2005E, 2009E, 2011E, 2013E OD: 2003C, 2005C, 2009C, 2011C, 2013C, 2015C, 2017C TN: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B WB: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B
2003:E	2	2005:A	1	2005:B	2	
2005:C	3	2005:E	2	2009:E	2	
2009:A	1	2009:B	2	2009:C	3	
2011:A	1	2011:B	2	2011:C	3	
2011:E	2	2013:A	1	2013:B	2	
2013:C	3	2013:E	2	2015:A	1	
2015:B	2	2015:C	4	2015:E	1	
2017:A	1	2017:B	2	2017:C	4	
2017:E	1					

Table 7.24: Item table for VDF and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:E	1	2003:B	3	2005:E	1	AP: 2011A, 2017A KA: 2009A, 2011C, 2013A, 2015A, 2017C KE: 2009A, 2013A, 2015A, 2017A TN: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B WB: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B ANI: 2003B, 2005B, 2009B, 2011C, 2013B, 2017C
2005:A	1	2005:B	3	2009:B	3	
2009:A	2	2009:E	1	2009:D	1	
2011:A	2	2011:B	2	2011:E	1	
2011:C	2	2011:D	1	2013:B	3	
2013:A	2	2013:D	1	2013:E	1	
2015:A	2	2015:E	1	2015:B	2	
2015:C	1	2015:D	1	2017:A	2	
2017:B	2	2017:C	2	2017:D	1	
2017:E	1					

Table 7.25: Item table for MDF and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:E	2	2003:B	2	2003:A	3	AP: 2003E, 2005E, 2015B GU: 2003A, 2005A, 2009A, 2011A, 2013A, 2015A, 2017A KA: 2003D, 2005D, 2009D, 2011D, 2013D, 2015D, 2017D MA: 2003D, 2005D, 2009D, 2011D, 2013D, 2015D, 2017D KE: 2003B, 2005B, 2009B, 2011B, 2015B, 2017B WB: 2003A, 2005A, 2009A, 2011A, 2013A, 2015A, 2017A ANI: 2003A
2003:D	2	2005:E	2	2005:D	2	
2005:A	2	2005:B	2	2009:B	2	
2009:A	2	2009:E	1	2009:D	3	
2011:A	2	2011:B	2	2011:C	1	
2011:D	3	2013:B	1	2013:E	1	
2013:A	2	2013:D	4	2015:A	2	
2015:B	3	2015:D	3	2017:D	3	
2017:A	2	2017:B	2	2017:C	1	

Table 7.26: Item table for OF and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:E	2	2003:B	2	2003:A	2	AP: 2003E, 2005E, 2011C, 2015C, 2017C GU: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B KA: 2011C, 2015C, 2017C MA: 2009E, 2011E, 2013E, 2015E, 2017E KE: 2003A, 2005A, 2009A, 2011B, 2013B, 2015B, 2017B OD: 2003E, 2005E, 2009E, 2011E, 2013E, 2015E, 2017E TN: 2003B, 2005B, 2009B, 2011C, 2013E, 2015C, 2017C WB: 2003A, 2005A, 2009A, 2011B, 2013B, 2015B, 2017B
2003:D	1	2003:C	1	2005:E	2	
2005:D	1	2005:C	1	2005:A	2	
2005:B	2	2009:B	2	2009:C	1	
2009:A	2	2009:E	2	2009:D	1	
2011:E	2	2011:B	3	2011:C	3	
2013:B	3	2013:E	3	2013:C	1	
2013:D	1	2015:B	3	2015:E	2	
2015:C	3	2017:E	2	2017:B	3	
2017:C	3					

Sorted frequent number table creation: Here, all the attribute values are sorted in increasing order and listed in Table 7.28. The experiment has shown that the increasing order of sorted frequent datasets causes a lower number of nodes in suffix-tree formation. Here, a unique number is given to all the items in this table for generating the number table.

Sorted frequent dataset (SFD) creation: Here, re-scanning has been done to the dataset (shown in Table 7.10 to 7.18) and each frequent item will be replaced with an item number as depicted in Table 7.28 (Sorted Frequent Number Table). A new set of tables will be created named as sorted frequent datasets which will be the input for suffix-tree generation. The sorted frequent datasets are shown below in Tables 7.29 to 7.36 for the actual data presented in Tables 7.1 to 7.7.

7.3.2 Constructing frequent generalized itemset suffix-tree

The suffix-tree generated from MDM is depicted in Figure 7.1 . From Table 7.29, the first string is for AP and the itemset is {3, 14, 15, 16, 31}. The suffixes for this string are {3, 14, 15, 16, 31}, {14, 15, 16, 31}, {15, 16, 31}, {16, 31}, and {31}. So, all these will be included as the branches in the final generalized suffix tree of MDM (as shown in Figure 7.1). Similarly,

Table 7.27: Item table for TF and the list of frequent items

Attribute-Value	Support	Attribute-Value	Support	Attribute-Value	Support	Frequent item list
2003:E	3	2003:B	3	2003:A	2	AP: 2003E, 2005E, 2009E, 2011E, 2013E, 2015C, 2017C GU: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B MA: 2003E, 2005E, 2009E, 2011E, 2013E, 2015E, 2017E KE: 2003B, 2005B, 2009B, 2011B, 2013B, 2015B, 2017B OD: 2003E, 2005E, 2009E, 2011E, 2013E, 2015E, 2017E TN: 2003B, 2005B, 2009B, 2011B, 2013E, 2015C, 2017C WB: 2003A, 2005A, 2009A, 2011A, 2013B, 2015B, 2017B ANI: 2003A, 2005A, 2009A, 2011A
2003:D	1	2005:E	3	2005:B	3	
2005:D	1	2005:A	2	2009:B	3	
2009:D	1	2009:A	2	2009:E	3	
2011:A	2	2011:B	3	2011:E	3	
2011:D	1	2013:B	3	2013:E	4	
2013:D	1	2013:A	1	2015:A	1	
2015:C	2	2015:D	1	2015:B	3	
2015:E	2	2017:B	3	2017:E	2	
2017:C	2	2017:A	1	2017:D	1	

Table 7.28: Sorted Frequent Number Table for all item tables(Table 7.19 to Table 7.27)taking minimum support 2

Attibute-Value	Support	Item Numbers	Attibute-Value	Support	Item Numbers
2003:A	2	1	2011:E	3	18
2005:A	2	2	2013:C	3	19
2013:A	2	3	2015:B	3	20
2003:B	2	4	2003:C	3	21
2005:B	2	5	2009:E	3	22
2009:B	2	6	2009:C	3	23
2003:E	2	7	2011:C	3	24
2011:B	2	8	2013:B	3	25
2003:D	2	9	2011:D	3	26
2005:E	2	10	2015:D	3	27
2005:D	2	11	2017:B	3	28
2017:E	2	12	2009:D	3	29
2015:E	2	13	2017:D	3	30
2009:A	3	14	2017:A	4	31
2011:A	3	15	2017:C	4	32
2015:A	3	16	2015:C	4	33
2005:C	3	17	2013:D	4	34
			2013:E	4	35

Table 7.29: Sorted frequent dataset of MDM

AP	3, 14, 15, 16, 31
GU	1, 2, 3, 14, 15, 16, 31
OD	1, 2, 14, 15, 31
ANI	16, 31

Table 7.32: Sorted frequent dataset of GA

AP	7, 10, 18, 22, 32, 33, 35
GU	17, 19, 21, 23, 24, 32, 33
KA	17, 19, 21, 23, 24, 32, 33
MA	7, 10, 18, 22, 35
OD	17, 19, 21, 23, 24, 32, 33
TN	4, 5, 6, 8, 20, 25, 28
WB	4, 5, 6, 8, 20, 25, 28

Table 7.35: Sorted frequent dataset of OF

AP	7, 10, 24, 32, 33
GU	4, 5, 6, 8, 20, 25, 28
KA	24, 32, 33
MA	12, 13, 18, 22, 35
KE	1, 2, 8, 14, 20, 25, 28
OD	7, 10, 12, 13, 18, 35
TN	4, 5, 6, 24, 32, 33, 35
WB	1, 2, 8, 14, 20, 25, 28

Table 7.30: Sorted frequent dataset of OM

AP	3, 4, 5, 14, 15, 16, 31
MA	3
KE	2, 14, 15, 16, 31
WB	4, 5, 14, 15
ANI	2

Table 7.33: Sorted frequent dataset of VDF

AP	15, 31
KA	3, 14, 16, 24, 32
KE	3, 14, 16, 31
TN	4, 5, 6, 8, 20, 25, 28
WB	4, 5, 6, 8, 20, 25, 28
ANI	4, 5, 6, 24, 25, 32

Table 7.31: Sorted frequent dataset of TM

AP	14, 15, 31
GU	4, 5, 6
KE	31
OD	14, 15, 31
ANI	4, 5, 6, 15

Table 7.34: Sorted frequent dataset of MDF

AP	7, 10, 20
GU	1, 2, 3, 14, 15, 16, 31
KA	9, 11, 26, 27, 29, 30, 34
MA	9, 11, 26, 27, 29, 30, 34
KE	4, 5, 6, 8, 20, 28
WB	1, 2, 3, 14, 15, 16, 31
ANI	1

Table 7.36: Sorted frequent dataset of TF

AP	7, 10, 22, 18, 32, 33, 35
GU	4, 5, 6, 8, 20, 25, 28
MA	7, 10, 12, 13, 18, 22, 35
KE	4, 5, 6, 8, 20, 25, 28
OD	7, 10, 12, 13, 18, 22, 35
TN	4, 5, 6, 8, 32, 33, 35
WB	1, 2, 14, 15, 20, 25, 28
ANI	1, 2, 14, 15

all the suffixes for GU {1, 2, 3, 14, 15, 16, 31}, OD {1, 2, 14, 15, 31}, and ANI {16, 31} will be generated and appended. Hence, the final generalized suffix-tree has been formed for MDM (Figure 7.1).

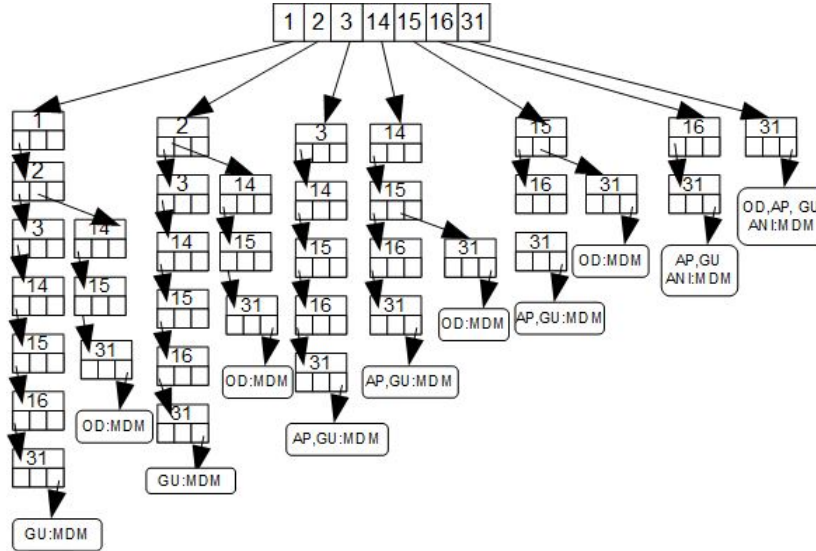


Figure 7.1: Structure of generalized suffix-tree for MDM (input taken from Table 7.29)

Considering Table 7.30, all the suffixes generated from OM are as follows: for AP, {3, 4, 5, 14, 15, 16, 31}, {4, 5, 14, 15, 16, 31}, {5, 14, 15, 16, 31}, {14, 15, 16, 31}, {15, 16, 31}, {16, 31}, {31}; for MA, {3}; for KE, {2, 14, 15, 16, 31}, {14, 15, 16, 31}, {15, 16, 31}, {16, 31}, {31}; for WB, {4, 5, 14, 15}, {5, 14, 15}, {14, 15}, {15}, for ANI, {2}.

All these suffixes get merged with the generalized suffix tree of MDM (Figure 7.1), and we obtain the following Figure 7.2.

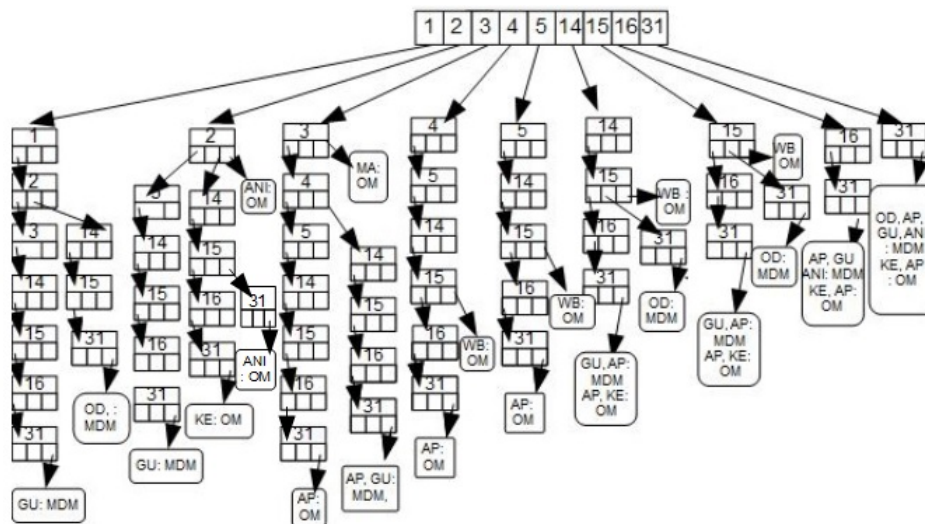


Figure 7.2: Suffix forest by merging MDM and OM (combined input from Table 7.29 and 7.30)

The next step is updation [37]. For each item in the tree, all the branches are compared to others. Every branch is collected following all the itemsets and the object lists in a depth-

first search. Taken any two branches at a time, intersection operation between the itemsets and union operation between the object lists are performed. Using updation, object lists get modified. For example, considering two branches, in Figure 7.2, itemset1 = {1, 2, 3, 14, 15, 16, 31}, and itemset2 = {1, 2, 14, 15, 31}; object list1 = {GU: MDM}, and object list2 = {OD: MDM}; hence, $itemset1 \cap itemset2$ gives object list1 \cup object list2; i.e. $\{1, 2, 3, 14, 15, 16, 31\} \cap \{1, 2, 14, 15, 31\}$ generates $\{GU: MDM\} \cup \{OD: MDM\}$, i.e. $\{1, 2, 14, 15, 31\}$ would contain the new object list {GU, OD: MDM}. Therefore, in Figure 7.3, the updated branch of the suffix forest will be {1, 2, 14, 15, 31, GU, OD: MDM}. The updated suffix-forest is shown in Figure 7.3.

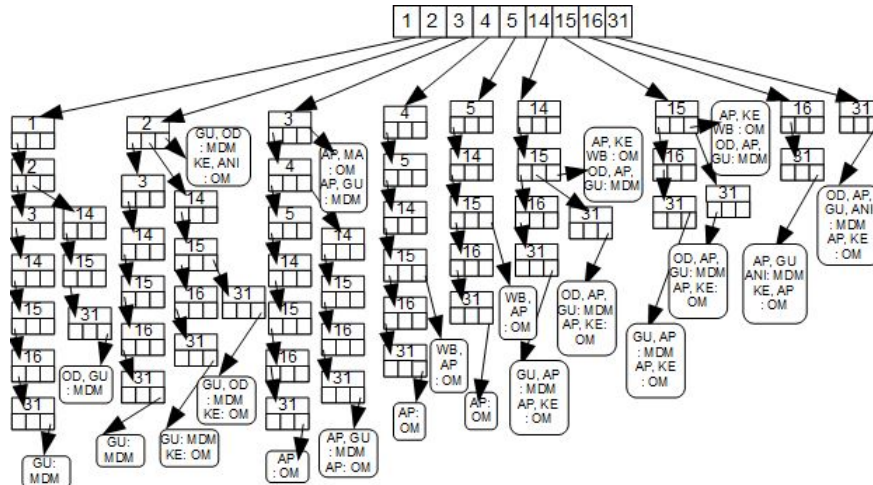


Figure 7.3: Updated suffix-forest obtained from Figure 7.2

Next, pruning is done to remove the redundancy [37]. This step removes nodes that are non-essential for knowledge pattern generation. These nodes represent itemsets that are infrequent and not closed. For any two branches, if the itemsets have the subset-superset relation and the object lists are the same, then the branch having a subset of itemsets can be pruned. For example, consider two branches {1, 2, 3, 14, 15, 16, 31} and {2, 3, 14, 15, 16, 31}, in Figure 7.3, where object lists are same, {GU:MDM}. Pruning can eliminate the branch: {2, 3, 14, 15, 16, 31, GU:MDM}. As it can be seen that only {1, 2, 3, 14, 15, 16, 31, GU:MDM} is present in Figure 7.4. The derived suffix forest is shown in Figure 7.4.

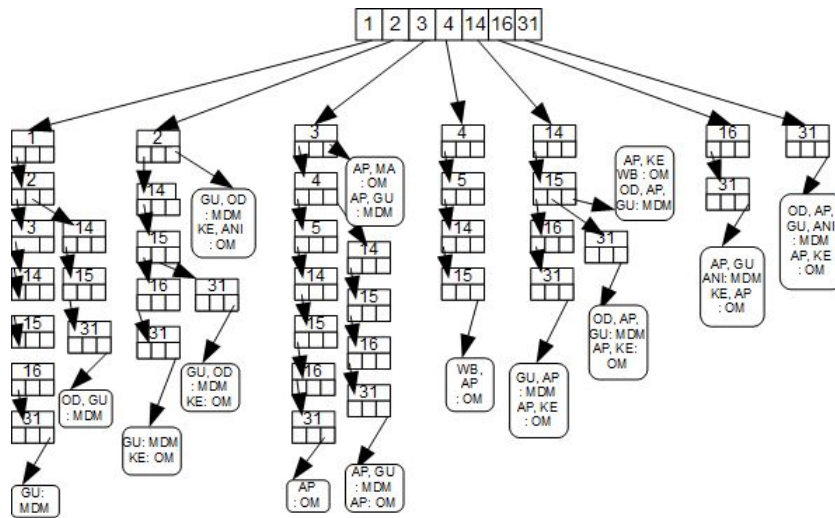


Figure 7.4: Pruned suffix-forest obtained from Figure 7.3

Now, merging the dataset of TM above the merged suffix forest of MDM and OM (Figure 7.2), followed by update and prune, generate the suffix forest that is depicted in Figure 7.5.

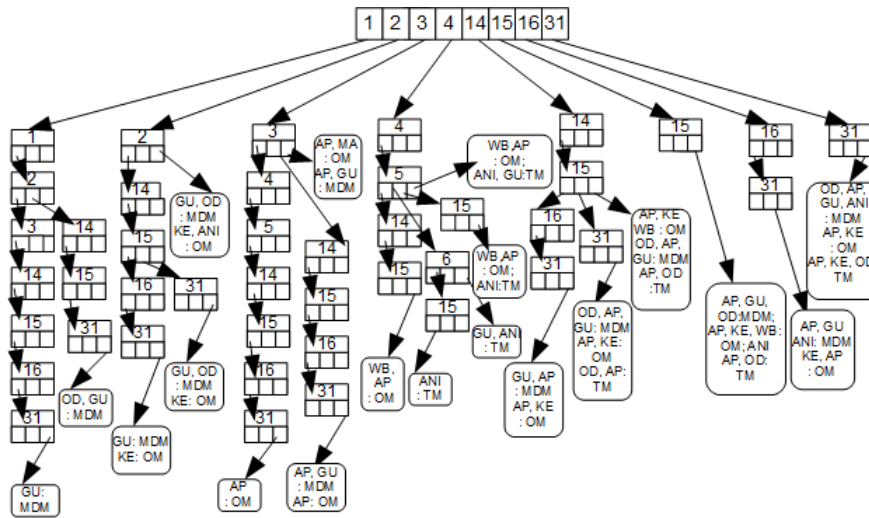


Figure 7.5: Pruned suffix-forest of *MDM*, *OM*, *TM*

Thus, the resultant suffix forest will continue to grow as more datasets are added. This suffix forest generation can be user-driven. That is, the merging of datasets can be processed based on the dataset requirement as per user query. From the pruned suffix-forest obtained in Figure 7.5, we can extract the tri-clusters from the leaf of the data structure.

7.3.3 Building frequent generalized suffix-forest

Building a suffix tree from a single SFD has furnished the idea for building a suffix forest when multiple SFDs are accumulated. SFD dataset contains multiple rows. All are in

ascending order according to their frequency of occurrence. To build a suffix tree from SFD, all rows are scanned one by one. Suffixes are created by deleting one front item in each rotation and appending it to the tree. Thus, for a single itemset, that corresponds to a particular row, the number of branches created, merged, or updated equals the number of items of that itemset. Each branch represents a set of numbers in ascending order. The leaf node shows data that satisfies each element in a particular branch and thus gives clustered results. So, in our example, if we consider the suffix tree generated for MDM, the internal nodes are for the status of year-wise forest presence and leaf nodes are for the states having similar forest statistics. This can be treated as the result of bi-clusters. To merge multiple SFDs, we perform a cumulative addition of new suffixes over the suffix tree built from the first SFD.

7.4 Tricluster generation and data interpretation

We can visualize the tri-clustering notion provided in Figure 3.1 with the help of the created suffix forest in Figure 7.5. The leftmost branch (Figure 7.5), 1- 2- 3- 14- 15- 16- 31, is generating a row-major tri-cluster. As only one dataset is able to extract the object GU: MDM to cluster it down, we alternatively describe it as a row-major cluster (referred to in Table 7.37). This branch reveals that Gujrat has been showing consistency in medium-dense mangrove coverage since 2003. We get another tri-cluster 1- 2- 14- 15- 31- GU, OD: MDM. It explains that, in the years 2003, 2005, 2009, 2011, and 2017, Odisha and Gujrat have the 'A' category of medium-dense mangrove. One may find in the original database that in the years 2013 and 2015, medium-dense mangrove cover area in Odisha decreased from its consistent coverage happening during 2003 to 2011.

Table 7.37: Row major cluster generated from the intermediate suffix tree

State	2003	2005	2009	2011	2013	2015	2017
AP	?	?	A	A	A	A	A
GOA	?	?	?	?	?	?	?
GU	A	A	A	A	A	A	A
KA	?	?	?	?	?	?	?
MA	?	?	?	?	?	?	?
KE	?	?	?	?	?	?	?
OD	A	A	A	A	?	?	A
TN	?	?	?	?	?	?	?
WB	E	E	E	E	E	E	E
ANI	B	B	B	B	B	A	A
DD	?	?	?	?	?	?	?
PU	?	?	?	?	?	?	?

Another example of tri-cluster, as seen in Figure 7.5, is, 14- 15- 31- OD, AP, GU: MDM; AP, KE: OM; OD, AP: TM. This is an irregular sort of tri-cluster, as each slice for MDM, OM, and TM is not homogeneous with regard to the four states we're considering (AP, GU, OD, KE). However, by merely looking at AP, it can be shown that it has the same value for each slice (MDM, OM, TM) and year (2009, 2011, and 2017). Therefore, 14- 15- 31- AP: MDM; AP: OM; AP: TM, would be an example of regular tri-cluster 7.6.

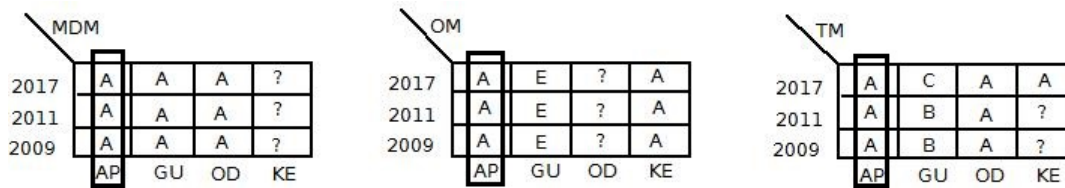


Figure 7.6: Example of regular tricluster

The final suffix forest in our example is shown in Figure 7.5.

Table 7.38 shows the clusters that might be obtained using the branches 14-15-16-31 and 16-31 in Figure 7.5. Attribute, object, and condition are the three dimensions

Table 7.38: Example of extracted Tri-clusters

Tree branch	Tricluster		
	Attribute list	Object list	Condition list
14- 15- 16- 31	2009:A, 2011:A, 2015:A, 2017:A	AP, GU: MDM; AP, KE:OM	MDM, OM
16- 31	2015:A, 2017:A	AP, GU, ANI: MDM; KE, AP: OM	MDM, OM

7.5 Extracting biclusters from suffix forest

We already know that adding dimension to bi-clusters results in tri-clusters. Once we have this pruned suffix forest, we can also extract the bi-clusters. Following the 14-15-16-31 branch (Figure 7.5), for example, we get the tricluster: AP, GU: MDM; AP, KE: OM. Each part individually is representing a bicluster that can be obtained from the individual suffix trees of MDM and OM. For example, AP, GU: MDM is a bicluster that can be seen in the suffix tree of MDM, as shown in Figure 7.1.

The closed itemsets for MDM can be listed as shown in Table 7.39 (following Table 7.11).

Table 7.39: Closed itemset of MDM (Figure 7.11)

Serial no	ClosedSet	Support	Object List
1	2003:E, 2005:E, 2009:E, 2011:E, 2013:E, 2015:E, 2017:E	1	WB
2	2003:B, 2005:B, 2009:B, 2011:B, 2013:B, 2015:A, 2017:A	1	ANI
3	2013:A, 2003:A, 2005:A, 2009:A, 2011:A, 2015:A, 2017:A	1	GU
4	2013:A, 2009:A, 2011:A, 2015:A, 2017:A	2	AP, GU
5	2003:A, 2005:A, 2009:A, 2011:A, 2017:A	2	GU, OD
6	2009:A, 2011:A, 2017:A	3	AP, GU, OD
7	2015:A, 2017:A	3	AP, GU, ANI
8	2017:A	4	AP, GU, OD, ANI

Now the frequent closed itemsets or the bi-clusters would be from serial numbers 4 to 8, which can be obtained from the final suffix-forest (Figure 7.5). Serial numbers 1, 2, and 3 are discarded since their support value is less than the minimum threshold of 2.

Table 7.40: Biclusters generated from the suffix tree of MDM (Figure 7.5)

Serial no	Tree branch	Bicluster	
		Attribute list	Object list
4	3- 14- 15- 16- 31	2013:A, 2009:A, 2011:A, 2015:A, 2017:A	AP, GU: MDM
5	1- 2- 14- 15- 31	2003:A, 2005:A, 2009:A, 2011:A, 2017:A	GU, OD: MDM
6	14- 15- 31	2009:A, 2011:A, 2017:A	AP, GU, OD: MDM
7	16- 31	2015:A, 2017:A	AP, GU, ANI: MDM
8	31	2017:A	AP, GU, OD, ANI: MDM

Frequent closed itemsets or biclusters for all other datasets can also be obtained in the same way. From Table 7.40, we can conclude that the suffix-forest data structure can efficiently locate bi-clusters as well.

7.6 Identifying rules from the clusters

Rule generation from the biclusters [92, 132, 133] could be useful for analyzing and predicting information related to the forest status. If we consider the closed itemsets in Table 7.39, we may build a large number of association rules by separating the elements into antecedent and consequent. For example, considering the serial number 6, rules could be like, 2009:A, 2011:A → 2017:A, where the supporting objects are states: AP, GU, and OD; or, 2009:A, 2017:A → 2011:A, where the supporting objects are states: AP, GU, and OD; or, 2011:A, 2017:A → 2009:A, where the supporting objects are states: AP, GU, and OD; revealing that the forest cover status of these states is closely related for the mentioned years.

7.7 Estimation of co-related parameters from triclusters

Observing branch 3- 14- 15- 16- 31- AP, GU: MDM; AP: OM; and branch 14- 15- 31- GU, AP, OD: MDM; AP, KE: OM; AP, OD: TM; it is possible to conclude that Odisha’s medium dense forest cover is experiencing an unusual phenomenon. Review of multiple data sources reveals that in 2013 and 2015, Odisha met a loss in forest cover. A survey has shown that 2017 has met the recovery in forest cover loss. In this regard, we want to delineate the relation between forest cover, forest type, and carbon stock which shows the effect of change in forest cover and forest type on carbon stock. Figure 7.7 has shown a pictorial view of reports generated by the Forest Survey of India. The graph shows that Odisha lost a significant amount of its total carbon stock during the year 2013 (indicating the loss of forest cover), which recovered again in 2017. Though Andhra Pradesh also lost its total carbon stock in 2013 than 2004, it is an effect of decreased total geographical area.

Mangroves play an important role in capturing atmospheric carbon and storing it in the form of biomass, soil carbon, and other forms. As a result, mangrove destruction increases atmospheric carbon, decreases carbon stock, and protects coastal areas from cyclonic damage. Though Odisha restored its mangrove cover in 2017, the goal is to have forest cover rather than green cover. As a result, the loss behind the forest must be investigated in or-

der to prevent further loss. So, looking at forest loss, gain, and consistency in forest cover state-wise and year-wise, will help to identify minute details of facts that are difficult to find from datasets only.

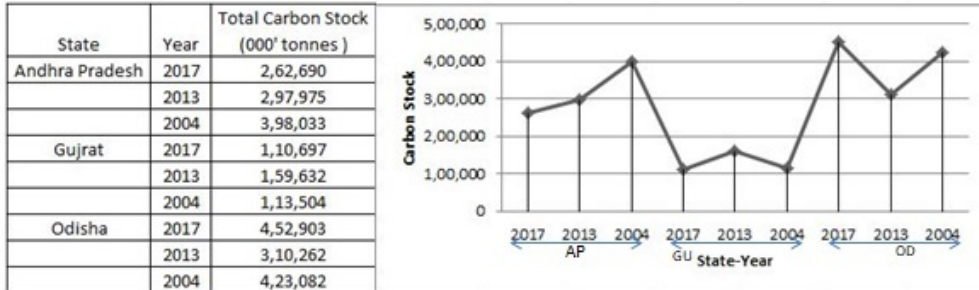


Figure 7.7: Carbon stock change in proportion to forest cover

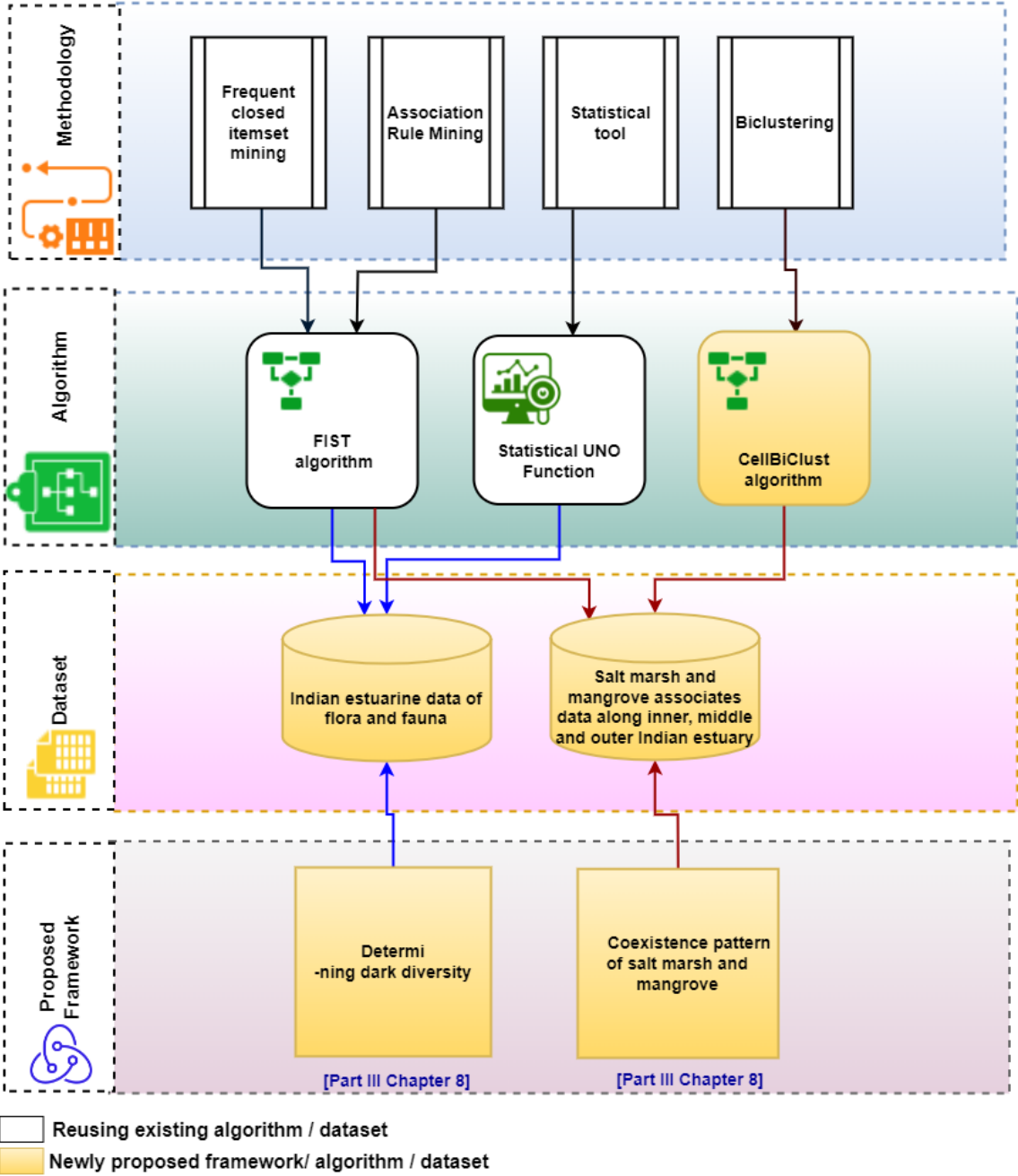
7.8 Summary

The used algorithm for extracting the tri-cluster uses a unique suffix forest data structure that has been proven for effective tri-cluster extraction. Manually curated datasets on forest cover along the time dimension have been presented here. Following the algorithm, presented in [60], the tri-cluster is represented by the nodes of the final suffix forest that we receive after merging the generalized suffix trees, updating, and pruning them sequentially. The goal is to extract more enriched information considering 3-dimensions of the data. In bioinformatics, biclustering has played an essential role in biological sequence analysis. Here, we intend to reveal the usefulness of tri-clusters in biodiversity analysis. In particular, in this work, we tried to establish relationships along varying years, among different types of forest covers, and multiple states.

Part III

Proposed Framework

The methodology, algorithm, and dataset utilised in each proposed framework described in the chapters are shown below:



Outline for Part III

DATA AUGMENTATION USING DARK DIVERSITY FOR FINDING SPECIES ASSOCIATION

8.1	Introduction	95
8.2	Proposed framework	96
8.2.1	Dataset	96
8.2.2	Applying UNO function	97
8.2.3	Normalization and binarization of the dataset	98
8.2.4	Applying data mining methodology	98
8.3	Result and discussion	98
8.4	Summary	102

8.1 Introduction

Computational Biodiversity can broadly be understood as the effort of computational approaches for exploring, interpreting, and analyzing biodiversity data. An enormous load of growing biodiversity data needs algorithmic care for accurate data management, and therefore the term computational biodiversity comes. Instead of relying purely on presence data, the probabilistic forecast of member distribution including the regions of not occurrence can neutralize biodiversity loss by restoring potential ecosystems. This work is aiming at revealing the perspective of computational biodiversity as a counteract for biodiversity loss by correlating the concept of dark diversity. The computation of the dark diversity is accompanied by a data mining algorithm for establishing rules with more nobility to manage the depletion of biodiversity.

Background Study Biodiversity loss is a global threat [134, 135] which implies local loss or reduction of members in an individual habitat. To prevent this loss, endless studies and experiments on behalf of ecological researchers and conservationists are going on and different paths are followed. Ecological theory has plenty of research articles on measuring biological diversity [136, 133] within species, between species as well as between ecosystems. Hypothesis establishment for species richness, abundance, and modeling distribution patterns are attempted numerous times over a range of datasets [137, 138]. The influence of regional species richness on local species, co-existence statistics, diversification factors, etc. are attributed by multiple authors [139, 140]. Species diversity monitoring [141, 142] has been studying to assess the changes that occur over time and implementing suitable measures for different management strategies. Biodiversity hotspot conservation [143, 144] is treated as the most effective way to compensate for the loss. All of these research articles are directed at quantifying the gathered biodiversity information of the recorded individuals, or we can say all of these are dealing with the presence data of the individuals.

Motivation towards dark diversity The concern for the missing part of biodiversity data raises a few more questions like whether a particular site is capable of providing shelter for more species or whether is it possible to expand diversity as well as the richness of species. However absent data could not be estimated merely from survey data. It indicates all the absent members from the total species pool that occur in a particular geographical region having similar environmental conditions [145]. All these absent species are considered to have the potential for being present and expanding diversity. These locally absent species from the regional species pool are termed dark diversity [146, 147]. Quantifying dark diversity is proved to be [148] an additional information source for restoration ecology.

Contribution In this study we have dealt with a dataset of Indian Estuarine Biodiversity consisting of faunal data [110]. It contains species count for each faunal group at different estuaries. It can be noted that multiple estuaries have not even reported any member count from the respective faunal groups. Our objective in this study is to employ the computation of dark diversity upon the dataset before processing rule mining tasks so that all possible likelihoods of faunal presence can be studied. These rules would generate the potential habitat for the faunal groups. In this regard, we would adopt a rule mining task [92] which is solely based upon biclustering [36] and association rule mining [149].

8.2 Proposed framework

8.2.1 Dataset

Indian estuarine data of faunal groups is outlined in the book: “Indian Estuarine Biodiversity” [110]. This data is manually curated and detailed in Table 8.1 in our work. It refers to the number of species counted for each faunal group and stored estuary-wise.

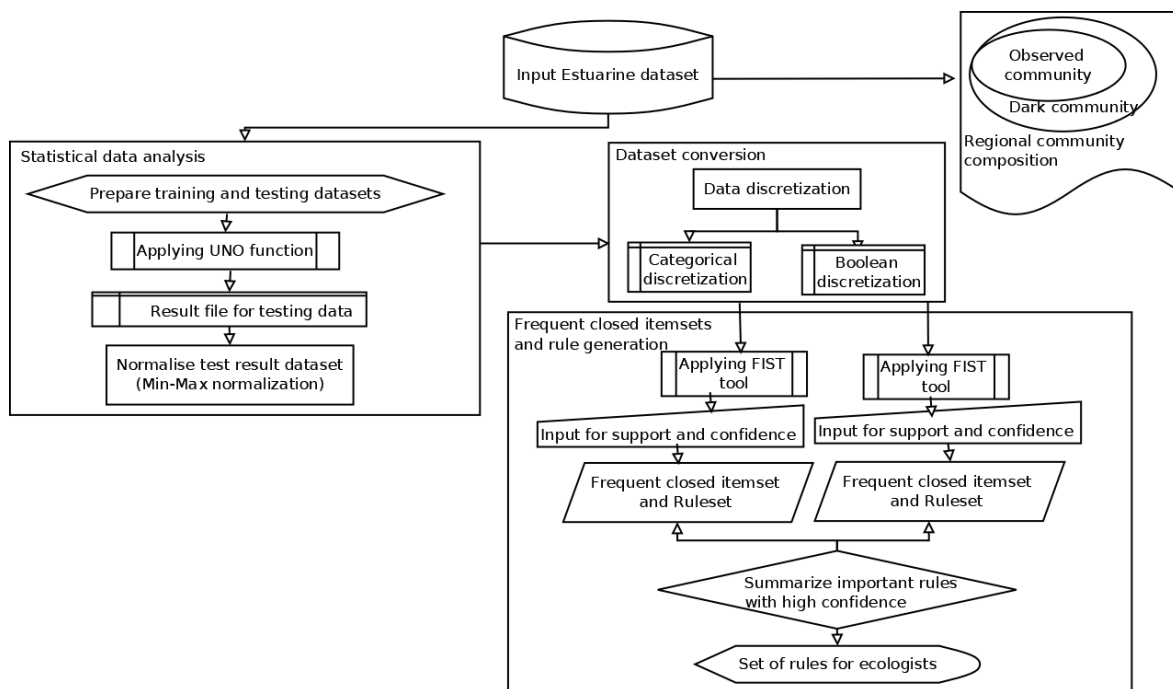


Figure 8.1: Framework of the proposed approach

In this work, the role of dark diversity may be thought of as the task to forecast the occurrence of a faunal group based on the composition pattern of the community. The composition pattern relates to the different faunal groups in a community. It may appear that some specific groups are not detected at all in some particular regions. Using the calibration dataset, the researchers have shown a unique approach (UNO) [150] for predicting probable occurrences in those regions. The most widely used Beals function [151] is already compared to the UNO approach and UNO proves its better efficiency. Therefore, we follow this approach for predicting the dark diversity of our dataset. UNO function uses correspondence analysis for the ordination of species samples. In this regard, ordination [152] is the task of discovering a gradient for the presence/ absence of samples along the geographical region. Primarily, it follows an exploratory analysis technique to reveal the distribution pattern of species. Correspondence analysis in the ordination method uses the weighted average of the data where the species occur and follows one scaling technique among some given choices [153]. UNO function provides 3 different choices and based on applicability, users specify anyone from those. The whole procedure is shown in the flowchart below in Figure 8.1.

Table 8.1: Faunal data at Indian Estuaries:- E1:Hooghly-Matla, E2: Subarnarekha, E3: Baitarani-Brahmani, E4: Mahanadi, E5: Rushikulya, E6: Bahuda, E7: Vamsadhara, E8: Nagavali, E9: Godavari, E10: Krishna, E11: Penner, E12: Ennore, E13: Adyar, E14: Veller, E15: Cauveri, E16: Cochin, E17: Zuari, E18: Mandovi, E19: Tapi, E20: Narmada

Taxonomic group	Major estuaries of India																			
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20
Faunal Groups																				
<i>Protozoa</i>					25	26					20	3		23					4	3
<i>Foraminifera</i>					5					47	11					73		14		
<i>Porifera</i>	1								2								2	2		
<i>Cnidaria</i>	24	12		11	20	5			13	3	10					34	3	3		
<i>Ctenophora</i>	1			1	2						1	1				1				
<i>Rotofera</i>	5								14		2	16		13						
<i>Nematoda</i>	2				11												20			
<i>Acanthocephala</i>									1											
<i>Sipuncula</i>	1	1																		
<i>Mollusca</i>	83	49	19	152	47		28	43	73	103	82	10		11	51	26	40	41	30	32
<i>Annelida</i>	91	37	11	34	19		13	4	70	45				24	48	47	70	3		
<i>Arthropoda</i>	377	53	88	45	159	99	24	17	88	118	125	56	58	35	55	167	72	21	25	60
<i>Bryozoa</i>	4																			
<i>Brachiopoda</i>	1	2								1										
<i>Chaetognatha</i>	4			3	6	2				1	2	3				4	6	6		3
<i>Echinodermata</i>	22	6	1						7	2	1									
<i>Hemichordata</i>	1																			
<i>Urochordata</i>				3	6	4					3					1				
<i>Pisces</i>	314	146	157	177	45	91	64	71	307	268	63	17	135	82	135	126	73	44	64	49
<i>Amphibia</i>	13	3	14							4		3				27				
<i>Reptilia</i>	57	5	45	6	2	1	1	1	1	10		1			4	7	7			
<i>Aves</i>	156	108	269	46	1			52	75	17						45	43	150	23	23
<i>Mammalia</i>	41	2	27	4					1	11		8			2	5				
Total	1198	424	631	478	352	228	130	188	652	630	320	118	193	188	291	560	336	291	146	170

8.2.2 Applying UNO function

We use 75% of the dataset as training data. Then we predict the probable occurrences of not-found classes at multiple estuaries for the remaining 25% of the dataset. UNO method uses any of the 3 different methods such as minobs(), minpred(), and binminpred() to calculate threshold value for predicting abundances. After having tested with all the 3 different methods, it has been found that minobs() and minpreds() are suitable for our case. minpred() uses the smallest predicted value for cases where a class is observed and minobs() uses the smallest positive value observed for a class. Data is first binarized in the case of binminpred(). Results obtained from the UNO function will be preprocessed.

It has been observed that minpred() and minobs() generate quite similar results and they scaled the data in proportion. To justify this observation, a graphical view is generated which shows the rescaled result of the first four classes in Figure 8.2. For each pair of values, the first one is from the observation of minpred().

From the achieved resultset, multiple sites with the potential presence of currently absent members could be found. Comparing with the initial dataset, one can identify the probable occurrence at different locations. All absent sites are not likely for each group under consideration. Like, in the case of *Protozoa*, it can be seen from the original dataset (Table 8.1) that this class is not reported from estuaries E1, E2, E3, and E4. However the application of the UNO function identifies E1 as most likely to be expected and E4 as likely to be expected (labeled as discretization referred to in Table 8.2). Considering *Cnidaria*, their original occurrence data and the data after computing dark diversity are displayed in Figure 8.3, where

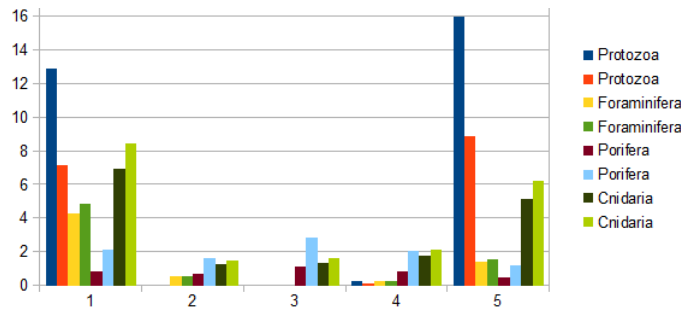


Figure 8.2: Result comparison for minpred()(first one from each pair) and minobs() for the first 4 classes: X-Axis shows 5 estuaries; Y-Axis shows the probability of occurrence

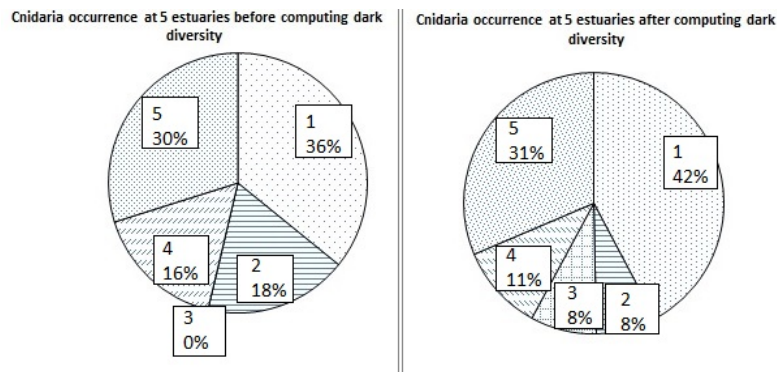


Figure 8.3: Result comparison between before and after computing dark diversity

E3 is turned up to have the probability of appearance.

8.2.3 Normalization and binarization of the dataset

We adopt the dataset obtained by applying minobs() function. Here, we feel the necessity of bringing down all the attributes on the same scale fluctuating between 0 and 1. Thus min-max normalization is performed on the result of UNO Function. We discretize all the obtained values and apply textual leveling to make the dataset comprehensible. The obtained result is presented in Table 8.2. This table also shows the binarized data based on the discretization (1 is replacing all other categorical values except NP, while NP is replaced with 0).

8.2.4 Applying data mining methodology

A combined approach of frequent closed itemset mining and association rule mining [37, 92] is applied to both the categorical discretized data and corresponding binarized data (Table 8.2).

8.3 Result and discussion

In this section, we will assess the consequence of measuring dark diversity before the rule-mining task. Here we justify the usage of the dark diversity function as it replaces most of

Table 8.2: Normalized and discretized dataset; discretization follows: 1= Maximum frequency (MP); $1 > X > 0.60$ = Most likely to occur (ML); $0.60 > X > 0.30$ = Less likely to occur (LL); $0.30 > x > 0.01$ = Least likely to occur (LTL); 0= Not probable (NP)

Class	Min-max normalization					Categorical discretization					Binarization				
	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
<i>Protozoa</i>	0.81	0	0	0.02	1	ML	NP	NP	LTL	MP	1	0	0	1	1
<i>Foraminifera</i>	1	0.12	0	0.06	0.33	MP	LTL	NP	LTL	LL	1	1	0	1	1
<i>Porifera</i>	0.59	0.29	1	0.53	0	LL	LTL	MP	LL	NP	1	1	1	1	0
<i>Cnidaria</i>	1	0	0.03	0.1	0.69	MP	NP	LTL	LTL	ML	1	0	1	1	1
<i>Ctenophora</i>	0.98	0	0.02	0	1	ML	NP	LTL	NP	MP	1	0	1	0	1
<i>Rotofera</i>	1	0.02	0.27	0	0.07	MP	LTL	LTL	NP	LTL	1	1	1	0	1
<i>Nematoda</i>	0.14	0.11	0	0.46	1	LTL	LTL	NP	LL	MP	1	1	0	1	1
<i>Acanthocephala</i>	1	0.7	0.31	1	0	MP	ML	LL	MP	NP	1	1	1	1	0
<i>Sipuncula</i>	1	0.43	0.95	0.25	0	MP	LL	ML	LTL	NP	1	1	1	1	0
<i>Mollusca</i>	0.81	0.42	0	1	0.34	ML	LL	NP	MP	LL	1	1	0	1	1
<i>Annelida</i>	1	0.33	0	0.8	0.17	MP	LL	NP	ML	LTL	1	1	0	1	1
<i>Arthropoda</i>	1	0	0.17	0.01	0.41	MP	NP	LTL	NP	LL	1	0	1	0	1
<i>Bryozoa</i>	1	0.24	0.74	0	0.16	MP	LTL	ML	NP	LTL	1	1	1	0	1
<i>Brachiopoda</i>	1	0.53	0.81	0.53	0	MP	LL	ML	LL	NP	1	1	1	1	0
<i>Chaetognatha</i>	0.96	0	0.32	0.1	1	ML	NP	LTL	LTL	MP	1	0	1	1	1
<i>Echinodermata</i>	1	0.33	0.74	0.19	0	MP	LL	ML	LTL	NP	1	1	1	1	0
<i>Hemichordata</i>	1	0.24	0.74	0	0.16	MP	LTL	ML	NP	LTL	1	1	1	0	1
<i>Urochordata</i>	0.46	0	0	0.19	1	LL	NP	NP	LTL	MP	1	0	0	1	1
<i>Pisces</i>	1	0.36	0.32	0.53	0	MP	LL	LL	LL	NP	1	1	1	1	0
<i>Amphibia</i>	1	0.15	0.7	0	0.06	MP	LTL	ML	NP	LTL	1	1	1	0	1
<i>Reptilia</i>	0.92	0.28	1	0	0.03	ML	LTL	MP	NP	LTL	1	1	1	0	1
<i>Aves</i>	0.76	0.37	1	0.18	0	ML	LL	MP	LTL	NP	1	1	1	1	0
<i>Mammalia</i>	1	0.21	0.85	0	0.07	MP	LTL	ML	NP	LTL	1	1	1	0	1

the null entries in the dataset, followed by extracting additional information. To accomplish this, we attempt to mine twice; on the binarized dataset obtained corresponding to the initial table and the datasets after applying the dark diversity function (Table 8.2). We carry out a comparison of the data generated in both cases. It is found that they differ significantly. Figure 8.4 determines the difference in the number of rules and confirms that larger rules are brought about where the dark diversity function is used. Here, we look at the scenario for the 5 estuaries that are taken as testing data. In Figure 8.4, P_MME , P_MMAS , and P_MMAP denote the preceding scenario, i.e. the rules generated without using the dark diversity function. F_MME , F_MMAS , and F_MMAP denote the rules generated by exploiting the dark diversity function. Here, MME corre-

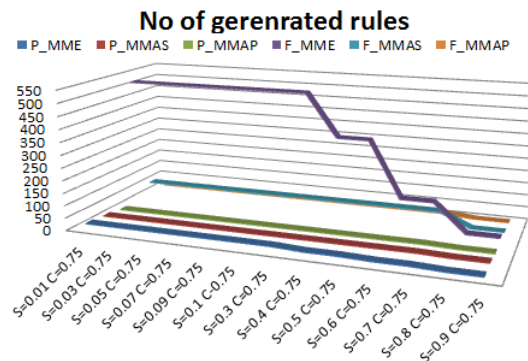


Figure 8.4: Study on the no of rules generated in the previous case and following the case of incorporating dark diversity function; x-axis: Threshold of support and confidence; y-axis: no of rules

Table 8.3: FCI obtained from the original and derived dataset

ClosedSet	Support	Object list
Result obtained from the original binarized dataset before applying dark diversity function		
FCI1: [E1=1]	19	3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23
Result obtained from derived dataset after applying dark diversity function		
FCI2: [E1=MP]	14	2, 4, 6, 8, 9, 11, 12, 13, 14, 16, 17, 19, 20, 23
FCI3: [E1=ML]	6	1, 5, 10, 15, 21, 22
FCI4: [E1=LL]	2	3, 18

sponds to exact rules with support value 1. MMAS and MMAP both conform to approximate rules where the support value is always greater than zero, asserting the possibility of appearing.

Knowledge retrieved by employing computational biodiversity: FCIs are able to find a large set of items that are taking place more times than user-specified support values in a dataset. Hence, FCIs are capable of deriving useful knowledge. Also, exact rules are derived from the facts that are demonstrated in the dataset, whereas approximate rules deal with the probability of occurring. Below, in this section, we are moving to discuss the information that we may derive from the FCIs and rules.

- Table 8.3 shows the resultset consisting of FCIs, before and after applying the dark diversity function. It is explicit from the available dataset that 19 members are identified to have habitat at Hooghly-Matla (E1) estuary. But the adaptation of the dark diversity function mentions 22 total members that may occur. Thus, information related to 6 more faunal data is obtained by having the occurrence data categorized as maximum frequency, most likely to occur, and least likely to occur. The illustration is given below,
 - Performing FCI2 - FCI1: 2, 8; It indicates that *Foraminifera* and *Acanthocephala* are the maximum probable species missing at E1.
 - Performing FCI3 - FCI1: 1; It shows that *Protozoa* is the most likely faunal group that should exist at E1.
 - Performing FCI4 - FCI1: 3; It suggests that *Porifera* may have the chance of occurrence, but it is less likely to occur.

The diversity of *Foraminifera* is inadequate. It is less available as its existence is reported merely from 5 estuaries (Rushikulya, Krishna, Penner, Cochin, and Mandovi). Being highly sensitive to physicochemical characteristics, this group can be chosen as an indicator of oceanic and climatic information. On that account, the resulting rule serves as noteworthy information from the ecological perspective. *Acanthocephala* is reported from only Godavari estuary. As Hooghly-Malta is a highly diverse estuary, it may be the apparent site next to Godavari estuary for this parasite. For a similar reason, *Protozoa* and *Porifera* can also be expected at estuary Hooghly-Malta.

- Findings based on presence-only data are noted in Table 8.4, where the antecedent part is the same for all the cases. Comparing the consequent and object list, we can draw the following conclusions:

Table 8.4: Rules (similar estuary list) from binarized derived dataset

Rule	Antecedent	Consequent	Sup	Conf	Estuary list
R1	[S6=1, S13=1, S17=1, S20=1, S21=1, S23=1]	[S5=1, S12=1, S4=1, S15=1]	3	0.75	[1, 3, 5]
R2	[S6=1, S13=1, S17=1, S20=1, S21=1, S23=1]	[S2=1, S7=1, S10=1, S11=1]	3	0.75	[1, 2, 5]
R3	[S6=1, S13=1, S17=1, S20=1, S21=1, S23=1]	[S3=1, S8=1, S9=1, S14=1, S16=1, S19=1, S22=1]	3	0.75	[1, 2, 3]

Table 8.5: Rules (Homogeneous antecedent) from binarized derived dataset

Rule	Antecedent	Consequent	Sup	Conf	Estuary list
R1	[S6=1, S13=1, S17=1, S20=1, S21=1, S23=1]	[S3=1, S8=1, S9=1, S14=1, S16=1, S19=1, S22=1]	3	0.75	[1, 2, 3]
R2	[S3=1, S8=1, S9=1, S14=1, S16=1, S19=1, S22=1]	[S6=1, S13=1, S17=1, S20=1, S21=1, S23=1]	3	0.75	[1, 2, 3]

- Estuary list (R1) - Estuary list (R2) = 3; also Antecedent R1 is the same as the Antecedent of R2. It implies that estuary 3 has the probability of finding all the faunal groups numbered S2, S7, S10, and S11 with a confidence level of 75% those are in the consequent part of R2.
- Oppositely, Estuary list (R2) - Estuary list (R1) = 2, and the antecedents of R1 and R2 are matched. So, like before, it can be estimated that estuary 2 has the probability of sustaining all the faunal groups present in the consequent part of R1 i.e. the groups numbered S5, S12, S4, S15 with a confidence level of 75%.
- Similarly, from rules R1 and R3, it can be stated that estuary 5 has the likelihood of occurrence for S3, S8, S9, S14, S16, S19, and S22 with a confidence level of 75%.
- Rule R2 and R3 could derive that estuary 3 has the chance to exhibit the groups numbered S3, S8, S9, S14, S16, S19, and S22 with a confidence level of 75%.
- Table 8.5 displays two rules where the object list is the same for all the cases. Comparing the antecedent and the consequent, we can claim that the antecedent and consequent parts are highly associated as they complement each other. These faunal groups are building a closed set with a 75% confidence level.
- An identical example is shown in Table 8.6. The estuary list is the same for all. It can be observed that antecedent and consequent together form a closed group. The occurrence of any class is linked to all others belonging to the same group with 75% confidence.

It has been found that [110] E1 (Hooghly-Matla) is a highly diverse estuary. Again, the estuaries E2 (Subarnarekha) and E3 (Baitarani-Brahmani); E4 (Mahanadi), and E5 (Rushikulya) are situated side by side and those are closely related regions. Therefore, they are expected to have similar physico-chemical properties. So, closely associated member lists are found which are justified in Tables 8.5,8.6.

Therefore, through this study, we have shown how data mining knowledge discovery can be associated with ecological research and discover new diversity patterns with adequate explanation. We have shown that with the help of the algorithmic solution, we can estimate the likelihood of being present in a locally absent faunal group and verify our findings from the standpoint of ecology. These data would certainly assist ecologists in practicing

Table 8.6: Rules from binarized derived dataset with similar object lists

Rule	Antecedent	Consequent	Sup	Conf	Estuary list
R1	[S4=1]	[S1=1, S18=1, S2=1, S7=1, S10=1, S11=1, S15=1]	3	0.75	[1, 4, 5]
R2	[S15=1]	[S1=1, S18=1, S2=1, S4=1, S7=1, S10=1, S11=1]	3	0.75	[1, 4, 5]
R3	[S2=1]	[S1=1, S18=1, S4=1, S7=1, S10=1, S11=1, S15=1]	3	0.75	[1, 4, 5]
R4	[S7=1]	[S1=1, S18=1, S2=1, S4=1, S10=1, S11=1, S15=1]	3	0.75	[1, 4, 5]
R5	[S10=1]	[S1=1, S18=1, S2=1, S4=1, S7=1, S11=1, S15=1]	3	0.75	[1, 4, 5]
R6	[S11=1]	[S1=1, S18=1, S2=1, S4=1, S7=1, S10=1, S15=1]	3	0.75	[1, 4, 5]

ecological restoration through habitat and range improvements for species under study. Contrarily it could also be stated that it nourishes the sustainability of an ecosystem under surveillance via assimilating new species.

8.4 Summary

This study incorporates data mining along with statistics and directs us toward a competent solution for biodiversity restoration specific to a particular region of study. This study has introduced the proposition of applying the dark diversity function to the presence-absence dataset before the process of rule mining. The reason behind this is to gain information related to the absent part of the occurrence data. The usefulness of deploying the dark diversity function is illustrated by visualizing the number of rules generated with and without applying the dark diversity function. It is understood that the underlying reason behind the greater number of rules is more non-zero values in the dataset. Our study helps in proper management in a survey or re-survey aiming at finding new sites for probable habitats for a particular faunal group. The generated results can suggest the likelihood of occurrence for specific faunal groups in a degraded estuary for the introduction of the members of the specific group or accelerating the restoration process.

CHAPTER 9

MANGROVE REGENERATION FRAMEWORK USING FREQUENT CO-EXISTENCE PATTERN

9.1	Introduction	104
9.2	Materials	108
9.2.1	Study area	108
9.2.2	Data gathering and preprocessing	110
9.2.3	Dataset description	111
9.3	Methods	112
9.3.1	Proposed framework	112
9.3.2	Validation through statistical approach: multidimensional scaling (MDS)	112
9.4	Result	114
9.4.1	Finding FCI on <i>BSM</i> dataset	114
9.4.2	Association rule generation on <i>OEBM</i> , <i>MEBM</i> , <i>IEBM</i>	117
9.5	Discussion	121
9.5.1	Finding inferences from exact association rules	122
9.5.2	Predicting novel associations from approximate association rules	124
9.5.3	Implications on restoration practitioners	126
9.5.4	Comparisons with parallel studies in mangrove restoration in terms of used methodology, findings, and limitations	127
9.6	Summary	130

9.1 Introduction

Climate-change-driven sea level rise causes an increase in salinity in coastal wetlands accelerating the alteration of the species composition. It triggers the gradual extinction of species, particularly the mangrove population which is intolerant of excessive salinity. Thus despite being crucial to a wide range of ecosystem services, mangroves have been identified as a vulnerable coastal biome. Hence restoration strategy of mangroves is undergoing rigorous research and experiments in literature at an interdisciplinary level. From a data-driven perspective, analysis of mangrove occurrence data could be the key to comprehending and predicting mangrove behavior along different environmental parameters, and it could be important in formulating a management strategy for mangrove rehabilitation and restoration. As salt marshes are natural salt-accumulating halophytes, mitigating excessive salinity could be achieved by incorporating salt marshes in mangrove restoration activities. This study intends to find a novel restoration strategy by assessing the frequent co-existence status of salt marshes, with the mangroves, and mangrove associates in different zones of degraded mangrove patches for species-rich plantation. To achieve this, we primarily design a novel methodological framework for the practice of knowledge discovery concerning the coexistence pattern of salt marshes, mangroves, and mangrove associates along with environmental parameters using a data mining paradigm of association rule mining. The proposed approach has the capability to uncover underlying facts and forecast likely facts that could automate the study in the field of ecological research to comprehend the occurrence of inter-species relationships. Our findings are based on published data gathered on the Sundarban Mangrove Forest, one of the world's most important littoral forests. The existing literature reinforces the findings that include all the sets of frequently co-occurring mangroves, their associates, and salt marshes along the salinity gradient of coastal Sundarbans. A detailed understanding of the occurrence patterns of all these, along with the environmental variables, would be able to promote a decision-making strategy. This framework is effective for both academia and stakeholders, especially the foresters/ conservation planners, to regulate the spread of salt marshes and the restoration of mangroves as well.

Mangrove restoration becoming a global issue In spite of the great ecological and economic services of the mangrove ecosystem, it is estimated that up to 35% of the mangrove area on the Earth has been lost since the 1980s, mostly as a result of various developmental activities on the coast [8, 9, 10]. However, concern over the disappearance of coastal mangrove regions has recently grown significantly [9]. These coastal habitats are said to be most fragile due to their susceptibility to climate change vis-a-vis sea level rise. The continuous loss and fragmentation of such habitats hinder species migration/ dispersal, obliterate local coastal resilience, and drive essential mangrove ecosystems into collapse [9].

The Sundarban, the largest deltaic region in the world is a home of nurturing wet coastal biodiversity, especially the mangroves. Sundarban spreads over both India and Bangladesh of which 40% (nearly 4000 sq km) covers the Indian part [154]. Besides ecological functions, it significantly supports socioeconomic stability and local livelihoods. The unique plant communities of the mangrove ecosystem thrive in a wide range of saline tidal inundation with a clay-silty loose substratum. Many mangrove species are intolerant of higher salt, while several others have a likeness to it [155, 156]. Periodic saltwater

inundation affects physiological functions such as germination, seedling growth, reproduction, rate of transpiration, and so on. Although moderate salinity is essential to stimulate the growth of mangroves, high salinity may cause adverse effects [157] and alters mangrove demography by causing toxicity-induction, and, in some cases, plant death [158]. At its most extreme level of degeneration, the homeostasis of the mangrove ecosystem fails to the point where it approaches extinction. [12] firmly shows saline intrusion as the main factor causing the deterioration of such littoral forests. From the river's edge to the landward side, salinity declines under the control and maintenance of tidal flushes. Many forests are devoid of some species because they are more sensitive to greater salinity levels. Salt-sensitive *Heritiera fomes* and *Phoenix paludosa*, for example, disappeared entirely from many blocks in the Indian Sundarban regions as a result of rising salinity [11, 12, 13, 14, 15, 16, 17]. *Phoenix paludosa* is currently categorized as near threatened and *Heritiera fomes* is listed as endangered by the International Union for Conservation of Nature [159, 160]. Contrarily, *Excoecaria agallocha* and *Avicennia* spp. expand predominantly into degraded forests due to their high resilient capability [161, 12, 162]. Thus the rapid alteration and loss of mangrove ecosystems are becoming a major global problem [160, 163, 12, 17]. In consequence, the mangrove habitat restoration is really an imperative issue for the sustenance of such a fragile ecosystem.

Role of salt marsh in mangrove restoration In subtropical coastal wetlands, an ecotone exists between salt marshes and mangroves [164, 165, 163]. Despite the vegetative variations of saltmarshes (which are dominated by herbaceous vegetation like forbs and clonally growing graminoids [166]), and mangroves (which are characterized by trees with a limited herbaceous vegetation [167, 160]), both co-exist in mostly similar physical conditions (dynamic intertidal zones), playing a pivotal role in establishing different ecological niches of the tidal wetland habitats [164, 168].

As the salt marshes are salt-accumulating halophytes [169] and grow in high-saline intertidal mudflats, they play an important functional role in the colonization of several species of mangrove [170, 171, 12, 163]. Salt marshes with their special physiological functions and morphological adaptation remove the extra salt from the soil and provide suitable habitats for many mangrove species of a varied range of salt tolerance. As the salt marshes are growing mostly in saline encrusted soil and gradually lessen the salt from the soil, there is a good possibility of the successional association of different species of mangroves and salt marshes [170, 171, 172]. Besides the salt-accumulating nature, the early colonization by saltmarsh vegetation has a significant facilitative effect on mangrove vegetation in saltmarsh-mangrove ecotones [173, 174]. There are two basic ways that salt marsh vegetation functions as a nurse species [163]. First, mangrove seedlings are physically trapped by the dense salt marsh vegetation [175]. The marsh species *Sesuvium portulacastrum* promotes mangrove recolonization by trapping the propagules of *Rhizophora* spp., *Avicennia* spp., while tidal exchange rapidly disperses mangrove seedlings on bare surfaces [176]. Second, salt marshes hinder incoming hydrodynamic energy, creating relatively calm physical conditions for mangrove species to root and structural support for mangrove seedlings that need to maintain an upright posture [163].

Therefore, in order to maintain, restore, and develop new coastal wetlands in the coming decades, it is essential to recognize the colonization establishment among the mangroves and salt marshes [163]. Although the behaviors between these foundation species have

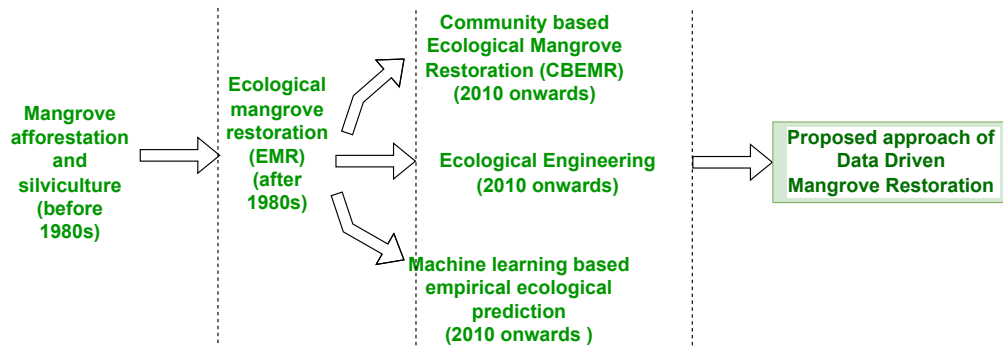


Figure 9.1: Major types of mangrove rehabilitation and restoration techniques followed in literature

been examined in a few works [177, 163], little research has been conducted to explore the co-existence pattern and potential interactions at varying environmental parameters from a regeneration and management standpoint.

Previous attempts made for mangrove restoration practices Restoration of ecosystems is becoming a more valuable technique in systems that have been destroyed, degraded, or agitated by both natural and human disturbances. Mangroves and salt marsh ecosystems have been subjected to a higher level of widespread disruption and loss [178, 179]. To compensate for this loss, large-scale coastal restoration activities have been attempted all over the world [180, 181, 172, 16, 182] through mangrove rehabilitation and restoration. Recovery and restoration techniques can broadly be categorized [183] (Figure 9.1) in 5 major directions.

1. Prior to the 1980s afforestation for silviculture was followed, adopting monoculture plantation [184, 185].
2. After that Ecological mangrove restoration (EMR) approaches, and later Community Based Ecological Mangrove Rehabilitation (CBEMR) has been adopted via including local communities. From the aspect of socio-ecology, Ecological Mangrove Rehabilitation (EMR) has been adopted to incorporate local people into the environment [186]. EMR methods involve altering the intertidal zone (e.g. dredging, filling) to keep biophysical conditions, notably inundation, within acceptable ranges for mangrove establishment, survival, and reproduction [187, 181, 188].
3. Eco-engineering approaches combines ecological concepts with modern environmental engineering technique. An example of an eco-engineering framework is the incorporation of mangroves into engineered hard coastal defense structures [189, 190, 191, 192].

4. Machine learning-based applications have been followed in parallel with the eco-engineering approach. This data-driven approach is another emerging trend having an integral role in this aspect. Multiple knowledge-based classifiers in machine learning have recently been used for the classification of mangrove species [193, 194, 195], identifying the changes in mangrove cover along time dimension [196, 197]. Mainly machine learning algorithms have been employed in this category of study to solve complicated analytical problems correctly without human intervention.

The proposed data mining approach is also a data-driven model that identifies previously unidentified, practically applicable, and easily accessible interpretations of knowledge required for decision-making. Examples of the application of data mining algorithms on mangroves in ecology are rare [48].

Potential for data mining approach: A data-driven adaptive management for mangrove rehabilitation Previous studies have shown that multiple efforts have been made for the rehabilitation and restoration of mangroves at the interdisciplinary levels [198]. It has been found that the loss or gain in mangrove cover, classification of mangrove cover, and restoration of mangroves by employing multiple eco-engineering approaches have been investigated by multiple research works. However, systematic research on the growth pattern of dominating species of salt marsh and mangrove in the coastal ecosystems is still unveiled. Since the groundbreaking work of [45] on mangroves and [199] on salt marshes, it is remarkable that this area has received such little attention. Studying the presence/absence data at the species level could hypothesize the process of ecosystem functioning, and could formulate better decisions to achieve better conservation policies. This could be thought of as data-driven adaptive management for mangrove restoration which includes the method of formulating hypotheses about the growth pattern of mangroves and salt marshes, afforestation via planting, observing outcomes, relating them to predictions, and altering decisions to more effectively accomplish conservation goals through enhanced understanding of ecological processes.

Therefore, we would like to employ a novel strategy for ecological rehabilitation and restoration by incorporating a data mining approach. Association rule mining is one of the distinguished data mining approaches and is well-known in the domain of market basket analysis [200], bioinformatics [132, 201] for analyzing data at the granular level. But in the ecological study, the application of association rule mining is very scarce. Although, the usefulness has already been justified in ecology [32, 62]. The approach of frequent closed itemset mining and rule mining [33, 37] can extract the frequent distribution patterns of mangroves, mangrove associates, and salt marshes with different environmental parameters. Both approaches accept records in two-dimensional matrix form. Thus records of species presence along sites can be simply converted into a two-dimensional matrix, with each site corresponding to a row index. The cells for a row denote the status of the considered species along with the columns. As a result, determining the frequently co-existing species in various sites could be an ecology-relevant query related to the task of frequent closed itemset mining. Alternatively, it could be, which sites exhibit a similar frequent growth pattern. Again, using rule interestingness measurements, the output of association rule mining can be used to uncover significant relationships between species. As a result, rules that are ecologically significant and valuable could be derived.

Contribution Ecological restoration entails rejuvenating native ecosystems in vulnerable areas while preserving the diversity of local flora and fauna through regeneration with a much shorter regeneration time. It takes longer for natural regeneration to occur. So, precisely, we could state that the main contributions are,

- Proposal of an excessive salinity-affected mangrove community restoration approach where hyper-salinity could be neutralized by growing suitable salt marshes.
- Case study on Sundarban coastal area considering major environmental/habitat factors, such as salinity, pH, soil texture, tidal amplitude, along with the occurrence data of mangroves, mangrove associates, and salt marshes and compile 3 different datasets for inner, middle, and outer estuarine species records.
- Establishing
 - salt marsh-salt marsh co-existence pattern along the salinity gradient
 - salt marshes, mangroves, and mangrove associates co-existence patterns with varying environmental factors
 - probable inter-species association from present co-existence data

Our overarching aim was to advance the understanding of the frequent co-existence pattern of salt marshes with mangrove, and mangrove associates along the salinity gradient. As the salt marshes can remediate excessive salinity from the soil, mangroves could be regained in their natural habitat. Hence, this study provides valuable information for selective plantations to coastal scientists and restoration practitioners. This analytical approach demonstrates a feasible way for the study-based suitable multi-species heterogeneous mangrove afforestation that could enrich the species, ecosystem, and overall biodiversity as well. This framework could be utilizable for other datasets also.

9.2 Materials

9.2.1 Study area

Blocks in Sundarban The study area is the The Indian part of Sundarban covering an area of almost 4000 sq km and lies between 21°13′ - 22°40′ North latitude and 88°05′ - 89°06′ East longitude [202]. The physiographic division of this part can be viewed as the outer estuary, middle estuary, and inner estuary [54] (Figure 9.2). According to [54], 22 blocks are identified that fall under these divisions. Pirkhali, Jhilla (Northern blocks); Arbesi, Khatuajhuri, Harinbhanga (Eastern blocks), and Goashaba (Central block) form the inner estuarine region. Panchmukhani (Northern block); Chamta, Chandkhali (Central blocks); Matla, Netidhopani (Western blocks); and Ajmalmari, Dhulibhasani (blocks of south 24 Parganas) are forming the middle estuarine region. The outer estuarine region is formed by Bagmara, Gona, Mayadwip (Southern blocks); Chulkati, Thakuran, Saptamukhi, Muriganga (blocks from South 24 Parganas); Chottohardi (Western part).

The outer, middle, and inner estuarine regions have distinctive features causing the zonation of the different types of mangroves and their associated plant communities [54, 179, 55]. Soil texture, pH, duration of tidal inundation, tidal water level, water salinity, and the mixing of freshwater and seawater are all key factors in mangrove zonation.

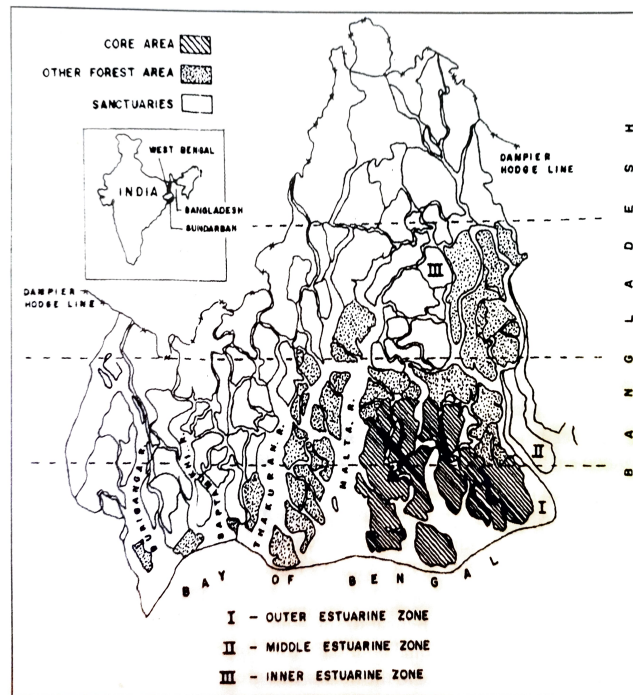


Figure 9.2: Zone wise division for Sundarban biosphere reserve (Collected from [54])

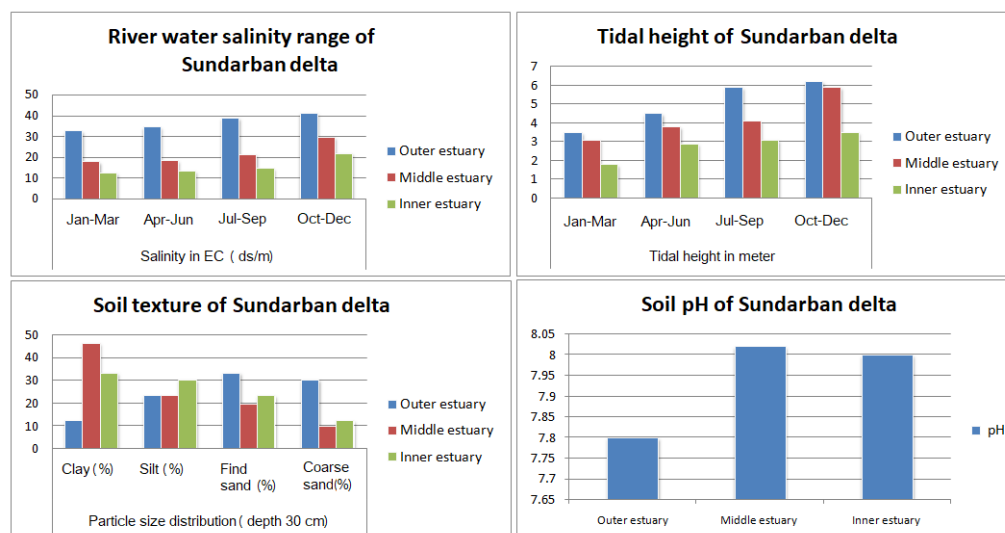


Figure 9.3: Graphical representation of numerical data for water salinity, tidal amplitude, soil texture and soil pH of Sundarban delta

9.2.2 Data gathering and preprocessing

Figure 9.4 summarises the steps for creating the input datasets. Outer estuarine blocks, middle estuarine blocks, and inner estuarine blocks were used to organize the species occurrence data. The data is mostly obtained from [203, 179, 55], which provides information on dominant species as elevation and distance from the estuary increase. In addition, we assess a few other growth-impacting variables (salinity, pH, tidal amplitude, soil texture) for the occurrence data, as shown in Figure 9.3. Rather than using quantitative data, we categorize the variables and assign suitable labels for better readability by decision-makers/other domain users. As shown in Table 9.1, each variable is listed in the column Indicator and has been divided into a number of categories (which are considered in generating association rules).

Electrical conductance (EC) is used to measure salinity, with decisiemens per meter (dS/m) as the unit of measurement. The following equations, 9.1 and 9.2, are used to convert it to TDS (Total dissolved solids made up of salts) [204].

$$TDS \text{ (mg/L or ppm)} = EC \text{ (dS/m)} \times 640, \quad (9.1)$$

where, $5 \text{ dS/m} > EC > 0.1 \text{ dS/m}$

$$TDS \text{ (mg/L or ppm)} = EC \text{ (dS/m)} \times 800, \quad (9.2)$$

where, $EC > 5 \text{ dS/m}$

After getting the TDS in ppm (parts per million), the data is labeled based on the salinity content [205] (Table 9.1). The soil texture triangle [206, 207] was used to label the data on soil texture. The data is labeled using the textual groups provided in the triangle based on the percentages of clay, silt, and sand (fine sand and coarse sand) (Table 9.1). Similarly, based on the study made in [208], soil pH data has been given some descriptive names (Table 9.1). For the tidal amplitude, the average range has been considered exhibited for all the estuaries throughout the year. Based on the average tidal amplitude, the outer, middle, and inner estuarine tidal levels could be discriminated (Table 9.1). Mangrove occurrence data can also be categorized as abundant, regular, infrequent, and rare depending on the percentage of species occurrence data [54] and it is shown in Table 9.1.

Table 9.1: Indicators and their categories considered in generating association rules

Indicator	Measure	Category/ Label	Definition	Estuary
Salinity	TDS in ppm	Freshwater	TDS up to 1.5	IE
		Brackish water	TDS 1.5 - 10	ME
		Seawater	TDS 10 - 45	OE
Soil texture	Particle distribution in percentage	Sandy loam	12.5% clay, 23.5% silt, and 64% sand	OE
		Clay	46.5% clay, 23.5% silt, and 30% sand	ME
		Clay loam	33.5% clay, 30.5% silt, and 36% sand	IE
Soil pH	pH unit	Slightly alkaline	7.4 to 7.8	OE
		Moderately alkaline.	7.9 to 8.4	ME, IE
Tidal amplitude	Height in meter	Maximum	5.02	OE
		Higher	4.22	ME
		Moderate	2.82	IE
Species occurrence	Presence Percentage	Abundant	76% - 100%	NA
		Frequent	46% - 75%	NA
		Occasional	16% - 45%	NA
		Rare	up to 15%	NA

Table 9.2: Structure of Experimental Databases

Datasets	#Rows	#Columns			
	# Blocks	# Salt marshes	# Mangroves	# Mangrove Associates	# Other factors
<i>BSM</i>	22	11	-	-	-
<i>OEBM</i>	8	11	5	0	4
<i>MEBM</i>	7	11	12	4	4
<i>IEBM</i>	7	11	16	7	4

Table 9.3: List of salt marshes, mangroves, mangrove associates, and other factors under study

Salt marshes	Mangroves	Mangrove associates
<i>Aeluropus lagopoides</i>	<i>Aglaia cuculata</i>	<i>Acanthus ilicifolius</i>
<i>Heliotropium curassavicum</i>	<i>Aegialitis rotundifolia</i>	<i>Acanthus volubilis</i>
<i>Salicornia brachiata</i>	<i>Aegiceras corniculatum</i>	<i>Brownlowia tersa</i>
<i>Sesuvium portulacastrum</i>	<i>Avicennia alba</i>	<i>Cerbera odollam</i>
<i>Suaeda maritima</i>	<i>Avicennia officinalis</i>	<i>Crinum defixum</i>
<i>Suaeda nudiflora</i>	<i>Avicennia marina</i>	<i>Clerodendrum inerme</i>
<i>Tamarix dioica</i>	<i>Bruguiera sexangula</i>	<i>Cynometra ramiflora</i>
<i>Tamarix gallica</i>	<i>Bruguiera cylindrica</i>	<i>Cyperus exaltatus</i>
<i>Tamarix troupii</i>	<i>Bruguiera parviflora</i>	<i>Derris trifoliata</i>
<i>Trianthema portulacastrum</i>	<i>Bruguiera gymnorhiza</i>	<i>Derris scandens</i>
<i>Trianthema triquetra</i>	<i>Ceriops decandra</i>	<i>Fimbristylis ferruginea</i>
	<i>Ceriops tagal</i>	<i>Finlaysonia obovata</i>
Other factors	<i>Excoecaria agallocha</i>	<i>Intsia bijuga</i>
Soil texture	<i>Heritiera fomes</i>	<i>Myriostachya wightiana</i>
pH	<i>Kandelia candel</i>	<i>Pentatropis capensis</i>
Salinity	<i>Nypa fruticans</i>	<i>Porteresia coarctata</i>
Tidal amplitude	<i>Phoenix paludosa</i>	<i>Sarcolobus globosus</i>
	<i>Raizophora apiculata</i>	<i>Scirpus littoralis</i>
	<i>Raizophora mucronata</i>	<i>Tylophora tenuis</i>
	<i>Sonneratia apetala</i>	
	<i>Sonneratia caseolaris</i>	
	<i>Sonneratia griffithii</i>	
	<i>Xylocarpus granatum</i>	
	<i>Xylocarpus mekongensis</i>	

9.2.3 Dataset description

Table 9.2 summarizes the structures for the input datasets. The occurrence records of 11 salt marshes along all the blocks are summarized in *BSM* (dataset of blocks versus salt marshes), where the rows represent the 22 blocks and the columns represent the 11 salt marshes. Besides the salt marsh records, three more datasets have been generated for inner, middle, and outer estuarine blocks [54]. These three datasets are denoted as *OEBM*, *MEBM*, and *IEBM* (Outer Estuarine Blocks Mangrove, Middle Estuarine Blocks Mangrove, and Inner Estuarine Blocks M, respectively) where each of these contains the presence record of the estuary-specific distinct salt marshes, mangroves, and mangrove associates data, and other environmental parameters (such as salinity, pH, soil texture, and tidal amplitude) across the columns. The rows represent the identified blocks for the outer, middle, and inner estuarine regions. The list of salt marshes, mangroves, and mangrove associates, that have been considered in this study, are listed in Table 9.3. The datasets we prepare are provided as supplementary material.

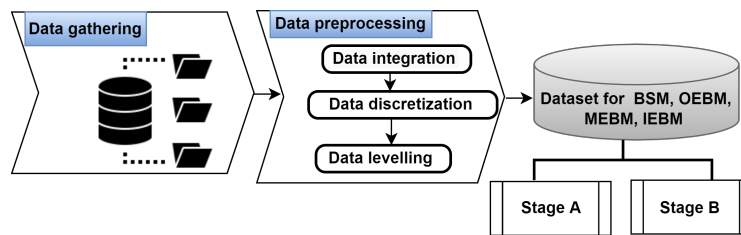


Figure 9.4: The proposed framework

9.3 Methods

9.3.1 Proposed framework

The proposed framework is shown in Figure 9.4. After data gathering and preprocessing, the next steps are summarised in Stage A and Stage B. Figures 9.5 and 9.6 depicts the procedures taken to accomplish Stage A and Stage B, respectively.

1. Stage A (Figure 9.5) consists of applying the data mining methodology to *BSM* dataset and extracting the frequent closed itemsets, as well as statistical analysis, visualization, and validation of the outcome.
2. Stage B (Figure 9.6) performs the knowledge discovery on all the datasets (*BSM*, *OEEM*, *MEBM*, and *IEBM*).

It follows the phases of algorithm application for rule mining (consisting of frequent closed itemset mining and rule development), result interpretation, discussion, and validation (from the domain expert 's perspective).

A descriptive data mining model, FIST [92] has been employed to extract the intrinsic structure and relationships of the data. In stage A (Figure 9.5), FIST was applied to a smaller dataset of salt marshes. The generated result was validated statistically in this stage only. Next, a similar approach has been used (Figure 9.6) upon homogeneous types of bigger datasets of salt marshes, mangroves, and mangrove associates. Here, the focus is solely on the discovery of knowledge from those larger pooled datasets.

Species association: a way for measuring relationship, and co-existence pattern among the species From the data mining perspective, the species association can be considered as an association rule. The species association represented using the association rule is a way to assess the relationship among species. It can be computed from their occurrence record and indicates real behavioral phenomena.

9.3.2 Validation through statistical approach: multidimensional scaling (MDS)

Multidimensional scaling (MDS) [209] visualizes the level of similarity/ dissimilarity, by the relative positions on a map, between any two objects in the dataset in two or three-dimensional pictures. For example, given a matrix of perceived proximity of various objects,

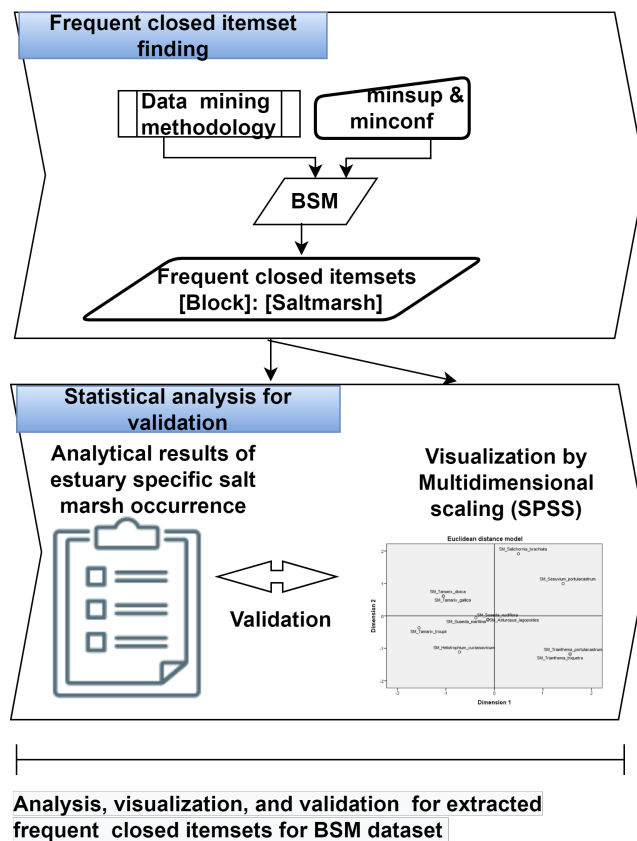


Figure 9.5: Stage A of the proposed framework

MDS plots the objects on a map in such a way that those having higher similarities are positioned near each other on the map. The objects exhibiting dissimilarity, are put far away from each other. MDS has been employed by adopting the ALSCAL (Alternating Least Squares Scaling) method of SPSS in our study. SPSS ALSCAL exploits the Euclidean model as a fundamental operation to figure out the optimal distances between objects in n-dimensional space [210]. The distance function for the Euclidean model is given in equation 9.3:

$$Distance_{ij} = \sqrt{\sum_{k=1}^n (P_{ik} - P_{jk})^2} \tag{9.3}$$

where $Distance_{ij}$ represents the squared euclidean distance between two points P_i and P_j . Here, ik and jk are the respective coordinates of axis k.

Validation of this MDS is measured via stress index and squared correlation index [211]. Stress is a loss function and it must be less than 0.2, whereas the squared correlation has a value greater than 0.8. Kruskal’s stress formula is used here. The stress value indicates the quality of the MDS measure. Hence, a higher value of stress gives a lower quality. The squared correlation index is the fraction of variance of the optimally scaled data that are considered for the respective distances by the MDS. The squared correlation index is a goodness-of-fit measure in statistics. Contrarily, stress indicates the not goodness-of-fit, i.e. the fraction of variance of the scaled data that is not considered for the respective distances by the MDS.

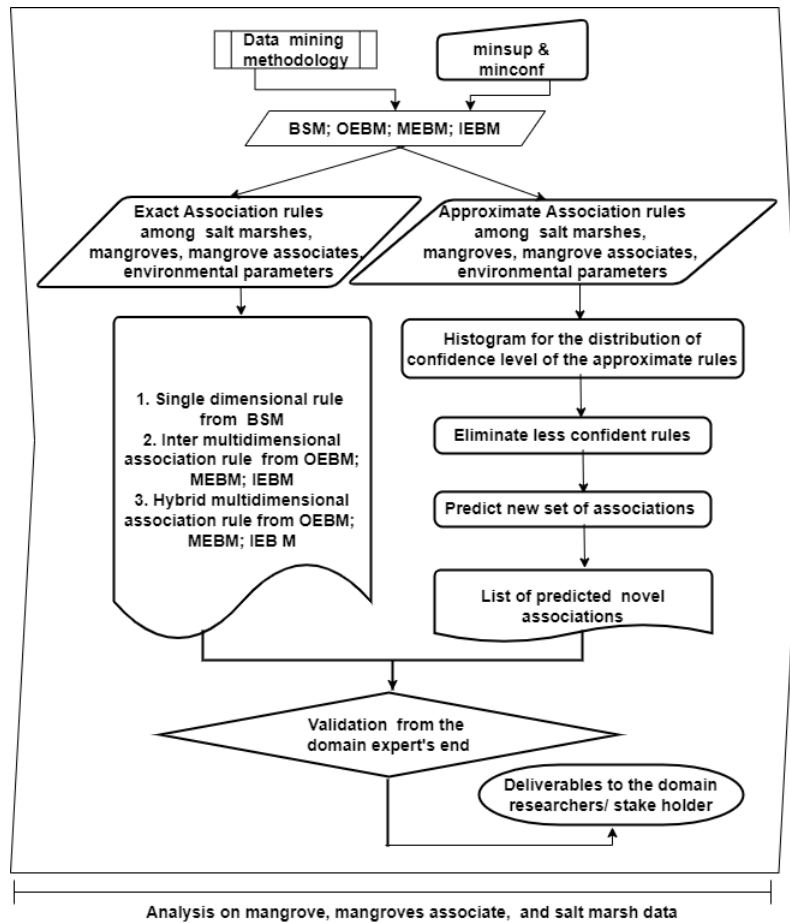


Figure 9.6: Stage B of the proposed framework

9.4 Result

9.4.1 Finding FCI on *BSM* dataset

For *BSM* dataset, the number of generated FCIs at varying minimum support are depicted graphically in Figure 9.7. It is obvious that as the value of minimal support is reduced, the number of FCIs increases. Basically, the minimum support prunes the candidates by mentioning the lower bound of the dataset rows to be considered in generating FCIs.

Statistics on the generated FCIs

For *BSM* dataset, the number of generated FCIs at varying minimum support is depicted graphically in Figure 9.7. It is obvious that as the value of minimal support is reduced, the number of FCIs increases. Basically, the minimum support prunes the candidates by mentioning the lower bound of the dataset rows to be considered in generating FCIs.

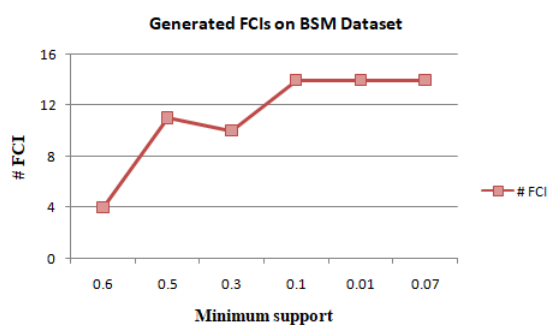


Figure 9.7: Generated frequent itemsets with varying minimum support count on *BSM* dataset

Illustration on the generated FCIs

FCIs for the *BSM* dataset (Table 9.4) are the cluster of salt marshes that have co-occurrence reports in a cluster of blocks. The minimum number of blocks in a cluster is the user-specified threshold to be considered frequent.

Figure 9.8 summarises the evidence recorded in Table 9.4.

The thirteenth row of Table 9.4 extracts the list of common zones where *Suaeda maritima* and *Suaeda nudiflora*, as well as three species of *Tamarix*, co-occur. All zones are from the inner estuary or the middle estuary. The co-occurrence data exclusively for *Suaeda maritima* and *Suaeda nudiflora* are derived from the twelfth row, which includes a larger number of blocks from all types of estuaries (inner, middle, and outer). Therefore, *Tamarix dioica*, *Tamarix gallica*, and *Tamarix troupia* can be said to thrive, particularly in regions of the inner and middle estuaries.

Our observation is supported by a study cited in [54]. According to [54], *Tamarix* species are characterized as species that flourish in areas where tidal flooding occurs at a regular interval. Therefore, the parts of the estuary with a greater elevation are favorable for this species (Region B of the venn-diagram in Figure 9.8). *Aeluropus lagopoides*, *Suaeda maritima*, *Suaeda nudiflora* are present in all types of estuaries, according to the sixth row of Table 9.4. The same is reported for *Aeluropus lagopoides*, *Heliotropium curassavicum*, and *Suaeda nudiflora* in the third row. Therefore, the most typically developing salt marshes in all sorts of estuaries are *Aeluropus lagopoides*, *Suaeda maritima*, *Suaeda nudiflora*, and *Heliotropium curassavicum* (Region A of the Venn-diagram in Figure 9.8).

The cluster of *Aeluropus lagopoides*, *Suaeda nudiflora*, and *Sesuvium portulacastrum* is found only in the outer region of tidal inundated areas (the tenth row of the same table), indicating the preferable zone for *Sesuvium portulacastrum* as *Aeluropus lagopoides* and *Suaeda nudiflora* have a wider range of occurrence. From the cluster of salt marshes present in the eleventh row (*Aeluropus lagopoides*, *Suaeda nudiflora*, *Trianthema portulacastrum*, *Trianthema triquetra*) data of *Trianthema portulacastrum*, and *Trianthema triquetra* can be obtained as *Aeluropus lagopoides*, *Suaeda nudiflora* are already recognized to have occurred in almost all blocks. So, it can be claimed that *Trianthema portulacastrum*, and *Trianthema triquetra* prefer to grow in the outer estuary (Region C in Figure 9.8). [54] describes its appearance in water-logged areas. So, all the salt marshes, under study, can be distinguished by their preferable zones of occurrences, as shown in Figure 9.8.

Table 9.4: Frequent closed itemsets generated from *BSM* dataset

S. No.	Cluster of salt marshes	Cluster of blocks
1	<i>Aeluropus lagopoides</i> , <i>Heliotropium curassavicum</i> , <i>Suaeda maritima</i> , <i>Suaeda nudiflora</i> ,	Arbesi, Chottohardi, Herobhanga, Jhilla, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
2	<i>Aeluropus lagopoides</i> , <i>Heliotropium curassavicum</i> , <i>Suaeda maritima</i> , <i>Suaeda nudiflora</i> , <i>Tamarix dioica</i> , <i>Tamarix gallica</i> , <i>Tamarix troupii</i> ,	Arbesi, Jhilla, Netidhopani, Panchmukhani, Phirkhali
3	<i>Aeluropus lagopoides</i> , <i>Heliotropium curassavicum</i> , <i>Suaeda nudiflora</i>	Arbesi, Ajmalmari, Chottohardi, Herobhanga, Jhilla, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
4	<i>Aeluropus lagopoides</i> , <i>Salicornia brachiata</i> , <i>Suaeda maritima</i> , <i>Suaeda nudiflora</i>	Bagmara, Gona
5	<i>Aeluropus lagopoides</i> , <i>Suaeda maritima</i>	Arbesi, Bagmara, Chamta, Chottohardi, Chulkati, Gona, Harinbhaga, Herobhanga, Jhilla, Khatuajhuri, Matla, Mayadwip, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
6	<i>Aeluropus lagopoides</i> , <i>Suaeda maritima</i> , <i>Suaeda nudiflora</i>	Arbesi, Bagmara, Chottohardi, Chulkati, Gona, Harinbhaga, Herobhanga, Jhilla, Khatuajhuri, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
7	<i>Aeluropus lagopoides</i> , <i>Suaeda maritima</i> , <i>Suaeda nudiflora</i> , <i>Tamarix dioica</i> , <i>Tamarix gallica</i> , <i>Tamarix troupii</i>	Arbesi, Harinbhaga, Jhilla, Khatuajhuri, Netidhopani, Panchmukhani Phirkhali
8	<i>Aeluropus lagopoides</i> , <i>Suaeda maritima</i> , <i>Tamarix dioica</i> , <i>Tamarix gallica</i> , <i>Tamarix troupii</i>	Arbesi, Harinbhaga, Jhilla, Khatuajhuri, Matla, Netidhopani, Panchmukhani Phirkhali
9	<i>Aeluropus lagopoides</i> , <i>Suaeda nudiflora</i>	Arbesi, Ajmalmari, Bagmara, Chottohardi, Chulkati, Dhulibhasani, Gona, Harinbhaga, Herobhanga, Jhilla, Khatuajhuri, Murganga, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
10	<i>Aeluropus lagopoides</i> , <i>Suaeda nudiflora</i> , <i>Sesuvium portulacastrum</i>	Gona, Murganga
11	<i>Aeluropus lagopoides</i> , <i>Suaeda nudiflora</i> , <i>Trianthema portulacastrum</i> , <i>Trianthema triquetra</i>	Saptamukhi, Murganga
12	<i>Suaeda maritima</i> , <i>Suaeda nudiflora</i>	Arbesi, Bagmara, Chandkhali, Chottohardi, Chulkati, Gona, Harinbhaga, Herobhanga, Jhilla, Khatuajhuri, Netidhopani, Panchmukhani, Phirkhali, Saptamukhi, Thakuran
13	<i>Suaeda maritima</i> , <i>Suaeda nudiflora</i> , <i>Tamarix dioica</i> , <i>Tamarix gallica</i> , <i>Tamarix troupii</i>	Arbesi, Chandkhali, Harinbhaga, Jhilla, Khatuajhuri, Netidhopani, Panchmukhani, Phirkhali
14	<i>Suaeda maritima</i> , <i>Tamarix dioica</i> , <i>Tamarix gallica</i> , <i>Tamarix troupii</i>	Arbesi, Chandkhali, Goashaba, Harinbhaga, Jhilla, Khatuajhuri, Matla, Netidhopani, Panchmukhani, Phirkhali

Validation of the generated FCIs through MDS

Here, the aim is to justify the outcome that has been derived from the frequent closed itemsets (Figure 9.8). The input to the dataset for the MDS is a square, symmetric correlation matrix showing the relationships among the set of items. The cells represent the support values (the frequency of occurring together) for the occurrences of the corresponding matrix elements. The correlation matrix for the salt marshes is given in Table 9.5. For MDS plotting, the cell values are subtracted from the maximum value. This implies that, if two salt marshes are experiencing a stronger correlation, the proximity between these two will be lower. This is resulting in a closer portrayal on the map. The used SPSS MDS model calculates the stress and the squared correlation as 0.163 and 0.86, respectively, assuring that the model is acceptable and significant. Once the correlation matrix is found, the spatial representation of these relationships could be established via MDS. In the multidimensional space, all the salt marshes are shown in Figure 9.9 as per their spatial relationship.

When comparing with Figure 9.8, which is generated from the result of FCIs of *BSM*, it can be stated that both the figures (Figure 9.8 and 9.9) exhibit quite similar proximity among themselves. All three species of *Tamarix* are exhibiting closer proximity and they are located far apart from *Trianthema* and *Sesuvium* (Figure 9.9). It was found that *Tamarix* was

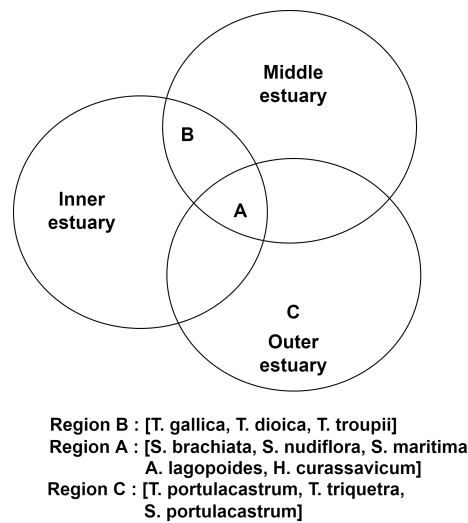


Figure 9.8: Distribution of Salt Marshes along the estuaries as derived from the result of frequent closed itemsets given in Table 9.4

reported from the inner and middle estuarine region, whereas *Trianthema* and *Sesuvium* were from the outer estuarine region (Figure 9.8). Hence, it justifies the observation made in Figure 9.8. As *S. nudiflora*, *S. maritima*, *A. lagopoides* are tolerating a wide range of salinity variations, they are occurring in all the three estuaries (Figure 9.8). Hence, these are comparatively closer to all other salt marshes, as appeared in Figure 9.9.

9.4.2 Association rule generation on OEEM, MEBM, IEBM

It is found that the result set of frequent closed itemsets obtained from *BSM*, can be validated through SPSS (as discussed in section 9.4.1). Therefore, we proceed further with homogeneous kinds of datasets for knowledge discovery via the generation of the rules for the salt marsh dataset alone and the combined dataset of mangroves, mangroves associates, and salt marshes.

Detailed statistics on the generated exact rule

The number of exact rules obtained (for *OEEM*, *MEBM*, and *IEBM*) has been shown via the primary vertical axis on the left side of the graph (Figure 9.10). The secondary vertical axis is showing the respective values for the minimum support and the minimum confidence for generating the rules. We consider the minimum support values of 0.1, 0.3, 0.5, 0.7, 0.9, and 1 depending on the dataset size. For each minimum support value, the rules are generated for the minimum confidence values of 0.1, 0.3, 0.6, and 0.9 (shown via the bars along the horizontal axis).

It is evident from the graph that *IEBM* generates the maximum number of rules for all combinations of minimum support and confidence values represented by *#IE_exact rules* in the graph. The reason can be understood as this dataset covers the maximum number of species data. When considering exact rules, the confidence value of a rule is 1, as stated before. Hence, for all the minimum support values, we obtain the same number of exact rules while varying minimum confidence.

Table 9.5: Correlation matrix for the salt marshes obtained from the support values

Salt Marsh	<i>Sesuvium portulacastrum</i>	<i>Aeluropus lagopoides</i>	<i>Suaeda maritima</i>	<i>Salicornia brachiata</i>	<i>Suaeda nudiflora</i>	<i>Tamarix dioica</i>	<i>Tamarix gallica</i>	<i>Tamarix troupii</i>	<i>Trianthema portulacastrum</i>	<i>Trianthema triquetra</i>	<i>Heliotropium curassavicum</i>
<i>Sesuvium portulacastrum</i>	0	2	1	1	2	1	1	0	1	1	0
<i>Aeluropus lagopoides</i>	2	0	17	2	17	8	8	8	2	2	10
<i>Suaeda maritima</i>	1	17	0	2	15	10	10	10	1	1	9
<i>Salicornia brachiata</i>	1	2	2	0	2	1	1	0	0	0	0
<i>Suaeda nudiflora</i>	2	17	15	2	0	8	8	8	2	2	10
<i>Tamarix dioica</i>	1	8	10	1	8	0	10	10	0	0	5
<i>Tamarix gallica</i>	1	8	10	1	8	10	0	10	0	0	5
<i>Tamarix troupii</i>	0	8	10	0	8	10	10	0	0	0	5
<i>Trianthema portulacastrum</i>	1	2	1	0	2	0	0	0	0	2	1
<i>Trianthema triquetra</i>	1	2	1	0	2	0	0	0	2	0	1
<i>Heliotropium curassavicum</i>	0	10	9	0	10	5	5	5	1	1	0

Detailed statistics on the generated approximate rules

The number of proper and structural bases for the approximate association rules have been shown in Figure 9.11. The same values of minimum support and minimum confidence are used, as before. The number of approximate rules obtained has been shown via the primary vertical axis on the left side of the graph. The secondary vertical axis is showing the respective values for the minimum support and the minimum confidence for generating the rules. As expected, the size of the structural bases is always much larger than the proper bases. As per our test parameters, for OEBM, the maximum number of approximate rules are generated for minsup and minconf values of (0.1, 0.1), respectively. For MEBM, the maximum number of rules are obtained in two cases, for minsup and minconf values of (0.3, 0.1) and (0.1, 0.1). For IEBM, the maximum number of rules are obtained for test parameters (0.5,0.1), (0.5,0.3), (0.3,0.1), (0.3,0.3), (0.1,0.1), (0.1,0.3), (0.1,0.6), (0.1,0.9).

Illustration for the generated rules on BSM, OEBM, MEBM, IEBM

This section highlights only a few rules from the generated ruleset. The highlighted rules can be considered to fall under different categories, viz. single-dimensional, inter-multidimensional, and hybrid-multidimensional.

Single dimensional association rule (uses only a single predicate in rule): Can be used to find the presence association among the salt marshes A rule on BSM as shown in Rule 9.4.1 uses a single predicate “presence”. It states that the presence of *Tamarix dioica* infers the presence of *Tamarix gallica*, *Tamarix troupii*, and *Suaeda maritima* in 10 blocks (Phirkhali, Panchmukhani, Jhilla, Chandkhali, Goashaba, Arbesi, Khatuajhuri, Har-

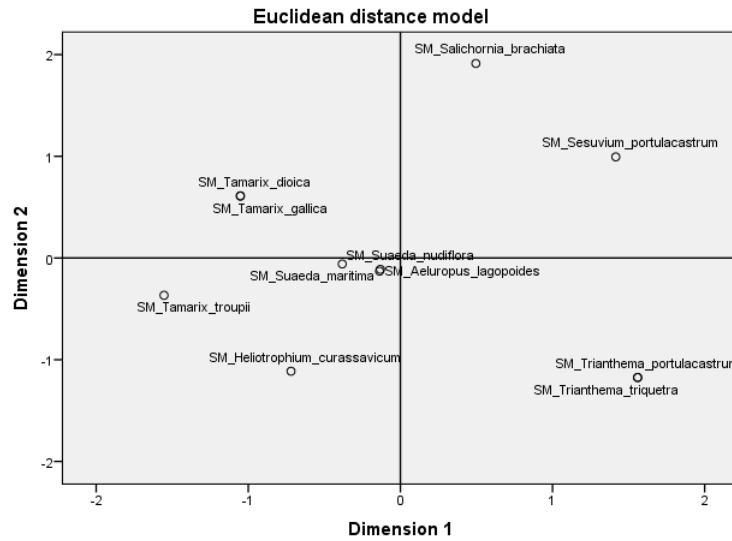


Figure 9.9: Derived stimulus configuration- MDS on salt marsh association matrix

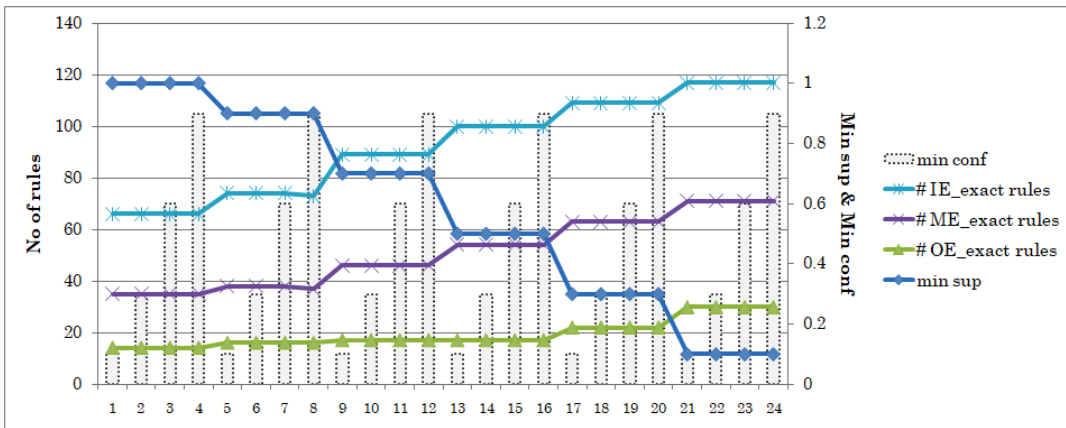


Figure 9.10: Bar plot showing the number of exact rules generated with varying minimum support and confidence in all three datasets for OE, ME, and IE

inhbaga, Matla, Netidhopani).

Inter-multidimensional association rule (uses multiple predicates, but no repetition in predicates in antecedent and in consequent): Can show the salt marsh association with mangrove and mangrove associates. Considering a rule on *IEBM* dataset (Rule 9.4.2), the used predicates are “salt marsh”, “mangrove”, and “mangrove associates”. Here, the predicate “salt marsh” is not repeated in the consequent part of the rule. Therefore, it can be termed as the inter-multidimensional association rule. It is stating the co-occurrence status for the salt marsh *Suaeda maritima* with mangroves and mangrove associates. The support value for the rule is 8, and the supporting blocks are Phirkhali, Jhilla, Goashaba, Arbesi, Khatuajhuri, Harinhbaga, and Herobhanga. Also, its requirement for pH, salinity, soil texture, and tidal height, are presented in the rule as multiple predicates.

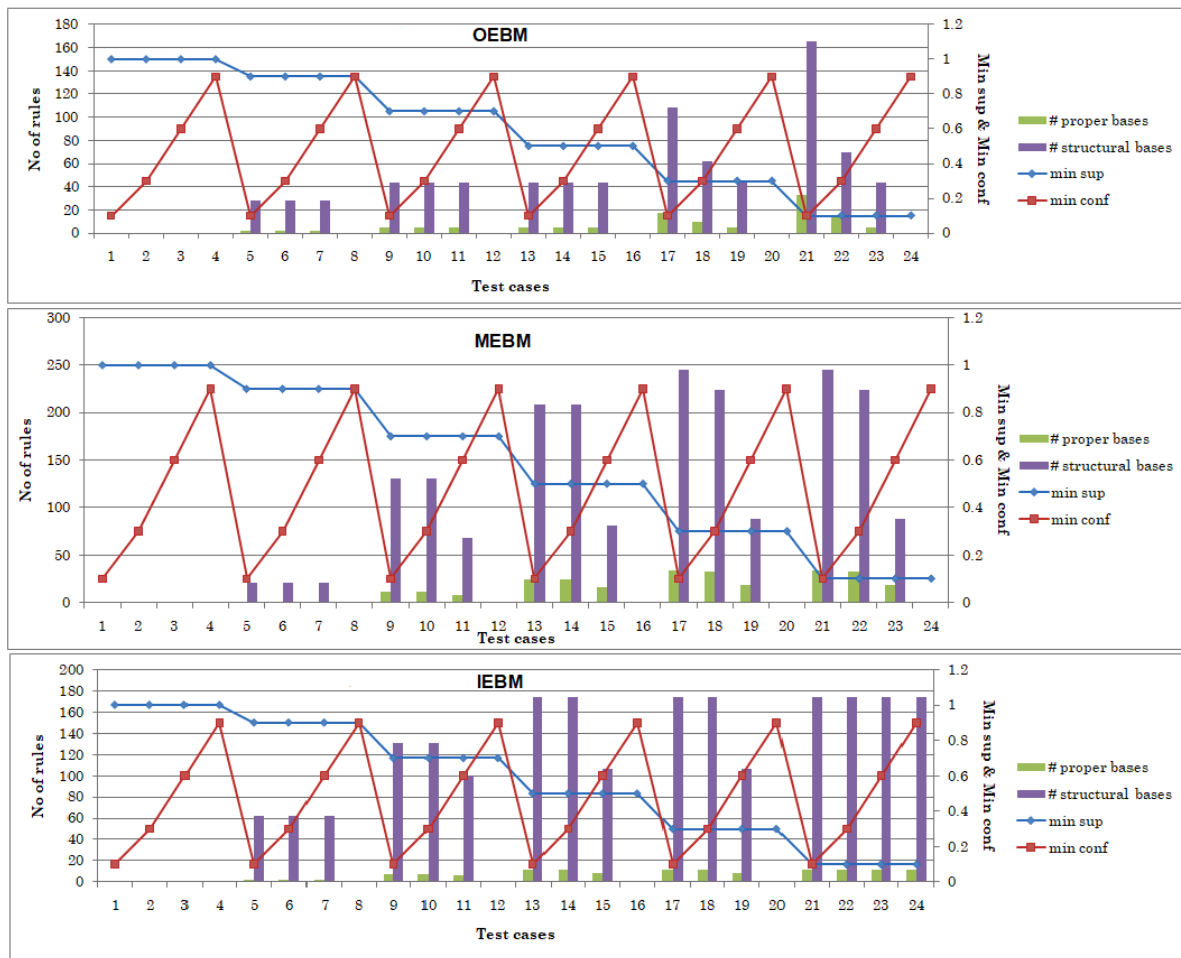


Figure 9.11: Bar plot showing the number of approximate rules (proper base and structural base) generated with varying minimum support and confidence in all three datasets for OE, ME, and IE

Hybrid multidimensional association rule: (uses multiple predicates with repetition in predicates in antecedent and in consequent): Can find mangrove association with other mangroves, mangrove associates, and salt marshes Considering a rule on *OEBM* dataset (Rule 9.4.3), the used predicates are “Mangrove ”and “Salt marsh ”. There is the repetition of the predicate “Mangrove ”in the antecedent and in the consequent as well. Hence, it is called the hybrid multidimensional association rule. It states the co-occurrence of *Sonneratia griffithii* with other mangroves, and salt marshes, where the support is 8, i.e., 8 blocks under the outer estuarine region holds this rule.

Rules on IUCN red-listed mangroves International Union for the Conservation of Nature and Natural Resources (IUCN) has categorized the red list of species into *Critically Endangered*, *Endangered*, *Vulnerable*, *Near Threatened*, *Least Concern*, *Data Deficient*. Among the Indian mangroves, *Phoenix paludosa*, *Brownlowia tersa*, *Aegialitis rotundifolia*, *Ceriops decandra* fall under *Near Threatened* category. *Sonneratia griffithii* is on the list of *Critically Endangered*, whereas *Heritiera fomes* is in the *Endangered* category [160]. This study enlists the rules explaining the co-existence status of the above-mentioned red-listed species. The

Rule 9.4.1: Rule for presence association among the salt marshes (example of single dimensional association rule)

[**Presence** *Tamarix dioica* = 1] \implies

[**Presence** {*Tamarix gallica* = 1, *Tamarix troupii* = 1, *Suaeda maritima* = 1}]

with support = 10, confidence = 1

Rule 9.4.2: Rule for salt marsh association with mangrove and mangrove associates (example of inter multidimensional association rule for *Suaeda maritima*)

[**Salt Marsh** *Suaeda maritima* = Present] \implies

[**Salinity**= Brackish water, **pH**= Moderately alkaline, **Soil texture**= Clay loam, **Tidal height**= Moderate,

Mangrove {*Avicennia officinalis*= Abundant, *Aglaia cuculata*= Occasional, *Bruguiera gymnorhiza*= Abundant, *Bruguiera sexangula*= Frequent, *Excoecaria agallocha*= Abundant, *Heritiera fomes*= Abundant, *Phoenix paludosa*= Abundant, *Nypa fruticans*= Frequent, *Sonneratia apetala*= Abundant, *Sonneratia caseolaris*= Frequent, *Xylocarpus mekongensis*= Frequent }

Mangrove Associate {*Acanthus ilicifolius*= Frequent, *Brownlowia tersa*= Frequent, *Cerbera odollam*= Occasional, *Clerodendrum inerme*= Abundant, *Crinum defixum*= Abundant, *Cyperus exaltatus*= Rare, *Cynometra ramiflora*= Frequent, *Derris trifoliata*= Abundant, *Fimbristylis ferruginea*= Rare, *Intsia bijuga*= Occasional, *Myriostachya wightiana*= Frequent, *Pentatropis capensis*= Abundant, *Porteresia coarctata*= Abundant, *Scirpus littoralis*= Rare }

with support = 8, confidence = 1

extracted facts on *Heritiera fomes* and *Brownlowia tersa* can be found in Rule 9.4.4 and Rule 9.4.5, respectively. Rule 9.4.6, 9.4.7, 9.4.8 are for *Phoenix paludosa*, *Ceriops decandra*, and *Aegialitis rotundifolia*, respectively. Co-existence status of *Sonneratia griffithii* has already been highlighted in Rule 9.4.3.

9.5 Discussion

It is clearly perceived that a study on the co-existence pattern of mangroves and salt marshes can improve our understanding regarding the ecosystem evaluation for the restoration of coastal wetlands. In this work, the proposed data mining-based algorithmic framework has shown an efficient way of identifying multi-species frequent co-occurrence

Rule 9.4.3: Rule for mangrove association with salt marsh and other mangroves (example of hybrid multidimensional association rule for *Sonneratia griffithii*)

[**Mangrove** *Sonneratia griffithii*= Abundant] \implies

[**Salt Marsh** *Aeluropus lagopoides*= Present,

Salinity in TDS (g/L)= Sea water, **pH**= Slightly alkaline, **Soil texture**= Silt Loam, **Tidal height**= Maximum,

Mangrove {*Bruguiera cylindrica*= Abundant, *Bruguiera parviflora*= Abundant, *Ceriops tagal*= Abundant, *Avicennia alba*= Abundant, *Avicennia marina*= Abundant, *Phoenix paludosa*= Abundant, *Aegialitis rotundifolia*= Abundant, *Excoecaria agallocha*= Abundant }

Rule 9.4.4: Co-existence pattern of *Heritiera fomes* (Endangered species) with salt marshes and other rare/ frequent/ abundant mangroves/ mangrove associates and influencing environmental factors:

[Mangrove *Heritiera fomes* = Abundant] \implies

[Salt Marsh *Suaeda maritima* = Present,

Salinity = Brackish water, **pH** = Moderately alkaline, **Soil texture** = Clay loam, **Tidal height** = Moderate,

Mangrove {*Avicennia officinalis* = Abundant, *Aglaiia cuculata* = Occasional, *Bruguiera gymnorrhiza* = Abundant, *Bruguiera sexangula* = Frequent, *Cynometra ramiflora* = Frequent, *Excoecaria agallocha* = Abundant, *Nypa fruticans* = Frequent, *Phoenix paludosa* = Abundant, *Sonneratia apetala* = Abundant, *Sonneratia caseolaris* = Frequent, *Xylocarpus mekongensis* = Frequent}

Mangrove Associate {*Acanthus ilicifolius* = Frequent, *Brownlowia tersa* = Frequent, *Cerbera odollam* = Occasional, *Clerodendrum inerme* = Abundant, *Crinum defixum* = Abundant, *Cyperus exaltatus* = Rare, *Derris trifoliata* = Abundant, *Fimbristylis ferruginea* = Rare, *Intsia bijuga* = Occasional, *Myriostachya wightiana* = Frequent, *Pentatropis capensis* = Abundant, *Porteresia coarctata* = Abundant, *Sarcolobus globosus* = Abundant, *Scirpus littoralis* = Rare }

support = 7, confidence = 1

data. The deliverable information is in the form of association rules (as shown in Rule 9.4.1 - 9.4.8) derived from the existing facts. The illustration for these rules and the similarities between these postulations with previous empirical findings have been discussed in section 9.5.1. Another important aspect can be drawn from the extracted frequent co-occurrence data of multi-species. That is, a new probable association can be inferred from the existing data. This has been shown in section 9.5.2. The significance of this study in restoration ecology is illustrated in section 9.5.3.

9.5.1 Finding inferences from exact association rules

As stated before, exact association rules uncover the underlying facts of the dataset. Therefore, the significance of this kind of rule can be well understood in the study of individual mangroves as its co-existence status can be visualized. Considering a species in the antecedent (the left part of the rule), the rules highlight the co-existence status of that species with the detailed occurrence data of others in the consequent part (the right part of the rule) and the conducive conditions as well.

In Rule 9.4.3, the example of *Sonneratia griffithii*, an IUCN red-listed critically endangered species [160], has been shown. It is found from the rule that it has a frequent co-existence with salt marsh *Aeluropus lagopoides*, and other mangroves in abundant quantity, (*Bruguiera cylindrica*, *Bruguiera parviflora*, *Ceriops tagal*, *Avicennia alba*, *Avicennia marina*, *Phoenix paludosa*, *Aegialitis rotundifolia* *Excoecaria agallocha*). A subset of these mangroves has appeared in a field survey report [212] of *Sonneratia griffithii* in Indian Sundarban. The report demonstrated and specified a similar co-occurred species set of *Sonneratia griffithii*. Along with the co-existing species list, preferable environmental conditions (high seawater inundation, slightly alkaline water pH, and silt-loam kind of soil texture) for the frequent occurrence of *Sonneratia griffithii* are also highlighted in the derived rule. It is believed that this kind of study would be helpful for field researchers, especially in mangrove restoration through afforestation.

Rule 9.4.5: Co-existence status of *Brownlowia tersa*, (Near Threatened) with salt marshes and other rare/ frequent/ abundant mangroves/ associates and influencing environmental factors:

[Mangrove *Brownlowia tersa* = Frequent] →

[SaltMarsh *Suaeda maritima* = Present,

Salinity = Brackish water, pH = Moderately alkaline, Soil texture = Clay loam, Tidal height = Moderate,

Mangrove {*Avicennia officinalis* = Abundant, *Aglaiia cuculata* = Occasional, *Bruguiera gymnorrhiza* = Abundant, *Bruguiera sexangula* = Frequent, *Clerodendrum inerme* = Abundant, *Heritiera fomes* = Abundant, *Intsia bijuga* = Occasional, *Nypa fruticans* = Frequent, *Phoenix paludosa* = Abundant, *Sonneratia apetala* = Abundant, *Sonneratia caseolaris* = Frequent, *Xylocarpus mekongensis* = Frequent},

Mangrove Associate {*Acanthus ilicifolius* = Frequent, *Cerbera odollam* = Occasional, *Crinum defixum* = Abundant, *Cynometra ramiflora* = Frequent, *Cyperus exaltatus* = Rare, *Derris trifoliata* = Abundant, *Excoecaria agallocha* = Abundant, *Fimbristylis ferruginea* = Rare, *Myriostachya wightiana* = Frequent, *Pentatropis capensis* = Abundant, *Porteresia coarctata* = Abundant, *Sarcobolus globosus* = Abundant, *Scirpus littoralis* = Rare }

support = 7, confidence = 1

Considering Rule 9.4.4, it could be used to identify the favorable regions of *Heritiera fomes* where the environmental conditions along with the co-existing plant species are given. The said rule has a support value of 8, i.e., similar data have been reported from eight numbers of blocks. It is found from the rule that *Heritiera fomes* has a preference for brackish water, i.e. less saline areas, which is agreed with the findings made in a similar kind of study cited in [213]. As per [213], in Bangladesh Sundarban, salinity had a negative impact on *Heritiera fomes*. The species was rare in high-salinity areas but common in low-salinity areas. In both moderate and low salinity conditions, the presence was abundant.

A study has shown that *Brownlowia tersa* has eventually been removed from the Indian Sundarban mangrove forest due to salinity intolerance [214]. Its preferable growing environment along with co-occurred species, as identified by our study, has been featured in Rule 9.4.5. Previous studies [12, 215] on *Phoenix paludosa* reported its intolerance to higher salinity. The identical scenario for *Phoenix paludosa* is reinforced by our findings (Rule 9.4.6). Our analysis of nearly threatened *Ceriops decandra* specifies that (Rule 9.4.7) salt-water preference with high tidal amplitude, clay type soil texture causes abundant growth of *Ceriops decandra*. *Excoecaria agallocha* also reported having abundant occurrence with *Ceriops decandra*. These findings are consistent with the previous research reports in [216]. Study results in [217] depicted the phenology of *Aegialitis rotundifolia* mentioning that *Ceriops decandra*, *Ceriops tagal*, *Bruguiera gymnorrhiza*, and *Excoecaria agallocha* all are growing alongside it. *Aegialitis annulata* and *Aegialitis rotundifolia* are the two species in the genus *Aegialitis* and they never occur together [218, 217]. In our study, *Aegialitis annulata*, does not have any presence report. The *Aegialitis* species prefers exposed areas and can withstand waves and tidal action, according to [218]. In addition, it is also stated that it can grow in severely saline soil. The habitat needs of *A. rotundifolia* has also been documented by [219], which stated that this species can also be found on the coastline, and grows in a saline environment. All these comply with our findings for *A. rotundifolia* (Rule 9.4.8).

Rule 9.4.6: Co-existence status of *Phoenix paludosa* (Near Threatened) with salt marshes and other rare/ frequent/ abundant mangroves/ associates and influencing environmental factors:

[Mangrove *Phoenix paludosa* = Abundant] →

[Salt marsh *Suaeda maritima* = Present,

Salinity = Brackish water, pH = Moderately alkaline, Soil texture = Clay loam, Tidal height = Moderate,

Mangrove {*Avicennia officinalis* = Abundant, *Scirpus littoralis* = Rare, *Sonneratia caseolaris* = Frequent, *Xylocarpus mekongensis* = Frequent}, *Nypa fruticans* = Frequent, *Aglaia cuculata* = Occasional, *Bruguiera gymnorrhiza* = Abundant, *Bruguiera sexangula* = Frequent, *Heriatiera fomes* = Abundant, *Sonneratia apetala* = Abundant},

Mangrove Associate {*Acanthus ilicifolius* = Frequent, *Brownlowia tersa* = Frequent, *Cerbera odollam* = Occasional, *Clerodendrum inerme* = Abundant, *Crinum defixum* = Abundant, *Cyperus exaltatus* = Rare, *Cynometra ramiflora* = Frequent, *Derris trifoliata* = Abundant, *Fimbristylis ferruginea* = Rare, *Intsia bijuga* = Occasional, *Myriostachya wightiana* = Frequent, *Pentatropis capensis* = Abundant, *Porteresia coarctata* = Abundant, *Sarcolobus globosus* = Abundant}],

support = 8, confidence = 1

Rule 9.4.7: Co-existence pattern of *Ceriops decandra* (Near Threatened) with salt marshes and other rare/ frequent/ abundant mangroves/ associates and influencing environmental factors:

[Mangrove *Ceriops decandra* = Abundant]

Salinity in TDS (g/L) = Salt water, pH = Moderately alkaline, Soil texture = Clay, Tidal height = Higher,

Mangrove {*Raizophora apiculata* = Abundant, *Raizophora mucronata* = Abundant, *Kandelia candel* = Abundant, *Aegiceras corniculatum* = Abundant, *Xylocarpus mekongensis* = Abundant, *Xylocarpus granatum* = Abundant, *Bruguiera gymnorrhiza* = Abundant, *Excoecaria agallocha* = Abundant, *Avicennia officinalis* = Abundant, *Avicennia alba* = Abundant, *Phoenix paludosa* = Abundant},

Mangrove Associate {*Finlaysonia obovata* = Frequent, *Derris scandens* = Occasional, *Tylophora tenuis* = Rare, *Sarcolobus globosus* = Abundant, *Acanthus volubilis* = Occasional}],

support = 7, confidence = 1

9.5.2 Predicting novel associations from approximate association rules

Consider a rule in the form of [Antecedent] – > [Consequent] with supporting object lists. Taking a set of rules, {n}, where the antecedent part is same for all the cases, and considering the combination of any two rules, say R1 and R2, from the set {n},

Object list (R1) - Object list (R2) = {r}, i.e., the set of r objects. As the antecedent part is the same for both cases, the elements in {r} should be held for the whole consequent part of R2. Hence, new associations could be alike, {Consequent (R2) - Consequent (R1)} – > {Object list (R1) - Object list (R2)} where, Antecedent(R1) = Antecedent(R2) and confidence of R1 and R2 > minimum threshold.

So, for the obtained set of approximate rules, we follow this procedure for the discovery

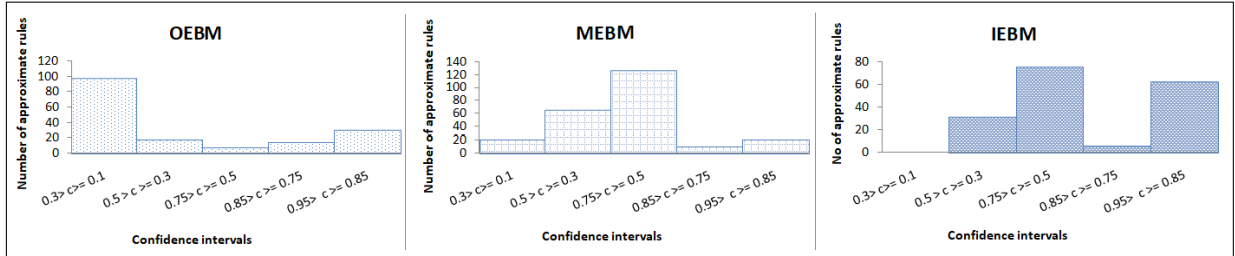


Figure 9.12: Histogram for the number of approximate rules (structural base) generated at varying confidence level

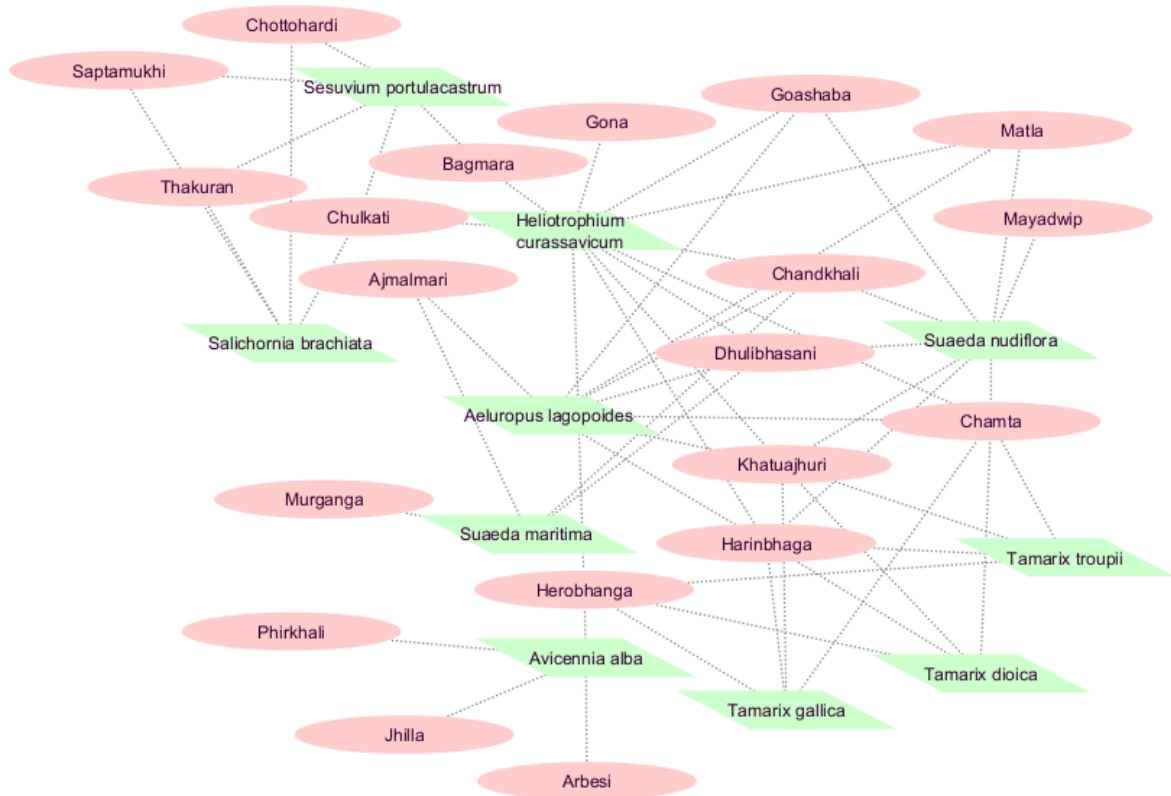


Figure 9.13: Network for the predicted occurrences for the species at multiple blocks

Rule 9.4.8: Co-existence pattern of *Aegialitis rotundifolia* (Near threatened) with salt marshes and other rare/ frequent/ abundant mangroves/ associates and influencing environmental factors:

[Mangrove *Aegialitis rotundifolia* = Abundant]

[Salt marsh *Aeluropus lagopoides* = Present,

Salinity in TDS (g/L) = Sea water, pH = Slightly alkaline, Soil texture = Silt Loam, Tidal height = Maximum,

Mangrove {*Bruguiera cylindrica*= Abundant, *Bruguiera parviflora* = Abundant, *Ceriops tagal* = Abundant, *Avicennia alba* = Abundant, *Avicennia marina* = Abundant, *Sonneratia griffithii* = Abundant, *Phoenix paludosa* = Abundant, *Excoecaria agallocha* = Abundant}],

support = 8, confidence = 1

of novel associations. Figure 9.12 shows the histogram for the confidence interval versus the number of generated approximate rules for the datasets *OEBM*, *MEBM*, and *IEBM*. Rules with very low confidence values are ignored for predicting new associations. Here, we consider 0.85 as the threshold. The new association prediction is visualized using Cytoscape software (Figure 9.13), with parallelogram and oval shapes representing the species and regions, respectively. A dotted line connecting a parallelogram with an oval shows the prediction of the presence of a species in a particular region. It is found that all the predictions are mainly for the salt marshes. As the salt marshes have scattered occurrence records from all three estuarine regions, the probability of finding new occurrences is higher for salt marshes. Contrarily, the records of the mangroves and mangrove associates are estuary specific. Hence, the probability of finding new occurrences is limited for them. In the case of our dataset, all the predictions on mangroves and their associates fall below the threshold except for *A. alba* (Figure 9.13). Salt marsh *A. lagopoides*, *H. curssavicum* are found to have major probable new associations in multiple blocks, followed by *S. nudiflora* and *S. maritima*. Our predicted associations (Figure 9.13) agree with the zonation pattern of salt marshes (Figure 9.8) that we have discussed in section 9.4.1. For example, the predicted blocks of occurrence for *Sesuvium portulacastrum* are from the outer estuary only (Chot-tohardi, Saptamukhi, Thakuran, Bagmara, Chulkati), as this species prefers tidal inundated areas, as stated before. Similarly, for *Tamarix*, the predicted blocks are only from the inner and middle estuarine regions.

9.5.3 Implications on restoration practitioners

Finally, the question will be which foundation species among salt marshes and mangroves should be planted in the new area under study to enhance community structuring in the salt marsh-mangrove ecotone?

Our findings clearly demonstrate the blockwise frequent itemsets of mangroves, their associates, and salt marshes. Extreme salt stress inhibits the growth and expansion of many mangroves, and high salinity restricts the structural development of mangrove forests. The responsiveness of mangrove species in the Sundarban has been found to differ considerably throughout the salinity gradient. Hence, depending on various environmental parameters, that differ along with the blocks, a frequent set of species can be identified. That frequent set

could be selected by restoration practitioners. A list of such frequent associations has been obtained here by the proposed methodology. The detail regarding any species under study can be extracted from the list. For example, if the planting of the endangered species *Heritiera fomes* is desired, our findings (Rule 9.4.4) demonstrate that an abundant quantity of this species are found in brackish water, where water pH is moderately alkaline, tidal amplitude is moderate, and soil texture is of clay loam type. In a scientific report [216] dealing with the habitat suitability model, *Heritiera fomes* shows a clear unfavorable impression of salinity. It is reported to be found in dense populations in less saline and freshwater-rich habitats, i.e, in brackish water. Additionally, information regarding native co-existing plants can be obtained through our study. *Suaeda maritima* is frequently found co-occurred salt marsh for *Heritiera fomes*. Other mangroves, such as *Avicennia officinalis*, *Bruguiera gymnorhiza*, *Bruguiera sexangula*, *Cynometra ramiflora*, *Excoecaria agallocha*, *Nypa fruticans*, *Phoenix paludosa*, *Sonneratia apetala*, *Sonneratia caseolaris*, *Xylocarpus mekongensis* and mangroves associates, such as, *Acanthus ilicifolius*, *Brownlowia tersa*, *Clerodendrum inerme*, *Crinum defixum*, *Derris trifoliata*, *Myriostachya wightiana*, *Pentatropis capensis*, *Porteresia coarctata*, *Sarcolobus globosus* can be planted alongside *Heritiera fomes* as abundance or frequent co-occurrence of these was detected (Rule 9.4.4).

The major problem of increasing soil salinity could be addressed by more plantations of compatible salt marshes as identified in the rule (Rule 9.4.4). Also, estuary-wise salt marsh growth patterns have been recognized from frequent associations among the salt marshes (Figure 9.8). Like, for *S. maritima*, co-occurred salt marshes are *S. nudiflora*, *S. brachiata*, *A. lagopoides*, and *H. curassavicum*. Hence, regeneration of *Heritiera fomes* could be achieved by introducing selected salt marshes to minimize hypersalinity. Plantations of other co-occurred mangroves can enrich biological diversity. This approach can be followed in any other dataset for particular species regeneration.

Finally, yet importantly, salt marshes were revealed to be superior species in [220] for fast-developing ecological structure at respective elevations due to its rapid expansion and recruitment rates in the coastal wetland in southeastern Louisiana (USA). [161] also agrees with the facilitative effect of salt marsh which is critical in assisting mangrove propagule colonization.

9.5.4 Comparisons with parallel studies in mangrove restoration in terms of used methodology, findings, and limitations

This section shows a comparative study on the failure and success of existing parallel mangrove restoration efforts from the insights of different published articles (Table 9.6, 9.7, 9.8, 9.9). The unique findings that could be gathered through our research have also been listed. In a nutshell, many initiatives have placed a focus on natural methods of restoration (Table 9.6). Some efforts have encouraged study-based plantations to follow systemic and effective restoration strategies (Table 9.7). Both aspects have been addressed in the present study. Interdisciplinary approaches of knowledge-based classifiers or fact-finding frameworks are not directly related to mangrove restoration operations (Table 9.8, 9.9). But, the knowledge-based decision on restoration activities helps to encompass a broader area of restoration while selecting proper sites and species as well, dominated by particular environmental conditions.

Unique characteristics of mangrove habitats are confronting multiple challenges such as

Table 9.6: Projects for the rehabilitation or restoration of mangrove: related amortization techniques throughout biogeographic regions

Project site/- country	Cause of impairment	Amelioration procedure	Limitation/ Remarks	Year	Ref no.
Prakasam, Guntur, Krishna, West Godavari, East Godavari, Visakhapatnam, Pulicat Lake, Andhra Pradesh, Kanchipuram, Tamil Nadu India	Exposed shores after tsunami or cyclones	- Forestation via considering three aspects, viz, ecological, social, and economic (financial support for plantations) aspects. - Plantation phase has been sub-divided into pre-plantation, plantation, and post-plantation phases.	- Need rigorous involvement of local people, even in decision making - Suitable sites and species selection, considering proper environmental parameters, prior to the plantation are necessary to save time, money, and effort.	2015	[221]
Bangladesh Delta	Loss of Mangrove Cover	- The assessment of mangrove regeneration via afforestation, - analyzing their species richness and community structure during a 40-year chronosequence in comparison to natural mangroves of that region	- Within 42 years, the species richness of natural mangroves was not attained through artificial regeneration. - The diversity of plantations may be increased by planting a variety of species.	2022	[189]

Table 9.7: Eco-engineering framework: the creation of resilient, natural and man-made ecosystems that integrate human society with its surroundings for the good of both.

Proposed framework	Findings	Limitation/ Remarks	Location	Year	Ref no
From grey to green: Assessing the methodology for nature-based vs. artificial coastal protection framework	- Evaluate the effectiveness of natural versus artificial coastal protection - Engineering solutions to safeguard coasts such as seawalls and breakwaters are becoming more and more unsustainable from an economic and ecological standpoint. - Saltmarsh planting is found to have most successful in mangrove habitat restoration projects.	- Instead of (or in addition to) artificial constructions, it has been suggested to create or restore natural ecosystems, such as sand dunes, salt marsh, mangroves, seagrass and kelp beds, and coral and oyster reefs. - Interdisciplinary research has been encouraged	Multiple areas all over the world	study 2018	[192]
Novel coastal protection approach of hybrid nature of mangrove plants and rock fillet habitats: By lowering rates of erosion, rock fillets safeguard riverbanks and promote the mangroves' natural re-growth.	- Investigates whether the functional similarities between natural mangroves and hybrid rock-fillet ecosystems (both the rehabilitated mangroves and rock-fillet coastline defense areas) are equal. - Compared to natural mangroves, eco-engineered mangroves provide diverse but functionally distinct habitats for estuarine organisms.	- The significance of natural habitats and their restoration is emphasized - Hybrid coastal defense structures provide a habitat that is not ecologically equivalent but rather enhances natural habitats - Due to the decline of mangroves globally and their growing susceptibility to climate change stressors, hybrid solutions should not be employed in place of replacing or restoring lost natural systems.	Three estuaries: Manning River, Wallis Lake, and Hunter River, Australia	2021	[222]

Table 9.8: Use of knowledge based classifier for mangrove assessment

Technique used	Type of data used	Task	Limitation/ Remarks	Location	Year Ref. No
K-nearest neighbor (KNN), Support vector machine (SVM), Classification, and Regression tree (CART)	Hyperspectral image data	Classification of Mangroves Species - Assessment of Tree Height	- Height information, in particular, proved useful for differentiating mangrove species with comparable spectral signatures - UAV hyperspectral imagery to classify mangrove species has been shown useful through experimental findings.	Zhuhai City, Guangdong, China	2018 [195]
Support Vector Machine (SVM)	Multispectral image data of Landsat, Sentinel 2A	- Loss or gain in mangrove cover(2000-2019) - mangrove fragmentation - Impact of fragmentation on Leaf Area Index, and Gross Primary Productivity	- Integrating fragmentation analysis with the results of the mangrove classification - Establishing uniform principles for the creation of better mangrove protection policies	Peninsular Malaysia	2021 [194]
Spectral Mapper Classification	Angle data of Landsat, Sentinel 2A	- Discriminating Mangroves Species and Assessing Health Mangrove canopy cover assessment - Mangrove forest structure assessment	- Assessment of species composition from satellite image is possible. - Interesting finding reported from the study: Avicennia stands landward, and Rhizophora seaward	Sundarban and Bhitarkanika	2019 [193]
Random forest classifier	Landsat data and two physical variables (Shuttle Radar Topographic Mission (SRTM), and Distance to Water).	- Study of the temporal and spatial distribution of mangrove and saltmarsh ecosystems - Study the transitional patterns of coastal wetlands into other land uses from 1991 to 2015	- Mangrove has lost 7.6%, and salt marsh has increased 20% due to the transformation of fresh/ brackish water to saline water	south-eastern Australia	2021 [196]

Table 9.9: Data mining applications on ecology and the proposed solution

Technique used	Type of data used	Task	Remarks/ Limitations	Location	Year	Ref. No
Association rule mining	Satellite-derived time series datasets	- Quantify the impacts of urbanization on mangrove changes using grid-based association rule mining - Non-numeric data of conservation policy can be incorporated into quantitative analysis	- The drawback is that the produced rules are highly dependent on the correctness of the source datasets, the calculation of the grid size, and the classification of the indicators.	China	2021	[48]
Association rule mining	Species presence/absence data	- Mangrove restoration via mitigating excessive soil salinity by salt marsh plantation - Finding association rules among co-existing mangroves and salt marshes, environmental parameters - Predicting probable species occurrence at particular sites	- Encourage a natural method of restoration - Proposes a unique data-driven management strategy for restoration - As rules are dependent on the correctness of the datasets, the generated rules have been validated against existing literature. - Study-based decision before plantation saves human resources.	Indian Sundarban Mangrove	2022	Present work

frequent inundation by seawater, causing changes in soil salinity; invasive species occurrence; anthropogenic disturbances. This study has presented a new restoration strategy for dealing with excessive salinity-affected mangrove habitats. A data mining approach has been followed to address this issue, and can efficiently automate the process of extracting frequent associations, and novel occurrences of species data. This approach shows an analytical way of study for mangrove forest restoration. The findings are supported by statistical visualization and past research findings on this topic. A potential strategy for ecosystem growth and restoration involves choosing and implementing multi-species in accordance with their tolerance of salt level and giving structure to frequently found species data at a study area. Thus introducing multi-species and maintaining biodiversity richness could be employed to stimulate the natural healing of ecosystem services. Furthermore, the paucity of species data in a particular ecosystem might lead to concerns about aspects like whether that habitat can accommodate more species or whether it is viable to increase species diversity in that specific area. The proposed framework also addresses this issue by predicting novel associations from association rules. This study contributes to the knowledge of the traits of co-occurring species and their potential relationship, which could help with both short- and long-term restoration efforts. It is, therefore, necessary to study the establishment, and long-term functioning of salt marsh and mangrove ecosystems using a comprehensive, interdisciplinary approach that takes into account both ecological and physical thresholds and bottlenecks. The only source of data for this study was field survey reports. Obtaining real-time data from inaccessible areas will significantly aid in the compilation of datasets. Remote sensing techniques may be useful in acquiring more accurate data analysis in the future. In the future, we would like to focus on accumulating data from satellite imagery and employing a data-driven learning algorithm to extract more accurate facts. A hybrid strategy of using image and text data in a data-driven learning framework should yield superior results. From the methodological point of view, the result set of generated rules can be optimized further. Domain knowledge can be utilized to exclude a key set of association rules, in addition to the support and confidence values for trimming association rules.

9.6 Summary

Unique characteristics of mangrove habitats are confronting multiple challenges such as frequent inundation by seawater, causing changes in soil salinity; invasive species occurrence; anthropogenic disturbances. This study has presented a new restoration strategy for dealing with excessive salinity-affected mangrove habitats. A data mining approach has been followed to address this issue, and can efficiently automate the process of extracting frequent associations, and novel occurrences of species data. This approach shows an analytical way of study for mangrove forest restoration. The findings are supported by statistical visualization and past research findings on this topic. A potential strategy for ecosystem growth and restoration involves choosing and implementing multi-species in accordance with their tolerance of salt level and giving structure to frequently found species data at a study area. Thus introducing multi-species and maintaining biodiversity richness could be employed to stimulate the natural healing of ecosystem services. Furthermore, the paucity of species data in a particular ecosystem might lead to concerns about aspects like whether that habitat can accommodate more species or whether it is viable to increase

species diversity in that specific area. The proposed framework also addresses this issue by predicting novel associations from association rules. This study contributes to the knowledge of the traits of co-occurring species and their potential relationship, which could help with both short- and long-term restoration efforts. It is, therefore, necessary to study the establishment, and long-term functioning of salt marsh and mangrove ecosystems using a comprehensive, interdisciplinary approach that takes into account both ecological and physical thresholds and bottlenecks.

KNOWLEDGE DISCOVERY IN BIODIVERSITY (KDB): WEB APPLICATION PROTOTYPE IN ECOLOGY

10.1 Introduction	133
10.2 Exploring KDB	134
10.3 Application	139
10.4 Datasets	139
10.5 Comparison with data mining tool	142
10.6 Comparative discussion on multiple previous case studies	142
10.7 Summary	143

10.1 Introduction

Biodiversity and Ecosystem Informatics [223] is a new and interdisciplinary field. Its promising advances have already been recognized by biologists, natural resource managers, and computer scientists. Data mining includes algorithmic processes and computational paradigms that assist computers in recognizing patterns in databases, performing predictions and estimates, and often improving their behavior through data cooperation. Data mining is becoming increasingly important in engineering and information systems, and it has been effectively used to address a wide range of scientific and technical concerns. Because of their ability to extract knowledge, frequent itemset mining, frequent closed itemset mining, and other methods are widely used in data mining [132, 133, 37, 62, 61].

As of now, the significance of frequent itemset mining in the field of biodiversity knowledge extraction has not been clearly defined. However, beyond the market basket analysis, it has clear uses in the field of bioinformatics [37] for discovering noteworthy patterns. In bioinformatics, GUI-based apps have been discovered to be common data mining tools [224]. This type of approach has yet to be discovered for the analysis of biodiversity data. We coined the term "computational biodiversity" [56], which refers to the application of computational methods to ecosystem conservation and biodiversity research. The objective is to identify underlying knowledge that may be beneficial to the scientists in the field of biodiversity and ecology, foresters, stakeholder groups, and others. Originally, this concept relied on a variety of computational methodologies applied to main species diversity data pertaining to occurrence and presence/absence.

The purpose of the proposed software system is to introduce researchers to the facility that may be accessible by using algorithmic techniques to analyze biodiversity data. At first, the system merely delivers a rich selection of sophisticated algorithms for data mining jobs. Furthermore, KDB serves as a digital data repository that domain researchers can use as a data repository. Statistics and visualization tools, as well as data preprocessing operations, may be introduced in the near future to facilitate the use of graphical user interfaces.

The goal of this work is to introduce a prototype for incorporating data mining tasks into primary biodiversity data analysis. Figure 10.1 depicts the general block diagram for showing KDB's working flow. To begin, three fundamental algorithmic operations have been provided in this section to highlight the power of data mining algorithms. There are three types of mining: frequent itemset mining, frequent closed itemset mining, and association rule mining. For binary datasets, frequent closed itemset mining is synonymous with biclustering. Several assembled datasets are also included. By selecting a dataset, potential users can apply a specific algorithm.

In summary, the main contribution made to this work is:

1. presenting an experimental platform for researchers in the biodiversity sector.
2. including techniques that are specifically designed for working with binary species occurrence datasets, although they are relevant to other datasets as well.
3. providing primary biodiversity data those have been digitised

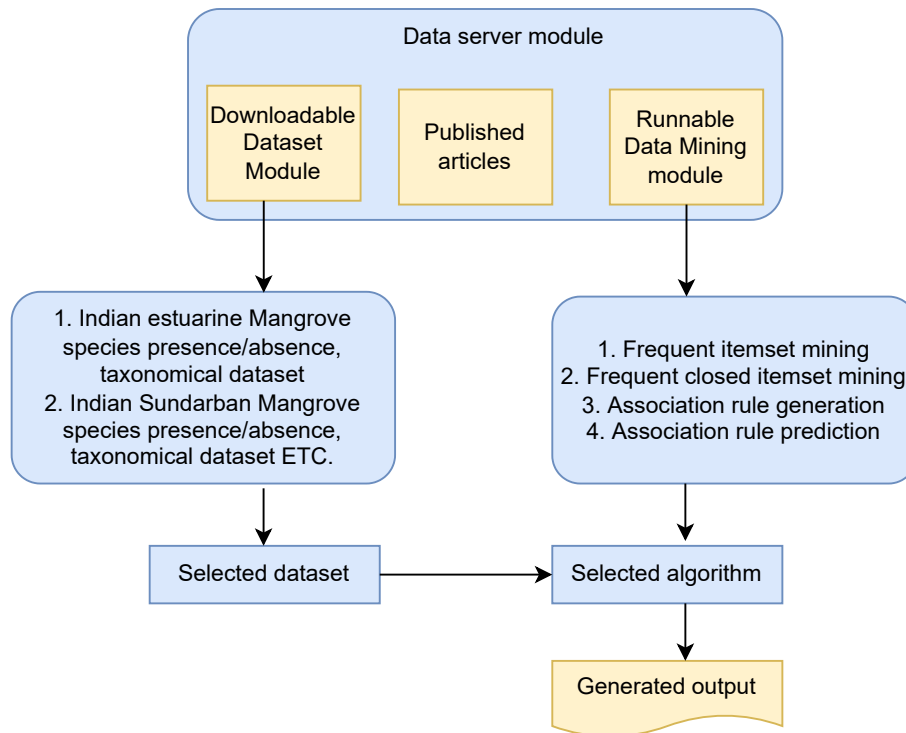


Figure 10.1: WorkFlow of the proposed work

10.2 Exploring KDB

The model is available at: <https://knowledgedb.ml>. It has five unique panels that are brought in by tabs at the top and perform five different functions. Figure 10.2 depicts the web application screenshot for the Home page. The web page for the Technical Documentation is highlighted in Figures 10.3, 10.4, and 10.5. Figures 10.6, and 10.7 depict web application screenshots for the Datasets, and Algorithms, respectively. The relevant journal and conference publications are listed and shown in Figures 10.8 and 10.9.

Figure 10.2 shows the Home screen, which presents the notion of Computational Biodiversity. The suggested field of work has been detailed in the About Us section. This panel also includes the section Technologies, which lists the technologies that were used to create this website. The members of this project are listed in the next section, Our team. A brief theoretical overview of the data mining approach is provided in the Technical Documentation panel that is displayed in Figures 10.3, 10.4 and 10.5. Introduction, Motivation and contribution, Data mining approach, Indian mangrove, and Sundarban mangrove are all aspects of this panel. Figure 10.6, Datasets panel, shows the options for uploading and downloading datasets. We began by compiling a few datasets. These datasets are intended for Indian mangrove and Sundarban mangrove primary biodiversity data. Figure 10.7 shows the Algorithms panel, which contains a list of data mining activities. It is structured in such a way that the user can select a specific data mining algorithm that corresponds to a specific operation. The user can then select the dataset to which the algorithm will be applied. The articles related to the data mining algorithms and application on biodiversity are listed under the Publications tab, which is seen in Figures 10.8 and 10.9.

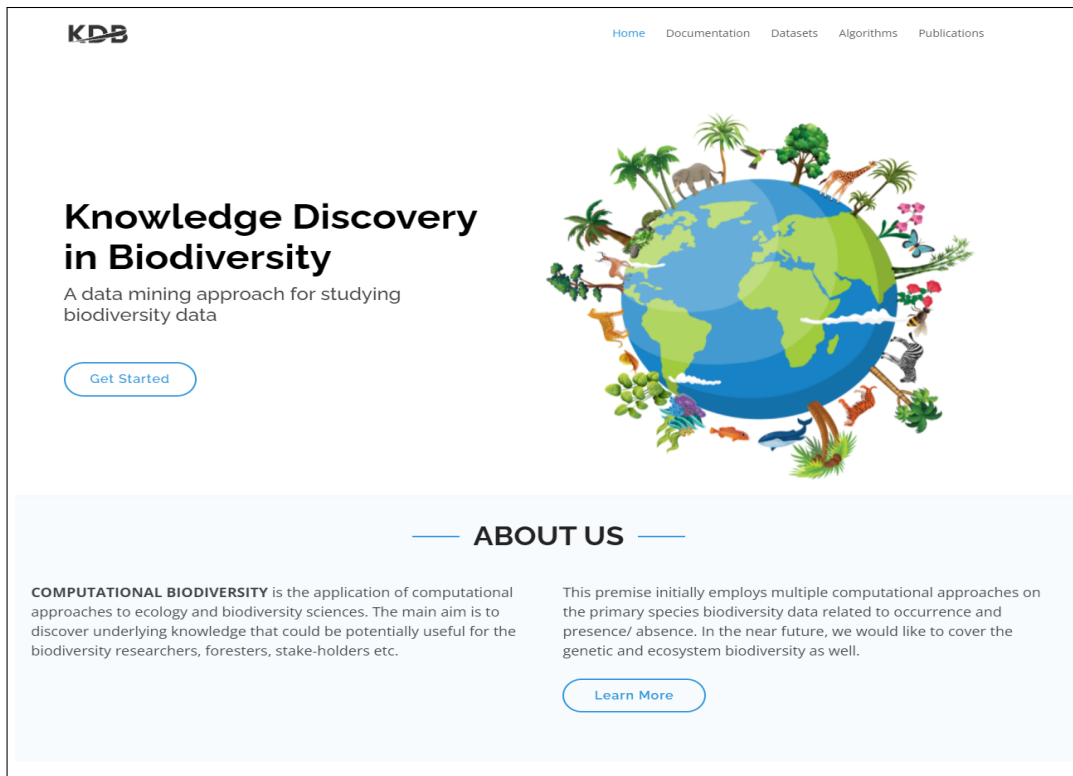


Figure 10.2: Homepage tab

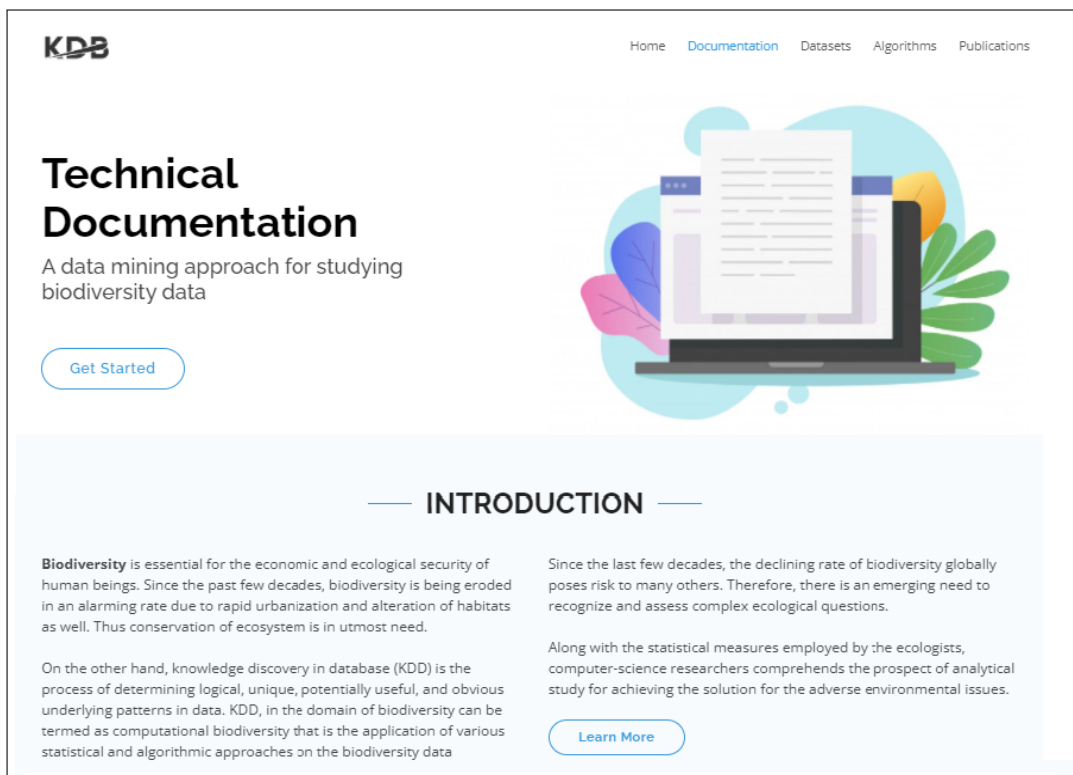


Figure 10.3: Technical Documentation tab (Introduction)

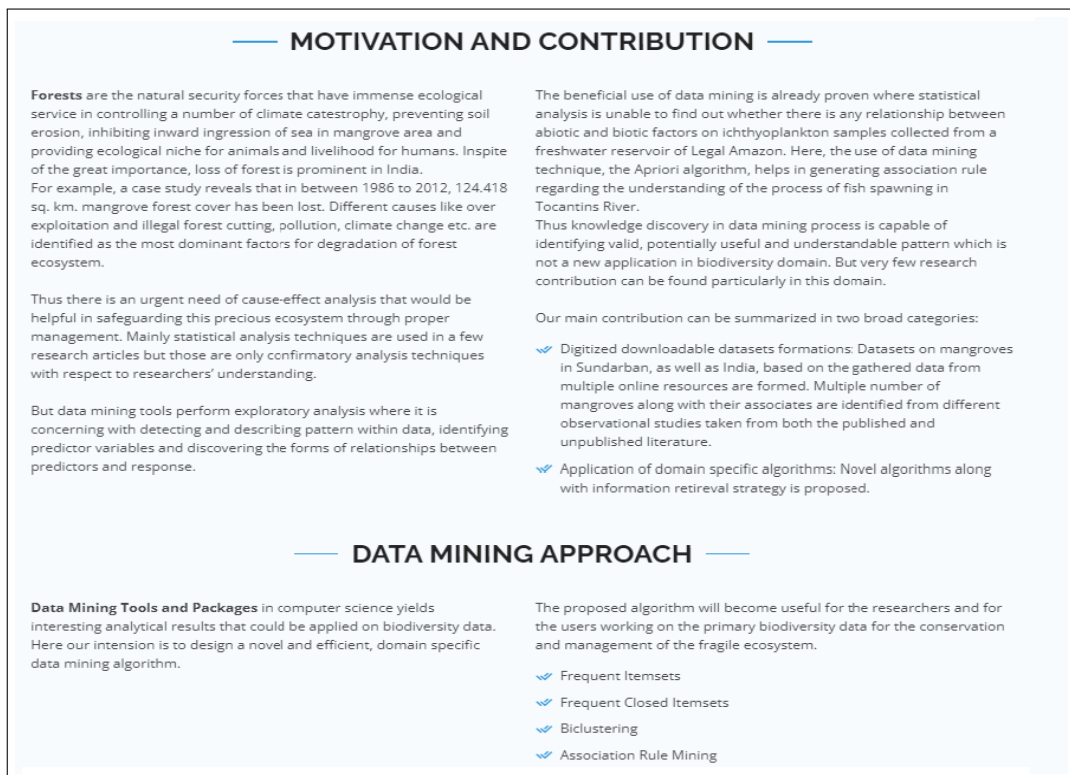


Figure 10.4: Technical Documentation tab (Motivation and contribution)

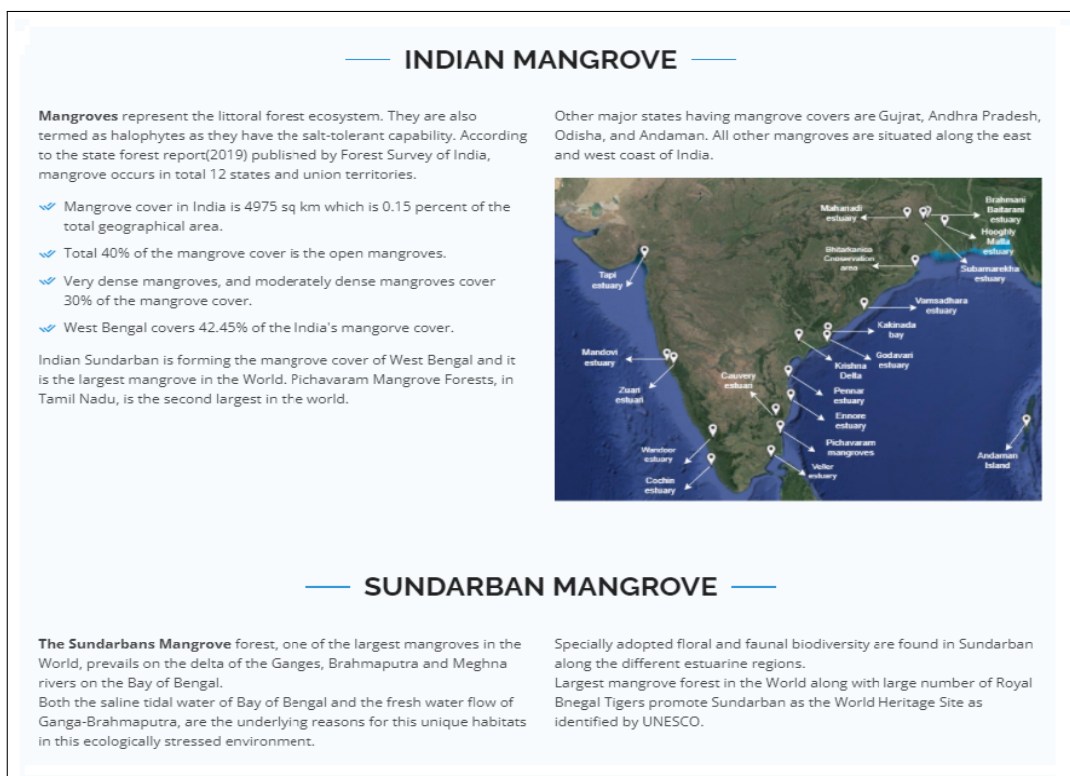


Figure 10.5: Technical Documentation tab (Indian and Sundarban Mangrove)

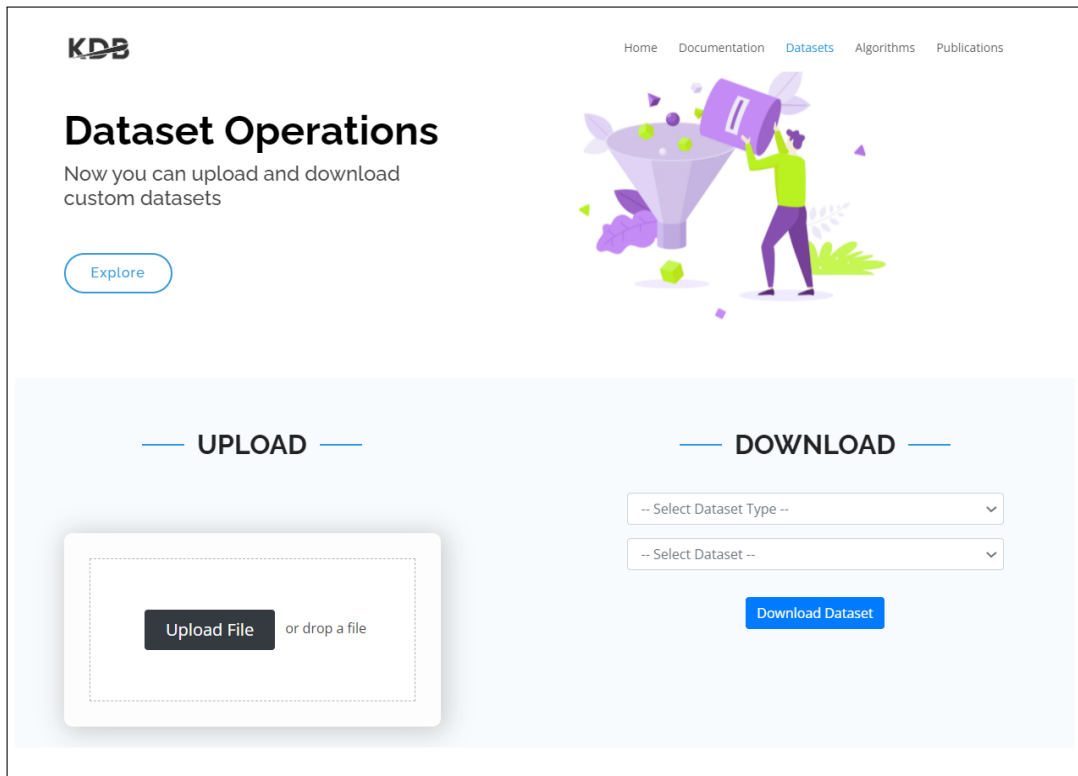


Figure 10.6: Datasets tab

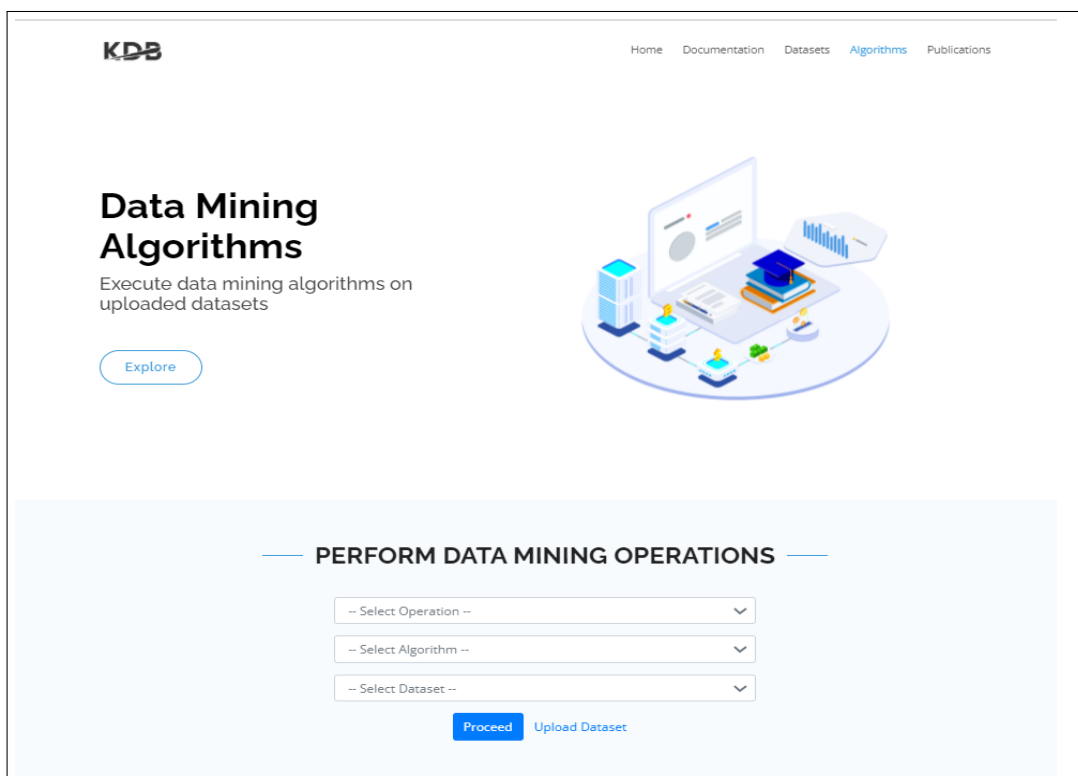


Figure 10.7: Algorithms tab

The screenshot shows the 'Publications' tab on the KDB website. At the top, there is a navigation bar with links for Home, Documentation, Datasets, Algorithms, and Publications. Below the navigation bar, the page title is 'Published Papers' with a subtitle 'Published papers related to the field of data mining and biodiversity data'. There is a 'Get Started' button and an illustration of a person looking at a large document. The main content area is titled 'JOURNALS' and lists four publications:

- Under Review (2022)**: CellBiClust: A Novel ICLA based Distributed FP-Tree Approach for Inclusion Maximal Biclusters. Authors: Moumita Ghosh, Pritam Sil, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)
- Under Review (2022)**: Frequent Itemset Mining Using FP-Tree: A CLA-based Approach and Its Extended Application in Biodiversity. Authors: Moumita Ghosh, Pritam Sil, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)
- Springer (2022)**: Knowledge Discovery of Sundarban Mangrove Species: A Way Forward for Managing Species Biodiversity." SN Computer Science, 3, no. 1 (2022): 1-14. Authors: Moumita Ghosh, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)
- Under Review (2021)**: An Irregular CLA-based Novel Frequent Pattern Mining Approach. Authors: Moumita Ghosh, Sourav Mondal, Harshita Moondra, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)

Figure 10.8: Publications tab (List of journals)

The screenshot shows the 'Publications' tab on the KDB website, specifically the 'CONFERENCES' section. It lists four conference publications:

- Springer (2021)**: Analysis of Indian Estuarine Data of Flora & Fauna", 2nd International Conference on Data Science and Applications (ICDSA 2021). Authors: Moumita Ghosh, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)
- Springer (2021)**: Computational Biodiversity", AISC Springer Series, Proceedings of International Conference on Advanced Computing Applications - ICACA 2021, Springer, 2020, pp. 1-6. Authors: Moumita Ghosh, Kartick Chandra Mondal. [View Publication](#)
- Springer (2021)**: Determining Dark Diversity of Different Faunal Groups in Indian Estuarine Ecosystem: A New Approach with Computational Biodiversity" In: J. K. Mandal and D. De (eds) Advanced Techniques for IoT Applications. EAIT 2021. Lecture Notes in Networks and Systems, vol 292, Pages: 147-158, Springer, Singapore. DOI: https://doi.org/10.1007/978-981-16-4435-1_16, Print ISBN: 978-981-16-4434-4, Online ISBN: 978-981-16-4435-1. Authors: Moumita Ghosh, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)
- Springer (2021)**: FCA based Constant and Coherent Signed Bicluster Identification and its Application in Biodiversity Study", AISC Springer Series, Proceedings of International Conference on Advanced Computing Applications - ICACA 2021. Authors: Moumita Ghosh, Anirban Roy, Kartick Chandra Mondal. [View Publication](#)

Figure 10.9: Publications tab (List of Conferences)

10.3 Application

The fundamental goal of KDB was to extract knowledge from primary biodiversity data. This section focuses on a few examples of successful applications of the proposed data mining methodologies on primary biodiversity datasets.

In [61], the application of frequent closed itemset mining on mangrove occurrence data was presented. To construct an off-the-shelf method to assess biodiversity presence/absence data, we used the FIST [37] approach, which employs association rule mining and biclustering approaches. The impacts of soil pH and water salinity on mangrove communities and biodiversity indices are investigated in this study. The association rules can estimate potential sites for mangrove species expansion by estimating the likelihood of introducing a new species to a certain location. Our research generates lists of commonly co-occurring species and supportive regions. It could aid in the restoration of mangrove ecosystems by identifying the most likely species missing from a certain region, possibly owing to extinction.

We demonstrated the efficient implementation of the combined technique of bi-clustering and association rule mining in [59] on a manually curated real dataset of flora and fauna. We create a set of criteria that ecologists can use to get a summary of closely occurring member lists, a predicted list of sites for member expansion, and so on. As a result, our findings may help to preserve estuarine variety, paving the way for future regional investigations.

The research in [62] combines data mining and statistics to lead us in the right direction for biodiversity restoration in a specific study region. This work recommended using the dark diversity function for the presence-absence dataset prior to the rule mining procedure. The purpose is to collect data on the missing part of the species occurrence data.

In [66], constant and coherent signed biclusters are identified utilizing a novel strategy for mining a multiple-signed dataset. We concluded that identifying sensitive regions and unprotected or endangered species using a signed bicluster retrieval from a spatiotemporal dataset of species versus region would be advantageous for biodiversity conservation. It would also help conservators/foresters to save or restore a declining species, a community, or even an ecosystem. Recently, [63] presented another key idea for an excessive salinity-affected mangrove community restoration technique in which hyper-salinity might be reduced by the establishment of suitable salt marshes. A case study of the Sundarban coastal area has been conducted, taking into account significant environmental/habitat characteristics such as salinity, pH, soil texture, and tidal amplitude, as well as data on the occurrence of mangroves, salt marshes, and mangrove associates. This study demonstrates a coexistence pattern among salt marshes, namely among mangroves and mangrove associates. Interspecies connections have also been hypothesized based on co-existence evidence.

10.4 Datasets

Another purpose of the KDB is to give primary biodiversity datasets of species. Exploration of unpublished data/statistics must be strengthened in order to identify the prevailing gap in knowledge of biodiversity. Primary data on biodiversity is a critical necessity for successfully completing ecosystem conservation [225]. However, these data are typically unavailable or difficult to obtain, and even when they are available, they are dispersed and unsuitable for the intended application [226]. Even when policymak-

ers require an integrated dataset to develop a strategic action plan, [227] becomes difficult. Through web searches and literature studies, several datasets, database systems, and articles were located using key terms such as “Indian mangrove”, “Indian estuarine mangrove biodiversity”, “Sundarban mangrove dataset”, and “Indian mangrove dataset”. The snowballing strategy was utilized to locate extra relevant study data. We looked at numerous prominent biodiversity data websites, such as Mangrove Reference Database and Herbarium (Dahdouh-Guebas F. (Ed.) (2021)), World Mangroves database (Accessed at <http://www.marinespecies.org/mangroveson2021-04-30.doi:10.14284/460>), Online database of Environmental Information System portraying mangrove cover of Indian states and territories (http://www.frienviis.nic.in/Database/Mangrove-Cover-in-India_2444.aspx). In addition, the most recent versions of biodiversity reports on Indian Mangroves were sorted for references to mangrove-specific information, including distinct statistics published by WWF on the State of the Art Report on Biodiversity in Indian Sundarbans [54], a report from Forest Survey of India (<http://www.frienviis.nic.in/Database/Mangrove-Cover-Assessment-20192489.aspx>), and [55] and [110] are two significant book sources of our findings.

KDB now has assembled and preprocessed datasets mostly consisting of Indian mangroves. Below in subsection 10.4, a brief description of the included datasets is given.

Dataset descriptions: The descriptions of the datasets are listed below and the brief metadata has been shown in Table 10.2.

1. *Presence/ absence of Indian estuarine mangroves:* This dataset contains presence/absence data of 34 mangroves along 19 estuaries situated on east and west coasts of India. These estuaries are Wandoor mangroves, Hooghly-Matla eatuary, Subarnarekha estuary, Brahmani-Baitarani estuary, Bhitarkanica estuary, Mahanadi mangroves, Vamsadhara estuary, Godabari estuary, Kakinada bay, Krishna delta, Pennar estuary, Ennore estuary, Cauvery estuary, Pichavaram Mangroves, Cochin estuary, Zuari estuary, Mandovi estuary, and Tapi estuary.
2. *Taxonomic details of Indian Estuarine Mangroves:* This dataset contains the taxonomic information (*Genus, Family, Order, Superorder, Class*) for 34 Indian mangroves. The data was organized primarily from the World Register of Marine Species (<http://www.marinespecies.org/aphia.php?p=taxdetails&id=211508>), the Integrated Taxonomic Information System (<http://www.itis.gov/>), the National Center for Biotechnology Information (<https://www.ncbi.nlm.nih.gov/taxonomy>), the Global Biodiversity Information Facility ([https://\(https://www.cbif.gc.ca/eng/\)](https://(https://www.cbif.gc.ca/eng/)).
3. *Presence/ absence of Indian Sundarban mangroves:* Multiple forest blocks comprise the Sundarban delta region of India [54]. The southern region is comprised of Bagmara, Gona, Mayadwip, and Ajmalmari. The northern blocks consist of Jhilla, Pirkhali, and Panchmukhani. Harinbhanga, Khatuajhuri, and Arbesi constitute the eastern blocks. Chottohardi, Matla, and Netidhopani form up the western blocks. Chamta, Chandkhali, and Goasaba comprise up the core blocks. The blocks that constitute the 24 Parganas (South) Forest Division are Herobhanga, Ajmalmari, Dhulibhasani, Chulkati, Thakuran, Saptamukhi, and Muriganga. 82 mangroves have been identified for this dataset.

Table 10.1: Dataset details

Sl No.	Dataset Name	Row data	No of Rows	Column data	No of Columns
1	Presence/ absence of Indian estuarine mangroves	Indian estuarine mangroves	34	Indian estuaries	19
2	Taxonomic details of Indian estuarine mangroves	Indian estuarine mangroves	34	Indian estuaries	19
3	Presence/absence of Indian Sundarban mangroves	Indian Sundarban mangroves	82	Indian Sundarban Blocks	22
4	Taxonomic details of Indian Sundarban mangroves	Indian Sundarban mangroves	82	Indian Sundarban Blocks	22
5	Indian estuarine data of fish species	Fish species	760	Estuaries	20
6	Indian estuarine data of flora and fauna	floral and faunal groups	26	Indian estuaries	20
7	Inner estuarine dataset of mangrove, saltmarsh, and environmental factors	Inner estuarine blocks	7	Salt marsh: 11, Mangrove: 16, Mangrove associates: 7, Other factors: 4	38
8	Middle estuarine dataset of mangrove, saltmarsh, and environmental factors	Middle estuarine blocks	7	Salt marsh: 11, Mangrove: 12, Mangrove associates: 4, Other factors: 4	31
9	Outer estuarine dataset of mangrove, saltmarsh, and environmental factors	Outer estuarine blocks	8	Salt marsh: 11, Mangrove: 5, Mangrove associates: 0, Other factors: 4	20

4. *Taxonomic details of Indian Sundarban Mangroves:* This dataset contains the taxonomic information (*Genus, Family, Order, Superorder, Class, Phylum*) for 82 number of Indian Sundarban mangroves. The taxonomic data was organized as mentioned before.
5. *Indian estuarine data of fish species:* We have curated a presence-absence binary dataset of fish from Indian coastal locations [110]. There are 20 major estuaries recognized throughout India's long coastal area, and 762 fish species occurrence data are displayed. 15 estuaries are taken from India's east coast. Among them are HooghlyMatla, Baitarani-Brahmani, Mahanadi, Rushikulya, Bahuda, Vamsadhara, Nagavali, Godavari, Krishna, Penner, Ennore, Adyar, Veller, and Cauveri. The west coast of India is defined by the rivers Cochin, Zuari, Mandovi, Tapi, and Narmada.
6. *Indian estuarine data of flora and fauna:* Along the long coastal area of India along the east and west coasts, 20 estuaries have been considered. Hooghly-Matla, Subarnarekha, Baitarani-Brahmani, Mahanadi, Rushikulya, Bahuda, Vamsadhara, Nagavali, Godavari, Krishna, Penner, Ennore, Adyar, Veller, and Cauveri: these 15 estuaries are from the east coast. Cochin, Zuari, Mandovi, Tapi, and Narmada are situated at the west coast of India. Along the estuaries, occurrence data of 23 faunal groups and 3 floral groups have been curated [59].
7. *Inner, middle and outer estuarine dataset of mangrove, saltmarsh, and environmental factors*

These datasets of outer estuarine blocks, middle estuarine blocks, and inner estuarine blocks contain the existing record of estuary-specific distinct salt marshes, mangroves, and mangrove associates data, as well as other environmental parameters (such as salinity, pH, soil texture, and tidal amplitude) across the columns. The rows represent the recognized blocks for the outer, middle, and inner estuarine zones.

Table 10.2: General characteristics of a few data mining tools

Tool name	Source	Prog. lan- guage	GUI/ Command line	Main purpose
WEKA	https://www.cs.waikato.ac.nz/ml/weka/	JAVA	Both	Mining of general data
RAPID MINER	http://rapidminer.com	JAVA	GUI	Mining of general data
Orange	http://orange.biolab.si	JAVA	Both	Mining of general data
FIST	https://sites.google.com/site/mrkartickchandramondal/direct-links/implementations-and-tools/trovefist?authuser=0	JAVA	Command Line	Mining of bioinformatics data
KDB	https://knowledgegedb.ml	JAVA	GUI and Manual	Mining primarily biodiversity data

10.5 Comparison with data mining tool

The modern world is heavily reliant on data, the majority of which is available in both structured and unstructured formats. Because much of the data is unstructured, a process and system are required to extract the information that is useful from the data. Material is also necessary to convert it into a comprehensible and useful format. There are several tools available for data mining activities that extract data efficiently. Some useful open-source data mining tools are listed below and briefly highlighted in Table 10.2 follows :

Weka has four different functionalities. These are the command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. Weka is also primarily focused on classification and regression problems rather than descriptive statistics and clustering methods. Support for big data, text mining, and semi-supervised learning is also currently restricted [228].

RapidMiner has a visually appealing and user-friendly graphical user interface. It provides several statistical graphs as well as an application wizard that provides pre-built workflows for a variety of data mining activities [228].

Orange Canvas provides an organized view of supporting features divided into multiple categories, such as classification, regression, evaluation, association, etc.

FIST [92] supports the data mining application of frequent closed itemset mining and association rule mining for both binary and textual datasets. Primarily, it was derived for the genetic dataset in the bioinformatics domain. But it can be applied to any other dataset provided in a specific format.

KDB is under development and primarily built for the biodiversity domain. The proposed data mining algorithms can be applied to any other dataset too. The user-guided R snippets are specifically related to the data mining tasks on species data.

10.6 Comparative discussion on multiple previous case studies

This section briefly discusses the comparison among the reported results from multiple case studies. This part is listed in Table 10.3. The goal of the study, datasets under consideration, findings, and significance are all underlined here.

Table 10.3: Comparative discussion among the reported results from multiple studies

Reference	Aim	Dataset used	Findings	Significance
[63]	The goal of this research is to find a novel mangrove restoration approach by evaluating the frequent co-existence status of salt marshes, mangroves, and mangrove associates in various zones of deteriorated mangrove patches for species-rich propagation.	Datasets on mangrove, saltmarsh, and mangrove associates. Refer to dataset 7, 8, and 9 in Table 10.1	<ul style="list-style-type: none"> - Co-existence pattern of salt marsh and salt marsh along the salinity gradient - Co-existence patterns of salt marshes, mangroves, and mangroves associates with various environmental conditions - Likely inter-species association based on current co-existence data 	Understanding the distribution characteristics of salt marsh, mangrove, and mangrove associates, as well as environmental factors, can aid in decision-making. This paradigm is valuable for both academia and stakeholders, particularly environmentalists and protection authorities, in controlling salt marsh expansion and mangrove restoration.
[229]	By linking the concept of dark diversity, this work aims to reveal the standpoint of computational biodiversity as a counterweight to biodiversity loss.	Indian estuarine data of fauna is studied here. Refer to dataset 6 in Table 10.1.	Using dark diversity computation on the dataset before processing rule mining jobs allows us to examine the likelihood of faunal occurrence in a totally absent part. These rules would create possible habitat for several faunal groups.	This research has presented the proposition of using the dark diversity function to the presence-absence dataset before the rule mining procedure. The reason for this is to obtain information about the missing portion of the occurrence data.
[61]	Interpreting biodiversity data, exploratory data analysis via box plot visualization, z-score normalization, histogram analysis, and data discretization, use of data mining on species data	Use of Indian Sundarban data on mangroves at 29 sites and the environmental factors. [Data source: [105]]	<ul style="list-style-type: none"> - The mangrove vegetation is more influenced by salinity than pH. - Finding information about the mangrove ecology in particular, - Identifying potential plantation locations based on salt content, - Conducting a species diversity analysis, and observing how salinity affects N2Div modulation 	Interdisciplinary collaboration has already been adopted in research. The current study stresses and attempts to demonstrate the utility of data mining in biodiversity data analysis. A complete knowledge of raw data, exploratory data analysis, and therefore data conversion in system-relevant format is demonstrated.
[59]	Understanding the significance of the estuarine ecosystem, this work focuses on Indian estuarine data analysis of flora and fauna and demonstrates the effective application of data mining in such analysis.	Indian estuarine data of fauna (occurrence data and presence/absence data) is studied here. Refer to dataset 6 in Table 10.1.	In addition to the discretized categorical data, we created and evaluated the dataset displaying presence-only data. This study reveals finer details for each class concerning their level of occurrences, and co-occurrences, and for each estuary concerning their consistency in diversity.	This kind of information assists ecologists in maintaining estuarine biodiversity by formulating policies and appropriate procedures.

10.7 Summary

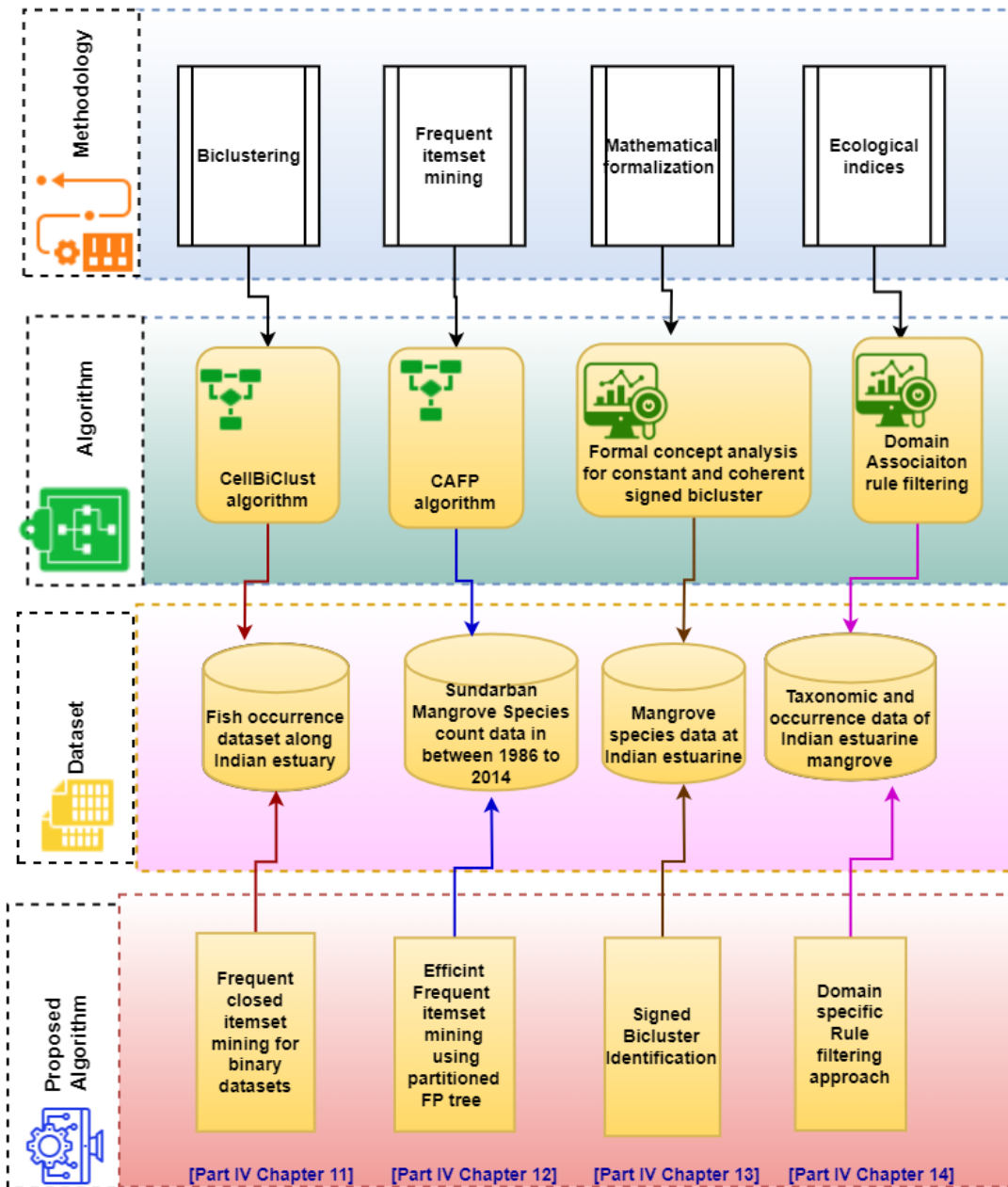
To summarize the subject, computational methods adoption in the field of biodiversity is extremely substantial and in demand, since it entails the ability to deal with heterogeneous and big-scale data. These methods provide precise, accurate indicators as well as the ability to deal with large amounts of data efficiently. In a nutshell, KDB can be viewed as a building prototype for the application of data mining approaches to ecology.

Despite the fact that Weka, Orange, and RapidMiner provide the majority of the desired qualities for a fully functional and adaptive platform, KDB can be considered of as a new addition to this set of data mining techniques when it comes to ecological datasets. Unlike other tools, KDB also acts as a data repository of primary biodiversity data.

Part IV

Proposed Algorithm

The methodology, algorithm, and dataset utilised in each of the proposed algorithm described in the chapters are shown below:



- Reusing existing algorithm / dataset
- Newly proposed framework/ algorithm / dataset

Outline for Part IV

CHAPTER 11

 FP TREE-BASED FREQUENT PATTERN MINING
 USING CLA

11.1	Introduction	148
11.1.1	Problem and motivation	148
11.1.2	Previous studies	149
11.1.3	Contribution	150
11.2	Approach	150
11.2.1	Algorithmic explanation	151
11.3	Completeness and correctness of the proposed algorithm	159
11.4	Result discussion and evaluation	161
11.4.1	Compared algorithms	161
11.4.2	Theoretical performance analysis	161
11.4.3	Test datasets	162
11.4.4	Running environment	163
11.4.5	Empirical performance comparison	163
11.5	Summary	167

11.1 Introduction

Mining frequent itemset is the fundamental task for association rule mining in data mining. Knowledge enrichment can be facilitated by finding interesting associations among the elements of a transaction dataset. Making decisions and creating patterns from the vast amounts of data thus becomes simple. One of the main issues with frequent itemset mining methods is their high resource requirements. Thus it become unsuitable for larger datasets having high cardinality. In recent years, a few efficient data structures have been proposed for mining frequent itemsets showing a cohesive mining technique for decreasing execution time while still providing the possibility for additional improvements in memory requirements.

In this work, we propose a novel technique for frequent itemset mining that is effective in terms of time and memory needs by utilizing multiple FP tree structures and cellular learning automata (CLA).

The effectiveness of the suggested method has been thoroughly tested by comparing it to the top algorithms, as well as utilizing real and synthetic datasets that are readily accessible to the public and created especially for pattern mining algorithms. It is clear that the suggested method is memory-efficient and executes in a comparable amount of time regardless of the size and density of the input datasets, proving its robustness. Along with the novel methodology's proposal for frequent itemset mining, its potential domain-specific use in the study of species biodiversity data has also been considered. Large occurrence records of species databases can be used to determine whether groupings of species are closely related. Understanding species co-occurrence in different locations may help with afforestation and reforestation-related ecology-related problems. It might be a step in the right direction for the beneficial application of computer science in the field of biodiversity.

11.1.1 Problem and motivation

Frequent pattern mining [230] is one of the essential tasks in multiple data mining algorithms (e.g. association rule mining, episode mining, classification, clustering) for extracting relevant patterns from large datasets.

Frequent itemset mining had been first proposed by [33]. It focuses on finding all those items that are bought together by customers from the customer transaction dataset. Since the proposal for frequent itemset mining, multiple studies have been performed in various directions ranging from the proposal of scalable, efficient data mining techniques to miscellaneous applications with diverse data types [231]. From market basket analysis [232, 233, 234] to text mining [235], in analyzing social network activities [236], in the bioinformatics domain for identifying the relevant patterns that occur frequently [92, 37, 237, 238, 133], in all real-life data, the usefulness is perceived [239]. From the perspective of algorithmic development, many algorithms have been proposed addressing various constraints along with various improvements as well as applications. Though the prominence of frequent itemset mining is noticeable to the data mining research, still it is a computationally expensive process.

11.1.2 Previous studies

The problem of mining frequent itemsets can be studied in two basic subparts to illustrate two main followed approaches in the last years. The first one is the breadth-first search-based approach that is used by Apriori-like algorithms for candidate generation and testing. The other is the depth-first search-based approach that uses a tree data structure, where the dataset is transferred to a compressed tree structure and then mined.

Frequent itemset mining with candidate generation In the apriori approach [33], the k^{th} itemset is used in this case to find all the $k + 1^{th}$ frequent itemsets. Thus the process of candidate generation and testing required multiple times dataset scans.

Improvements for apriori-based algorithms: In partitioning-based approach [240], first, the algorithm logically divides the dataset into several non-overlapping partitions and these are mined individually to generate large itemsets. It requires at most two passes. In sampling-based approach [241], a random sample is generated as a base for determining other frequent itemsets from the whole dataset in one pass only. But for accuracy, more passes are followed. Hashing-based algorithm Direct Hashing and Pruning (DHP) [242] proposed for efficient large itemset generation. Apriori scans the full dataset for every pass whereas, DHP reduces the number of times the dataset needs to be scanned and also reduces the dataset size. ApprioriTid [243] overcome the repeated scanning of the dataset. IApriori optimizes [244] both of the aspects of Apriori that lower its performance. It addresses the problems of multiple dataset scanning and huge candidate generations. Other major extensions over Apriori are: [245] has devised correlation-based mining, [246] proposes SPADE—a fast and efficient sequence pattern discovery algorithm, etc. All of the above-mentioned approaches generate a large set of candidate itemsets, and, in most cases, multiple scans to the dataset are required.

Frequent itemsets mining without candidate generation Depth-first search-based approach of FP-Growth algorithm [34] has solved this issue. Here the compressed prefix tree data structure saves the cost of a dataset scan.

Improvements for the FP Growth algorithm: Tree projection algorithm [247] uses the framework of a lexicographic tree for finding the frequent itemsets. OpportuneProject [248] combines the depth-first approach along with the breadth-first approach. A disk-based data structure named CFP Tree (Condensed frequent pattern tree) [249] was proposed for storing and querying frequent patterns. Though the FP tree data structure represents the dataset in a compact form, the problem with the FP Tree data structure has been felt as the tree size is becoming larger for a large number of transactions. Thus, the tree can not be adjusted in the main memory. However, it achieves an influential efficiency due to the previously mentioned compact FP Tree data structure.

Recently, the PPC tree data structure has been invented which stores the prefix, and post-fix data related to each node. Along with this tree data structure, two algorithms, namely N-Lists [250] and Node-List [251] are proposed. [252] has proposed a more efficient structure with a more compressed representation of the tree. Hence, more efficiency is there in representing frequent itemsets. Bitmap technique [253, 254] in the frequent pattern mining process uses bit table data structure, which is an improvement over FP tree-based methodologies. A new data structure Dynamic superset bit-vector structure has been proposed

recently in [255]. It generates frequent closed itemsets and their lattice structure showing the parent-child relationship. Though Dynamic Bit table data structure is memory efficient, it could not be appropriate to some datasets specifically where the binary representation of all the rows is not possible.

Frequent itemset mining from the vertical layout of the dataset In ECLAT (Equivalence class transformation algorithm) [256], a vertical transaction list is associated with each item where it has occurred in the transaction. Along with the minimization of the required dataset scan, they also devise an efficient lattice-based searching procedure.

Parallel mining approach Parallel approach [257] address the problem of mining with a skewed dataset. It can be said that a parallel and distributed mining-based approach can deal with the problem of apriori, FP-growth, and bit-table-based approaches. In this case, parallelism may be secured either by partitioning dataset [258] or by partitioning the process of operation [259]. [254] proposes an efficient systolic-based parallel mining algorithm SABMA. [35] proposes a more efficient cellular learning automata-based parallel frequent itemset mining approach using the proximity list data structure. Here, the proximity list has the problem of redundancy and causing huge computation time and memory requirements. In most of the studies, algorithms are not memory-efficient when the itemsets are with larger cardinality.

11.1.3 Contribution

Though multiple studies have already been focused in this direction, the lacunae are not adjusted completely. A deep insight into the existing literature can emphasize the potency of the FP Growth algorithm. This motivates us to exploit the FP Tree in such a way that it can be employed in a memory-efficient manner. The major contribution can be listed as follows:

1. To propose a novel algorithm for finding frequent itemsets using a novel partitioned FP Tree structure based on the concept of cellular learning automata.
2. A Dense dataset (DDS) has been proposed for a compact representation of the larger input transaction dataset. This optimizes both time and memory as a larger dataset scan can be avoided in successive mining steps.
3. The proposed approach has been compared with the classical as well as recent leading algorithms.
4. Apart from the proposal of a new model for generating frequent itemsets, we demonstrate the use of frequent itemset mining in biodiversity data analysis. Here, the dataset is newly generated for studying the presence of mangrove species in Indian estuaries.

11.2 Approach

Our proposed approach to find frequent itemsets is shown in Figure 11.1 and is named CAFF, or Cellular Learning Automata Based Frequent Pattern Mining.

The entire approach is broken down into 4 parts including the preprocessing stage. The transaction dataset is compressed to a dense dataset during the preprocessing stage, and redundant information is eliminated. After preprocessing, the CLA environment begins to work and initializes cellular automata cells. The CLA environment reads itemsets from the dense transaction dataset and transfers the itemsets to the appropriate cells. Every cell can process a list of itemsets simultaneously and update themselves. In this case, the CLA adheres to linear reward inactivity, or functioning without a penalty. Here, the neighborhood of the cells is produced in accordance with the itemsets in transaction and follows the structure of irregular cellular learning automata (ICLA). Additionally, the learning process in each cell is identical. So, the proposed cellular automata is uniform. Each cell acts as a root in FP Tree forming a neighborhood that updates itself after receiving the transactions from the environment. Finally, each cell examines its FP Tree. At this point, a conditional pattern base is created for each frequent-1 item. It is then pruned based on the threshold value defined by the users, followed by finding all possible combinations of frequent patterns. Finally, the list of all frequent patterns is sent to the environment.

11.2.1 Algorithmic explanation

The proposed methodology is explained using the transaction dataset presented in Table 11.1.

Table 11.1: An example of transaction dataset

Row ID	Itemset
1	A, B, C, D
2	A, D, E, F
3	A, B, C, D, G, H
4	A, B, C, D, F, E
5	B, C, E, F
6	B, D, F, G
7	A, D, E, F, G
8	A, D, E, F, K
9	A, B, C, D, E, F, H

Extracting Frequent 1-itemsets or $Freq_1$ (Step 2): To extract all the frequent 1-itemsets, the number of occurrences of each individual item is counted. If the count value for an item crosses the user-defined minimum support value, i.e. minsup, that item will be treated as a frequent 1-item. Otherwise, that item will be removed and will not be addressed anymore in subsequent phases. The list of all frequent 1-item forms frequent 1-itemset or $Freq_1$.

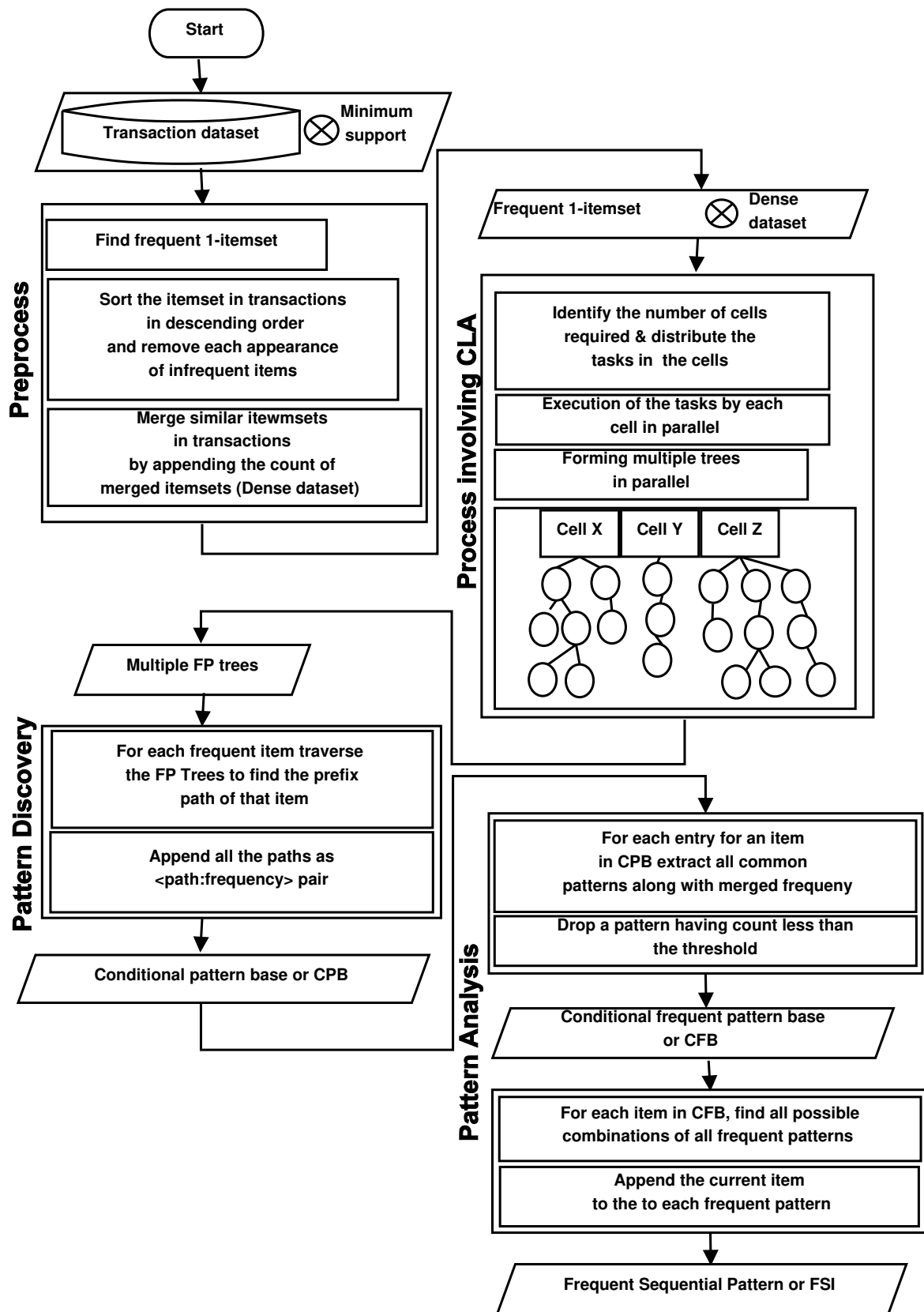


Figure 11.1: Flowchart for the proposed methodology

Algorithm 1 CAFPP**Input** TDB: Transaction dataset

Minsup: Minimum support threshold

Output FSI: List of all extracted frequent patterns in an itemset

```

1: procedure CAFPP
2:    $Freq\_1 \leftarrow$  The set of frequent 1-itemset with the count of occurrence of each item.
3:   DDS  $\leftarrow$  Dense dataset obtained from the original transaction dataset and the
    $Freq\_1$  as input.
4:   CellType  $\leftarrow$  A set containing the first item of each itemset in DDS.
5:   CLA  $\leftarrow$  Create cells for each of the distinct items in CellType
6:   Split DDS such a way that each cell is linked to a list of transaction itemsets starting
   with its CellType
7:   Each Cell process transaction itemsets in parallel and build and update FP tree as-
   sociated with each cell
8:   CPB  $\leftarrow$  Create an empty list to store the Conditional Pattern Base.
9:   for Each cell in CLA do
10:    Traverse along the corresponding FP Tree and find out the Conditional Pattern
    Base for each frequent 1-item and store it in CPB.
11:  end for
12:  CFB  $\leftarrow$  Create an empty list to store the Conditional Frequent Pattern Base.
13:  for Each item in CPB do
14:    Find the find out the Conditional Frequent Pattern Base and store it in CFB.
15:  end for
16:  FSI  $\leftarrow$  List of all frequent patterns FP generated from CFB.
17:  Return FSI
18: end procedure

```

The process for finding $Freq_1$ i.e. frequent 1-itemset:

1. An empty list of 1-itemset is created
2. For each itemset in the dataset
 - (a) For each item in that itemset
 - i. If the item is absent from the list, it is added with a count of 1
 - ii. Else the count of the item in the list is increased by 1
3. Once all the itemsets are scanned, thresholding is applied and the items whose frequency is less than the min sup value are removed from the list
4. The resulting list is the list of frequent 1-itemset

The list of all items along with their frequencies is given in Table 11.2. Considering the

minsup value as 5, the list of frequent 1-itemset along with their frequencies is given in Table 11.3.

Table 11.2: List of 1-itemsets

Item Key	A	B	C	D	E	F	G	H	K
Count	7	6	5	8	6	7	3	2	1

Creating Pruned and Dense dataset DDS (step 3): Once the frequent 1-itemset has been obtained, the original dataset is compressed by pruning the infrequent items. Pruning is used to eliminate work that adds to processing time but has no relevance to the output results. After eliminating the infrequent items, all the similar transactions have been identified. Transactions T_i and T_j are considered as similar if they have the identical itemsets. All the similar transaction entries have been merged by appending the count of the merged itemsets and forming a dense dataset (DDS). Thereby, memory space is saved, and computation time as well.

To explain this step, consider the example dataset in Table 11.1 and $Freq_1$ list in Table 11.3. Considering the sixth row in Table 11.1, since G is not present in the list of $Freq_1$, it has been pruned and D, F, B added to the DDS. A similar operation has been performed for each row of Table 11.1 and hence the DDS is obtained.

The DDS i.e. dense dataset is obtained in the following manner:

1. (a) Each itemset in TDB is sorted in descending order according to the frequency count of each item. This generates an updated TDB.
- (b) An empty dataset is created named DDS
- (c) For each itemset in the updated TDB
 - i. Create an empty list called temp to store an itemset
 - ii. For each item in that itemset
 - A. If it is present in the list of $Freq_1$, it is added to the temp
- (d) If the temp is not present in the DDS then it is added to it with a count of 1
Else its count is increased by 1 in the DDS

Creation of Cells in CLA environment and building multiple FP Tree structure (step 4 to 7): Each unique item at the beginning of each itemset will be identified once the DDS has been formed. All the distinct unique items will correspond to a cell in the CLA environment. The CLA environment reads each itemset of the dataset one at a time and

Table 11.3: List of frequent 1-itemset

Item Key	D	A	F	B	E	C
Count	8	7	7	6	6	5

Table 11.4: Dense dataset

Row ID	Itemset	Count
1	D, A, B, C	2
2	D, F, B	1
3	D, A, F, E	3
4	D, A, F, B, E, C	2
5	F, B, E, C	1

transfers it to the appropriate cell (where the front item of the itemset matches with the cell). The cells can operate simultaneously with one another after each of them receives itemset of the DDS. Each cell will eventually lead to an FP Tree. Each node of an FP Tree can be considered as a cell that poses a learning automaton. The learning automata updates its state depending on: i. The new itemset read ii. The present state of the neighbors and iii. The present state of the cell itself. Here the neighborhood of a node is denoted by the successor node/ nodes in the FP Tree. The occurrence count, associated with a node, denotes the state of that node. According to the received itemset, the cells update/ create new neighborhoods, and update their states in the FP Tree. Finally, the frequent items are extracted from the FP tree and transmitted to the environment.

A cell performs the following when an itemset and its respective iteration number from the DDS have been supplied to it:

1. If the cell is empty, create a tree by scanning each item in the itemset and adding it to the tree along with its frequency
2. Else increase the frequency of the root by the frequency count of the itemset
 - (a) Check if the next item in the itemset is present as a neighbor of the current item,
If true, then set the next item as the current item and increase the frequency by the frequency count of the itemset.
Else create a branch for the next item and set the frequency to the frequency count of the itemset
 - (b) Repeat this step for each item present in the itemset

This process can be illustrated using the dense dataset shown in Table 11.4. Here, D and F are identified as the distinct unique items situated at the front location considering all the itemsets in DDS. Hence, the cellular automata (Figure 11.2) has 2 cells, D and F, which are the roots of two different FP Trees.

The environment reads {D, A, B, C} having count 2. It is then transferred to the cell D (as it matches the front item). As cell D was empty initially, all the items in the itemset added and formed the FP Tree (Figure 11.3, top-left). Here neighborhood has been formed, for example, the neighbor of cell D is cell A, the neighbor of cell A is cell B, and the neighbor of cell B is cell C. The second itemset of the DDS is {D, F, B} with count 1. It has also been

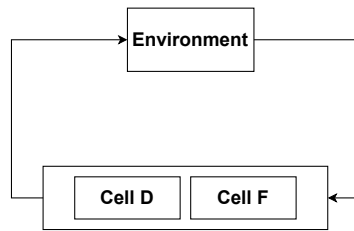


Figure 11.2: Cells based on unique items in frequent 1-Itemsets

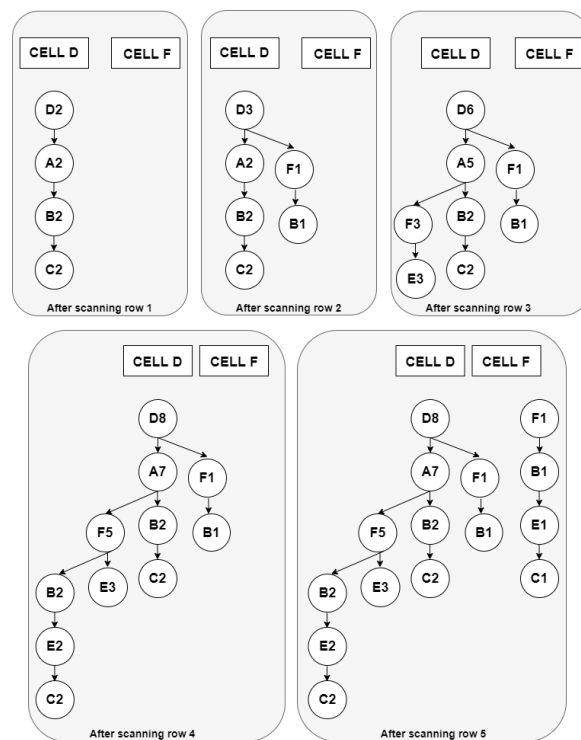


Figure 11.3: Generating CLA based multiple FP Tree structure

sent to cell D (Figure 11.3, top-middle) and the neighborhood has been updated accordingly. Here, both A and F are forming a neighborhood with D. Likewise, itemsets 1 to 4 in the DDS will be adjusted in the FP Tree rooted by cell D and itemset 5 will be adjusted in the FP Tree rooted by cell F. After reading the itemsets in consecutive order in the DDS, the state of the CLA is shown in Figure 11.3.

Generating and Pruning the Conditional Pattern Base CPB (Step 8 to 11): After all the itemsets have been added to the multiple FP Tree structure, the CPB is generated for each item present in $Freq_1$.

Table 11.5: Generating the CPB

Item	CPB
C	$\langle DAFBE : 2 \rangle, \langle DAB : 2 \rangle, \langle FBE : 1 \rangle$
E	$\langle DAFB : 2 \rangle, \langle DAF : 3 \rangle, \langle FB : 1 \rangle$
B	$\langle DAF : 2 \rangle, \langle DA : 2 \rangle, \langle DF : 1 \rangle, \langle F : 1 \rangle$
F	$\langle DA : 5, D : 1 \rangle$
A	$\langle D : 7 \rangle$
D	NULL

To generate CPB, the following steps are performed:

1. Create an empty list named CPB
2. For each item itm (in the reverse order of their frequency count) in the $Freq_1$
 - (a) Create an empty map itmCPB
 - (b) For each cell in the CLA
 - i. Traverse along the tree edges and keep on storing the item that is being visited in a list v-list
 - ii. If the current item is itm then update the map itmCPB with $\langle v\text{-list} : \text{count} \rangle$ where the count will be equal to the count for the itm.
 - (c) Add itmCPB to CPB

The generated CPB is shown in Table 11.5. CPB for frequent 1 itemsets is stored in the reverse order of their maximum frequency count. C is occurring at both of the FP Trees rooted by cell D and F. After scanning all the tree branches, pattern bases for C: $\langle DAFBE : 2 \rangle, \langle DAB : 2 \rangle, \langle FBE : 1 \rangle$, which are generated by following three paths containing C.

Generating the Conditional Frequent pattern base (CFB): (Step 12 to 15): CFB contains all the frequent patterns that are generated from the CPB.

To generate CFB, the following steps are performed:

1. Create an empty list named CFB
2. For each item itm (in the reverse order of their frequency count) in the $Freq_1$
 - (a) Create an empty map itmCFB
 - i. For all the entries e in itmCPB corresponding to the itm, find out all the common item patterns from all the sets or subsets of e, and the count would be the summation of all the respective subsets, and append them to itmCFB.
 - ii. Drop a pattern having a count less than the threshold.
 - (b) Add itmCFB to CFB

Table 11.6: Generating the CFB (threshold is 5)

Item	CFB
C	$\langle B : 5 \rangle$
E	$\langle DAF : 5 \rangle, \langle F : 6 \rangle$
B	$\langle D : 5 \rangle$
F	$\langle DA : 5, D : 6 \rangle$
A	$\langle D : 7 \rangle$
D	$\langle \rangle$

Here the CFB has been generated for threshold 5 and shown in Table 11.6. It contains all the entries as in CPB. Considering the entries for item C in CPB, $\langle DAFBE : 2 \rangle, \langle DAB : 2 \rangle, \langle FBE : 1 \rangle$; the common item patterns are: $\langle DAB : 4 \rangle, \langle FBE : 3 \rangle, \langle DAFBE : 2 \rangle, \langle B : 5 \rangle$. These could be found out from all possible combinations of the entries. The counts are summed up for the common item patterns. Infrequent item patterns are dropped. For the example considered here, only $\langle B : 5 \rangle$ is found to be the frequent pattern as it has a count value greater or equal to the threshold.

Generating the Frequent sequential itemset (FSI) from the CFB (Step 16 to 17):

To generate FSI, the following steps are performed:

1. Create an empty set FSI (Frequent sequential itemset)
2. For each item itm in itmCFB
 - (a) Retrieve the respective row in the CFB
 - (b) Find all possible frequent pattern bases of all the items present in that row
 - (c) Append itm to each combination and add it to FSI

We obtain the final list of FSI from CFB. To illustrate it, let us consider the CFB for item C. Only B is present. Hence, the frequent pattern would be BC. Considering item E, the corresponding entries in CFB are DAF, and F. Thus, the generated FPs will be DAFE, DFE, DAE, AFE, DE, AE, FE. The generated list of Frequent Patterns is stored in Frequent sequential itemset(FSI) (Table 11.7). As the sorted order of occurrences of all the items is maintained in all the steps, starting from the dense dataset formation, tree building stage, generation of CPB, CFB, and FSI, all the frequent patterns are in sequential order of the decreasing order of frequency. So, the set FSI contains the result of all the frequent sequential itemsets present in the input dataset.

Comparison with respect to FP-Growth Algorithm There are several advancements over the FP-growth approach that make our proposed algorithm memory-efficient without compromising the time requirement as compared to the classical as well as recent algorithmic developments for mining frequent itemsets.

1. It builds a highly compact dense dataset (DDS) from a larger input transaction dataset, which is considerably smaller than the initial dataset and therefore saves the cost of

Table 11.7: Generating the FSI:

Item	FSI
C	<i>BC</i>
E	<i>DAFE, FE, DAE, AFE, DFE, AE, DE</i>
B	<i>DB</i>
F	<i>DAF, DF, AF</i>
A	<i>DA, A</i>
D	–

larger dataset scans in the successive mining processes. DDS contains only the frequent items. Hence, the removal of infrequent items from the dataset saves processing time and memory as well.

2. It involves a partitioning-based method that dramatically reduces the number of transaction rows that have to be scanned to build a single tree. The proposed method distributes the transaction rows among multiple cells, which serve as the roots of the corresponding trees. Multiple trees can be built in parallel, which optimizes the time requirement.
3. Several optimizations have been incorporated in the algorithm implementation step. A hashmap data structure has been used that supports constant-time performance in $O(1)$ and internal usage of highly efficient hash tables. Instead of stack-based recursive methodology, iterator-based steps are followed for the intermediate power-set generation. This optimizes the memory usage.

Additionally, there is no extra overhead to maintain a single-rooted long FP Tree. In CAFPT, the itemsets are represented by a highly condensed and much smaller structure named DDS that can facilitate the mining process of frequent itemsets. Multiple FP tree structures can be built in parallel in our case, which lowers the execution time and those are implemented in an optimized way to extract the frequent itemsets.

11.3 Completeness and correctness of the proposed algorithm

There are several properties of multiple FP Tree structures that can be obtained from the FP Tree building process.

Property 1: *The items in the frequent itemset are ordered in the support descending order.*

Rationale: A higher frequency of an item indicates that more number of transactions are sharing the item. Hence, more tree branches are sharing the item, implying that they are situated closer to the root of the trees. As the algorithm is designed in such a way that the multiple FP Tree structure can be stored in a memory-optimized form, the support descending order is maintained in DDS, hence in tree branches.

property 2: *Count of any path along the tree branches is the count for the last item along the path.*

Rationale: As the items are stored in the support descending order in the FP Tree (property 1), the frequent patterns generated following the tree branches, would also have a sequence of support descending order. Hence, the last item along the path will have the lowest minimum support count with respect to the other items. So, it can be said all the items along the path have co-occurrence for minimum count times.

Property 3: For any frequent item f_i , all the possible frequent patterns in the frequent itemset containing f_i , can be obtained by following all the branches of all the trees, starting from the root to the node containing f_i .

Rationale: Say, for a frequent item f_i occurs p times in the transaction dataset where $p >$ threshold t of minimum support. Therefore, there are tree branches in such a way that the summation of the count for all paths containing f_i would be equal to p (as it is understood that the multiple FP Tree structure covers all the frequent item maps from lemma 1).

Now, the conditional pattern bases, i.e., all the sub-pattern bases or prefix paths containing p have been derived. Following this, the conditional frequent patterns are built from where all possible combinations of the items are extracted. Hence, appending f_i with all combinations of items would result in all frequent patterns.

Property 4: Non-redundant set of frequent patterns are generated by the algorithm.

Rationale: Considering a single path e_1, e_2, \dots, e_k . Therefore, any sequence of items in this path would generate a frequent pattern with their co-occurrence frequency of the minimum support of that item in the frequent pattern. Since the items present in a single path are unique, no redundant patterns can be generated along the path. Thus, the property holds.

Lemma 1: Given a transaction dataset TDB. DDS is the dense representation of TDB. $Freq_1$ denotes the frequent 1-itemsets. $Freq_1(T)$ is the set of all frequent 1-items in the itemset of a particular transaction T . i.e., $Freq_1(T) = T \cap Freq_1$, and can be referred to as the frequent item map of transaction T . For each transaction in TDB, its frequent item map is projected on one of the edges of the FP Trees in multiple FP Tree structures.

Rationale: Let us consider an edge on k nodes, n_1, n_2, \dots, n_k , starting from the root to a leaf node in a particular FP Tree, let, c_k denotes the frequency count at the node labeled n_k . Then, the edge contains the frequent item map of c_k transactions. On the other hand, each edge corresponds to an entry in DDS, and the respective count column entry of the DDS will be equal to the number of similar frequent item maps throughout TDB, and it will be c_k for the edge n_1, n_2, \dots, n_k . Since DDS contains the compact representation of all the frequent item maps of TDB, and the multiple FP Tree structure is formed from the DDS, the final structure covers all the frequent item maps from all the transactions in TDB.

Lemma 2: Given a transaction dataset TDB and a support threshold t . DDS is the dense representation of TDB and $Freq_1(T)$ is the frequent item map of transaction T . The maximum size of the FP Tree is bounded by $\sum_{T \in TDB} |Freq_1(T)|$ and the maximum height is bounded by $\max_{T \in DDS} |Freq_1(T)|$.

Rationale: Based on the multiple FP Tree structure, for any transaction T , there exists a path corresponding to all the frequent item entries, i.e., for all the frequent item maps for all the transactions.

As, DDS contains all the frequent item maps for all the transactions in a compact form, and the trees are built from the DDS, so, the maximum size of the multiple FP Tree structure will be limited to the number of transactions in TDB. Since there are multiple sharings of the frequent items among the transactions, the size of the tree will be much smaller than the

actual number of transactions in the TDB. In the worst scenario, the size of the structure will be larger but limited to $\sum_{T \in TDB} |Freq_1(T)|$.

Each transaction will contribute to at most one edge of the tree and the length equal to the number of frequent items in that transaction. Let, the maximum height of the tree is h . Say, the maximum number of frequent items that appear in the transaction T of TDB is n . Hence, the length of the longest frequent item map is n , i.e., we can write $max_{T \in DDS} |Freq_1(T)| = n$. Therefore, the maximum length of any edge of the tree would be n . So, the maximum height h will be the same as n .

Lemma 3: Consider two nodes of a branch, ni , and nj . If $ni, nj \in Freq_1$, and $count(ni) \geq count(nj)$, i.e., (nj) is the child of (ni) , then, their common occurrence in transaction row is equal to the count of (nj) .

As the tree branches are in support descending order (property 1), if, (nj) is the child of (ni) , then $count(ni)$ must be $\geq count(nj)$. The count of a path indicates the number of times items along the path follow the co-occurrence pattern (Property 2). Therefore, it can be said that the common occurrence of the items in the transaction rows is equal to the count (nj) .

11.4 Result discussion and evaluation

11.4.1 Compared algorithms

To evaluate the performance of the CAFPP algorithm, we choose Goethals-eclat [260], FP-growth* [261], LCM [262], dFIN [252], and PrePost [250]. Goethals-eclat is the state-of-the-art algorithm in the category of vertical mining algorithms, and FP-growth* is the state-of-the-art algorithm in the category of FP Tree-based mining algorithms. Another efficient method of LCM is considered that mines all the frequent closed itemsets and outperforms Fp-growth*, Eclat, Apriori[33], and Mafia [263], and is a fast maximal frequent pattern miner. dFIN and PrePost are two recently proposed efficient algorithms based upon node sets for representing the frequent itemsets. dFIN has proved its supremacy among its family and other families of frequent itemset mining algorithms at present. PrePost has been taken as it proposed a novel PPC tree structure and outperforms FP-Growth* [261] and Goethals-eclat [260]. FP-Growth* and Goethals-eclat are the optimized versions of their respective categories. Hence, our test algorithms perfectly cover the classical and recent approaches for frequent itemset mining algorithms.

11.4.2 Theoretical performance analysis

The complexity of CAFPP algorithm depends on two factors: the number of cells and the height of the FP Tree in each of these cells. Let there be n items in the dataset. Hence, the maximum number of cells will be n , i.e., one cell for each item. The maximum height of a tree can be at most n , as one branch might contain all the n items. Hence the time needed to find the conditional pattern base will be $O(\text{time needed to choose the correct cell} * \text{time needed to reach the required item})$. This boils down to $O(n * n)$ from the above-mentioned arguments. Thus space complexity could also be explained in a similar way and it will be also in $O(n * n)$.

Considering Goethals-eclat algorithm [260], in the worst situation, there are $2^n - 1$ itemsets in the search space for n -item set. So the time complexity for finding patterns is $O(2^n - 1)$

* $O(m)$ which we shorten to $O(2^n * m)$. Space complexity is also the same. In case of FP tree, the complexity depends on searching for paths in FP tree for each item in the header table, which depends on the depth of the tree. For each of the conditional trees, n sets an upper bound on the tree's maximum depth. As a result, the order is $O(n * n)$, i.e. (number of entries in header table * maximum depth of the tree).

Both Prepost [250] and DFIN [252] build PPC tree that occupies a larger space. The dFIN framework is composed of three sequential parts. Building the PPC tree and creating a list of all commonly occurring 1-itemsets and their accompanying Nodesets is the first step. The PPC tree is searched for all frequent 2-itemsets and related DiffNodesets in the second stage. In the third stage, all frequent $k(> 2)$ -itemsets are mined. The complexity is therefore $O(l(n + m))$, where m and n are the lengths of the corresponding Nodesets, and let l be the number of derived DiffNodesets.

On the contrary, the N-list is compact since transactions with identical prefixes share the same PPC-tree nodes.

The complexity of intersecting two N-lists can be reduced to $O(m + n)$ by an efficient approach, where m and n are the respective cardinalities of the two N-lists is the intersection of N-lists rather than the counting of itemset supports. PrePost exploits the single path property of N-list to quickly discover frequent itemsets without discovering candidate itemsets.

LCM [262] has a simple structure and its time and space complexity are theoretically limited by a linear function of the quantity of frequently occurring closed itemsets. There is no requirement for binary trees or other intricate data structures. LCM is only implemented via arrays. For some sparse datasets, LCM outperforms other approaches because it runs fast. Unlike many other existing algorithms, LCM does not provide a method for reducing the dataset. Since there are many unnecessary items and transactions in dense datasets with smaller minimum supports, LCM performs poorly in these situations. So, along with time complexity, space complexity also depends on the dataset density.

11.4.3 Test datasets

Several real datasets have been used for the comparison of the algorithms.

The datasets are openly accessible benchmark datasets that can be used to test how well the algorithms work. The performance of the algorithms in various dimensions can be accurately assessed with the aid of variable dataset densities, transaction lengths, and attribute counts.

These are designed particularly for mining tasks and can be downloaded from the FIMI repository (<http://fimi.ua.ac.be>). The characteristics of the datasets in terms of a number of transactions, total unique items, average length of the transactions, and dataset density are listed in Table 11.8.

We consider 6 real datasets and one synthetic dataset. The real datasets are denser compared to the synthetic dataset T10I4D100K. This synthetic dataset is created by the IBM generator and can be downloaded from <http://www.almaden.ibm.com/cs/quest/syndata.html>. The experimental datasets cover a broad range of characteristics. For example, T10I4D100K, and Retail have many items and transactions, but the dataset density is 0.06% for Retail and 1% for T10I4D100K dataset. These datasets are comparatively sparser. Chess, Mushroom, and Connect datasets are highly dense datasets with densities of 49.33%, 19.32%, and

Table 11.8: Dataset characteristics

Dataset	# Transactions	# Unique items	type	Avg transaction length	Dataset density (%)
Accidents	340183	468	Real	33.8	7.222
Chess	3196	75	Real	37	49.333
Connect	67557	129	Real	43	33.333
Mushroom	8124	119	Real	23	19.328
Pumsb	49046	2113	Real	74	3.502
Retail	88162	16469	Real	10.3	0.063
T10I4D100K	98487	949	Synthetic	10	1.053

33.33%, respectively. Accidents dataset and Pumsb dataset can be categorized as middle-dense datasets with densities of 7.22% and 3.5%, respectively.

11.4.4 Running environment

To get a proper, justified comparison result, we have conducted all the experiments in similar hardware and software environments. All the experiments have been performed on a computer with 16GB memory and Intel i7-10870H @2.21GHZ, 4-core CPU. The operating system is Windows 10, 64-bit. For generating the time and memory graph for the algorithms under study, we use the publicly available implementation of these algorithms from SPMF, an open-source data mining library ([https:// www.philippe-fournier-viger.com/spmf/](https://www.philippe-fournier-viger.com/spmf/)), written in Java.

11.4.5 Empirical performance comparison

We have used time and space utilization as the measure for determining the performance of our algorithms. Both parameters have been calculated for the above-mentioned datasets with varying minimum support values. Thus for all the datasets, the test algorithms have been evaluated and their time and space utilizations have been compared in the sections given below.

Runtime analysis The generated graphs for runtime comparison are shown in Figure 11.4. In this figure, the X and Y axes represent the minimum support and runtime, respectively. Important behaviors are discussed below.

As we can see (Figure 11.4), except for the chess dataset, the performance of CAFP always remains good. For the other datasets, CAFP always shows a comparable runtime and there is no significant difference in run time between CAFP and other algorithms with varying minimum support values. In the case of the Chess dataset, for lower minimum support values, CAFP takes a longer time than most of the others which reduces eventually with higher minimum support values. From the algorithmic point of view, it can be explained as a higher value of minimum support reduces redundancy by efficient use of the dense dataset. Here, the stronger constraint filters the transaction dataset and helps to have homogeneous row entries, thus fewer rows in the dense dataset with higher row counts. This causes the algorithm to run faster. The opposite scenario happens for a lower minimum support value, i.e., a larger size of dense transaction dataset. Here, the chess dataset has the highest density

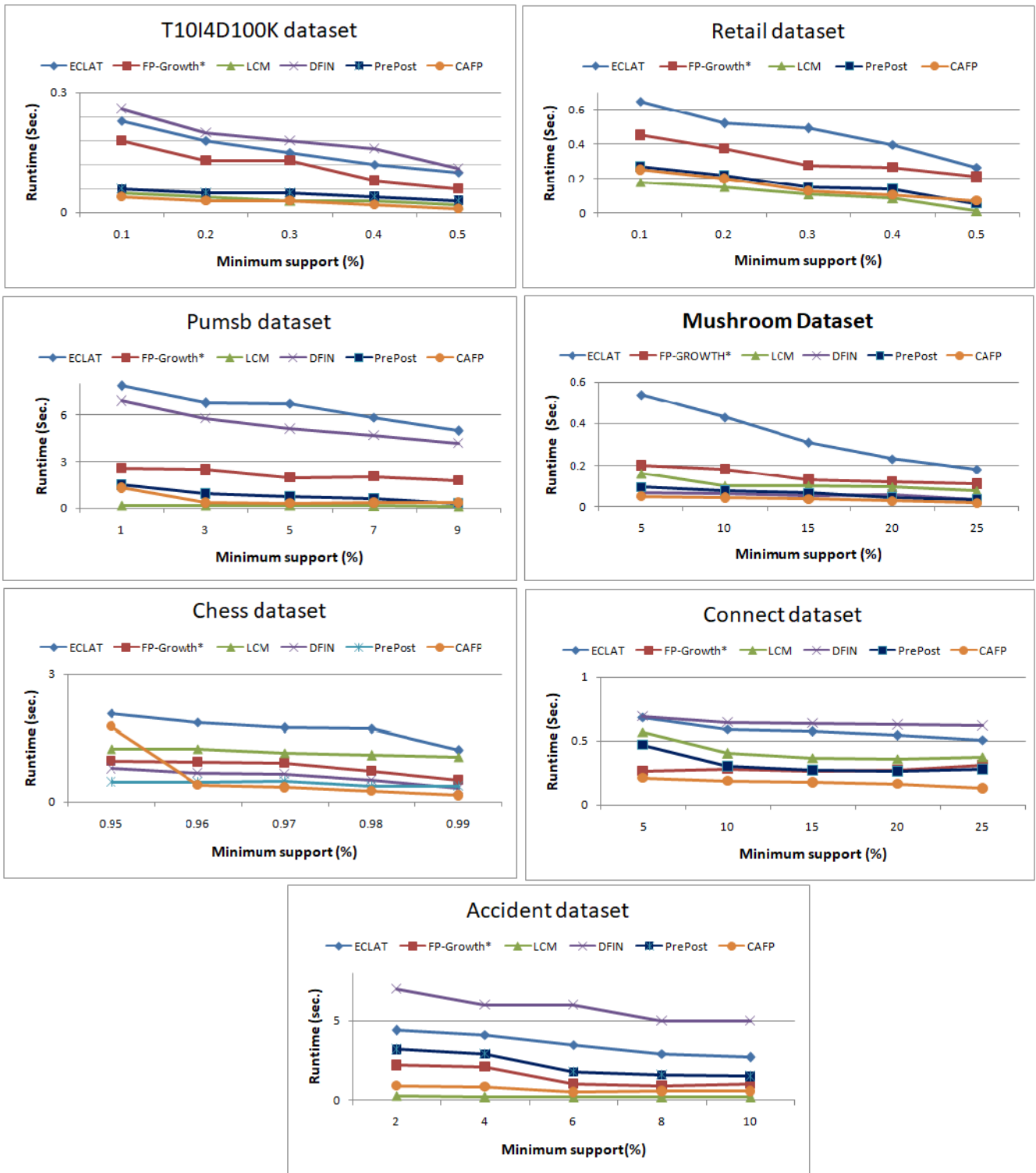


Figure 11.4: Result of execution time comparison on different datasets

of 49.33%. Hence, for lower minimum support values the dense transaction dataset is larger, resulting in larger computation time.

The average computation time for Pumsb is large. This is due to the large average length of the transaction row of the dataset. Accidents dataset has the second-largest average computation time. Both the dataset density and average transaction length are the influencing factors here. Although the Retail dataset has the maximum number of transaction rows and unique items, the average transaction length and density are low. Chess dataset has a much larger average transaction length and density. Hence chess dataset reports a comparable runtime with the Retail, although it has a much lesser number of rows and columns. Hence, it can be concluded that the runtime depends on multiple factors, such as the dimensions of the dataset, its average transaction length, and its density. The test datasets have covered a varying scenario mapping all these criteria for testing. In all cases, our approach proves its robust performance.

It has been seen that the performance of DFIN is poor where the number of unique items is large, for example, its performance degrades in Accidents, Pumsb, and T10I4D100K. Specifically, the Retail dataset, takes a huge time which is not comparable. Hence we did not plot it in the graph as for the retail dataset. ECLAT also takes a comparatively longer time for the lower minimum support count, which goes down for higher minimum support values. ECLAT stores a list of transaction IDs for each item. Thus it gives very fast support counting. But, for a larger number of transactions, intermediate transaction ID list generation becomes too large affecting the computation time. LCM has a lower runtime, but for some datasets (Retail, Pumsb, Accidents) it runs slightly faster than CAFP. A simple array data structure causes it to run faster. Prepost and FP-Growth* have comparable performance with CAFP. In some cases, FP-Growth* takes a longer time.

Memory requirement analysis Figure 11.5 shows the comparison for the memory requirement. Here, the memory consumption for each minimum support value is represented along the Y-axis. For all the minimum support values in the test cases, CAFP consumes much less memory than others in all the experimental datasets.

In ECLAT, for a larger number of transactions, intermediate transaction ID list generation becomes too large affecting the memory requirement. Thus, the performance of ECLAT is not good for larger datasets. The performance of LCM is better in a few cases. But, it is not significantly superior. Use Among the recent developments DFIN requires a little larger memory in most cases due to the complex and larger memory components of DiffNodesets. In most cases the memory requirement for DFIN and PrePost is almost the same. The memory consumption of CAFP always remains much less than dFIN and PrePost except for Accidents dataset. The huge length of the transaction and the higher average length of Accidents dataset create a huge memory component of the dense dataset along with the multiple FP Trees in CAFP. As all frequent 1-itemsets with associated DiffNodesets are generated by DFIN, along with the PPC tree, large memories are required when there are a lot of frequent 1-itemsets. As PPC tree contains both the pre-order and post-order information, its memory consumption is higher than FP Tree structure maintained by CAFP.

From the result of memory occupancy of all the algorithms, it can be seen that the performance of CAFP is always better than FP-Growth*. CAFP follows the rule for multiple smaller FP Tree creation rather than a single larger one. As the classical FP-Growth algorithm suffers from memory utilization problems, CAFP has shown a way out. DFIN and

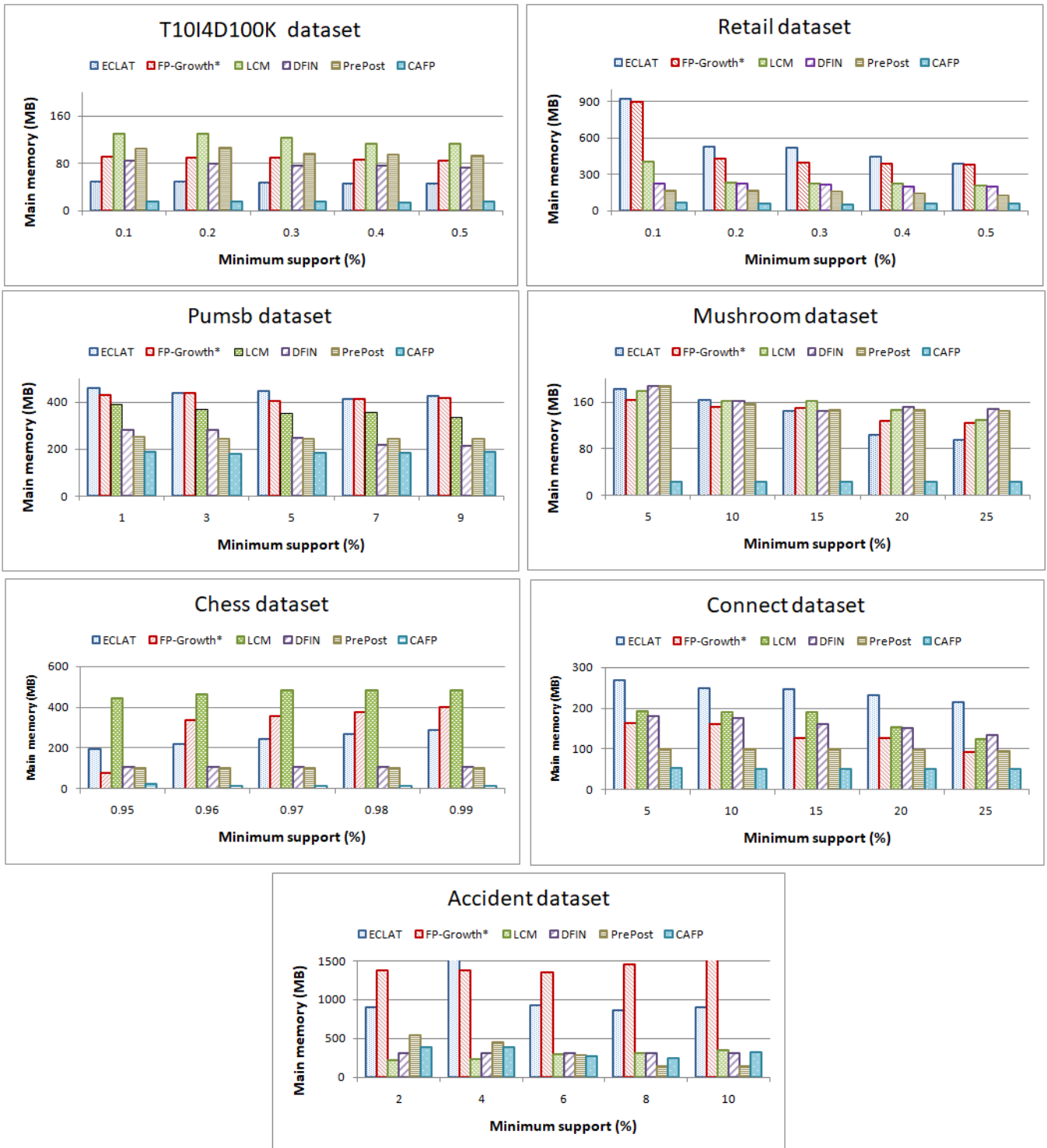


Figure 11.5: Result of memory requirement comparison on different datasets

PrePost need to build PPC tree which creates memory overhead for small datasets. Thus CAFPM requires less memory than DFIN and PrePost in smaller dense datasets. In the case of large datasets, if it is sparse then also CAFPM needs less memory than DFIN and PrePost. But for the large dense datasets, CAFPM requires comparatively more memory (e.g. Accidents dataset).

11.5 Summary

The importance of finding frequent itemset in data mining is well known. Still, the scope for making it faster with lesser memory consumption is a challenging task. We have proposed a novel and efficient parallel approach that wraps the concept of CLA in the multiple FP tree structure.

We have designed and implemented the proposed method, and examined its behavior both in priori and posterior analysis in comparison with several leading frequent pattern mining algorithms in large datasets. Our performance study illustrates that the proposed method mines patterns efficiently in large datasets, surpassing the memory requirements of all other leading algorithms.

FP-FOREST BASED FREQUENT CLOSED PAT- TERN MINING USING CLA

12.1	Introduction	169
12.1.1	Problem and motivation	169
12.1.2	Previous studies	170
12.1.3	Contribution	170
12.2	Proposed approach	171
12.2.1	Algorithm CellBiClust	171
12.2.2	Analysis of Algorithmic Result:	181
12.3	Performance evaluation	183
12.4	Application on ecological forecasting	187
12.5	Summary	190

12.1 Introduction

Biclustering is the process of simultaneous clustering of a set of samples along the rows and their attributes along the columns of a dataset. Binary datasets show a compressed and natural form of storing data about the relations between objects. Over the last few years, multiple biclustering algorithms have been built to be applied especially to binary datasets. Several methodologies, such as suffix trees, and matrix factorization, have been studied in literature to find useful biclusters from binary datasets. These discover interesting local coherent patterns along both the row and column dimensions.

In this work, we conceptualize, implement, and evaluate a novel and efficient algorithm CellBiClust that utilizes the tree data structure and exploits the concept of irregular cellular learning automata to incorporate parallelism in finding biclusters or frequent closed patterns from a binary dataset. Extensive experimental analysis on multiple datasets has proven the acceptability of our approach. The results achieved from diverse experiments with real as well as synthetic data expose the skillful behavior and robustness of CellBiClust to the density and size of experimental datasets. Finally, a comparison with respect to Bimax, the most cited and recognized binary biclustering algorithm for finding maximal biclusters, explains that the overall performance of CellBiClust is better while providing primarily similar results. Additionally, we show the utility of CellBiClust that can be extended to produce association rules. Both the biclusters and association rules, in turn, generate predictions for new relationships with certain probabilities. Although the market basket dataset and gene expression dataset are the two most commonly found application areas for frequent itemset mining and biclustering respectively, we exemplify it on the species occurrence dataset in ecology along with the validation from the end of domain expertise.

12.1.1 Problem and motivation

The storing of data in binary form is preferred in many research fields, covering data mining [264, 265], text mining [266], bioinformatics [92], biodiversity [110] etc. The values 0 and 1 are context-specific. For example, when carrying out with the sample of species, and features like occurrence sites, if a species s is found in a region that belongs to a region r , then the cell (s,r) of the dataset is equal to 1; otherwise, it is equal to 0 [110].

Although clustering is one of the most prevalent approaches to establishing the distribution patterns and underlying correlations in large datasets, the requirement of concurrent clustering along both the row and column dimensions, generates the concept of biclustering.

Considering a binary transaction dataset, where the items are along the rows and transaction IDs (TIDs) are across the columns, mining frequent closed itemsets (FCI) finds out all frequent itemsets across the row attributes for supporting column TIDs. A frequent itemset, listed in FCIs, is not included in a proper superset with the same support count. Similarly, for the same dataset, inclusion maximal biclusters, introduced with the Bimax algorithm [267], are the biclusters that are not entirely contained in any other bicluster. Hence, it can be said that in the case of a binary dataset, finding FCI is synonymous with finding all the maximal biclusters. Here, we are going to address the problem of mining inclusion maximal biclusters so that all the elements are 1. In addition to this, we discuss the usefulness of the extracted biclusters in finding association rules and predicting new associations.

12.1.2 Previous studies

A wide range of research contributions has been found in frequent pattern mining, though leaving it challenging with few research issues. The concept of frequent itemset mining was first introduced in [33]. After that, among several distinct variations, Apriori [268] generates a large number of candidates for generating frequent itemsets where multiple times dataset scans make it slower. [268, 240, 269, 270] follow apriori based methodologies. From the shortcomings of the candidate generation-based algorithms, FP Growth [34] is designed based on FP Tree data structure omitting candidate generation and thus carrying out the process in lesser computation time. FP tree is inefficient when the dataset is sparse [271]. Recently, cellular learning automata-based concept has been introduced in mining frequent itemsets for applying parallelism [35].

All of these generate a huge number of frequent patterns as these cover all subsets for all frequent patterns resulting in an exponential number of frequent patterns for large sets. Here, the concept for frequent closed pattern is suggested in [272]. The idea of frequent closed itemsets (FCI) overcomes the question of huge computational cost for frequent itemsets and therefore establishes a more efficient process without any information loss. [273] is one of the most efficient techniques for finding FCI where hashing is used on transaction IDs. [37, 133] are few important contributions in research for FCI where suffix-tree-based structure has been proposed. [274], [249] efficiently use FP Tree-like structures for finding frequent closed itemsets. A few other studies are mentioned in [231]. Independently of frequent closed itemset mining, the concept of biclustering came specially for gene-expression datasets. Multiple algorithms have been proposed in biclustering [36], and are used mainly in the bioinformatics domain. One of the most important concepts related to biclustering is the inclusion-maximal-biclusters [267], which came with Bimax algorithm. It reveals that for binary datasets, finding frequent closed itemsets is synonymous with finding maximal biclusters. Bimax follows divide-and-conquer approach with a rapid response. However, its performance degrades for dense datasets with a larger number of 1's, implying more recursive calls. Also, its subproblem formulation causes a large number of dataset splitting, hence memory usage is high.

Therefore, our motivation lies in proposing an approach that could prove its novelty in extracting biclusters in an optimized way.

12.1.3 Contribution

Encapsulating the above discussion, the contribution can be listed as:

1. Proposing an efficient and parallel mining algorithm for finding maximal biclusters/frequent closed itemsets on a binary dataset by taking advantage of FP-tree and irregular CLA.
2. Case study in one of the most promising employments of the proposed methodology by mapping it on the ecological aspect with a curated dataset of Indian estuarine fish occurrence and explaining from the ecological point of view.

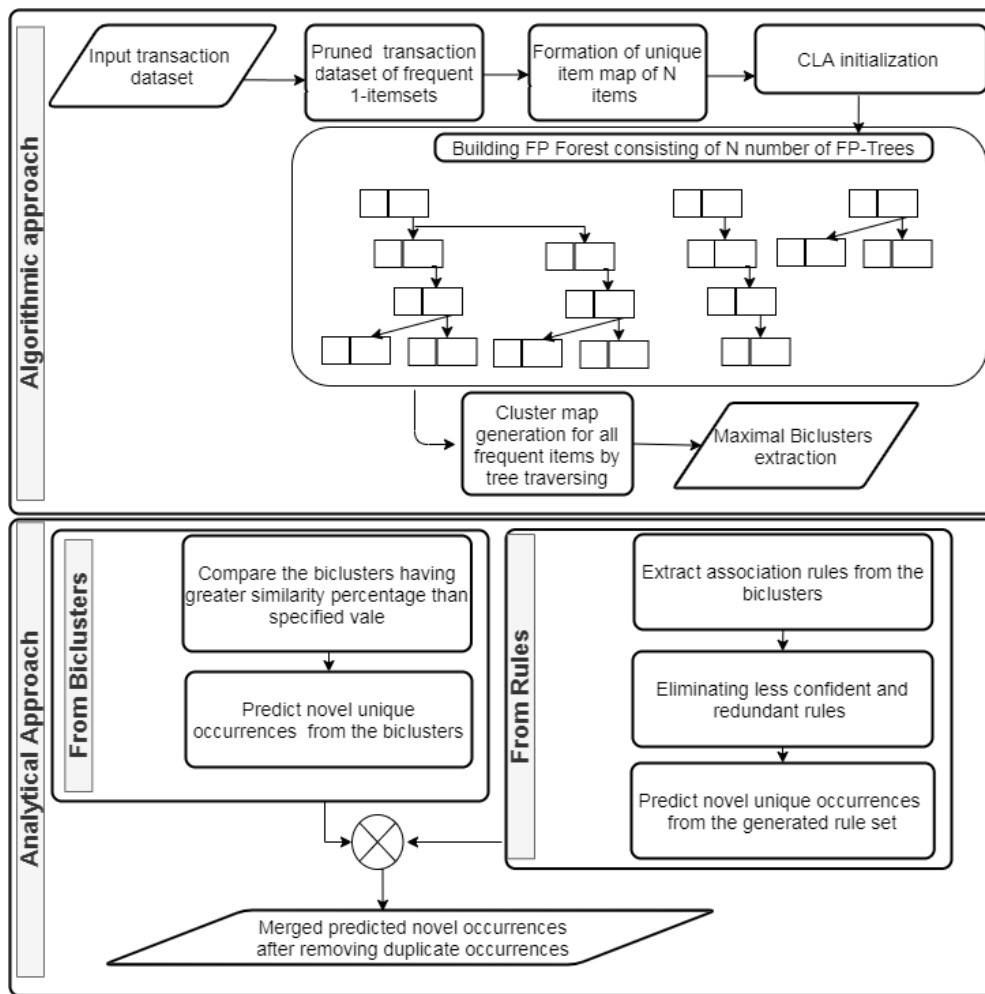


Figure 12.1: Outline for the experimentation performed

12.2 Proposed approach

The proposed algorithm is presented in the upper part of Figure 12.1 and the a respective explanation is given in section 12.2.1. The lower part of Figure 12.1 shows the procedure for analysis of the algorithmic result and is illustrated in section 12.2.2.

12.2.1 Algorithm CellBiClust

A tree-based data structure is always preferable to a matrix representation due to its compressed representation of the dataset. Additionally, parallelism is carried out in building n number of trees by adopting the concept of CLA. Therefore, we get twofold benefit by integrating CLA with a tree data structure for extracting the biclusters. The outline for our proposed algorithm is depicted in four major steps in Algorithm 2 which is further elaborated in the following four sections 12.2.1, 12.2.1, 12.2.1, and 12.2.1.

Algorithm 2 CellBiClust: Outline**Input** *TDS*: Transaction dataset

minItem: Minimum number of items in the cluster

minTID: Minimum number of supporting transactions for the items in the cluster

Output List of all maximal biclusters

- 1: **perform** preprocessing
- 2: **initialize** CLA environment and **construct** FP Forest
- 3: **extract** initial biclusters
- 4: **derive** final biclusters

Table 12.1: Binary representation of input transaction dataset: Items are denoted by the row names and the column names denote the transaction IDs.

Item\TID	2	3	4	5	6	7
M	1	0	1	0	1	0
N	0	1	0	1	0	1
O	1	0	1	0	1	0
P	0	0	0	1	1	0
Q	0	1	1	0	0	0
R	0	1	0	0	0	1
S	0	0	0	1	0	0
T	1	0	0	0	0	0

Perform preprocessing: Algorithm 3 shows the steps for data preprocessing. An example of binary input transaction dataset *TDS* is given in Table 12.1. The preprocessing steps are shown in Table 12.2 as per Table 12.1. The steps are as follows :

(Algorithm 3: Line no. 2) First, the count of occurrence for each item is computed that forms the list of *1-Itemsets*.

(Algorithm 3: Line no. 3-7) Then the items having a count greater than the user-defined threshold will form the *frequent 1-Itemsets* (*Freq_1-IS*). Here we use 2 as the threshold. Hence, the *Freq_1-IS* consists of M, N, O, P, Q, and R (Shown in Table 12.2).

(Algorithm 3: Line no. 8-12) After that, the *pruned and ordered dataset* (*PDS*) is formed according to the input transaction dataset. In *PDS*, each row contains the entry for a TID and its corresponding itemset (consisting of only the frequent items). Hence, S and T are infrequent and removed. The items of an itemset are stored in descending order of their frequency. For each of the input transactions (TID 1 to TID 7), *PDS* contains the corresponding frequent and ordered itemsets.

(Algorithm 3: Line no. 13-25) After the *PDS* creation, a map named *uniqueItemMap* is created having entries like < unique item: TIDs >, i.e. a unique item and all the TIDs where it occurs as the first item of an itemset. Here, M, N, and P are identified as unique items that occur in the front locations in the itemsets in *PDS*. Hence, *uniqueItemMap* has the entries

Table 12.2: Preprocessing steps of the input transaction dataset as given in Table 12.1

Item Key M	N	O	P	Q	R	S	T
Count 4	2	4	3	2	2	1	1
List of 1-Itemsets							
Item Key M	N	O	P	Q	R		
Count 4	2	4	3	2	2		
Frequent itemsets (<i>Freq_1_IS</i>): Threshold 2							
	TID	Itemset					
	1	M, O					
	2	N, Q, R					
	3	M, O, Q					
	4	P, N					
	5	M, O, P					
	6	P, R					
	7	M, O, Q					
Pruned and ordered dataset <i>PDS</i>							
Unique items present at front locations of the itemsets List of TIDs							
	M	1, 3, 5, 7					
	P	4, 6					
	N	2					
<i>uniqueItemMap</i>: unique items with TIDSs							

for M, N, and P along with the TIDs.

Initialize CLA environment and construct FP Forest: After the creation of *uniqueItemMap*, the CLA environment begins its processing. The process for this part is given in Algorithm 4. *Cells* are created for each unique item as in *uniqueItemMap*. Each *cell* has two fields: value and tid. Value contains an item in the itemset and tid consists of the related TID for the itemset. In this initialization step, the *cells* contains only the value fields, tid fields are kept null. We call these *parent cells*. Initially, 3 *parent cells* are created here, for M, N, and P. Each of these *cells* has local LA associated with them. LA of each *cell* knows the rule for building and updating the trees. The trees are built based upon the itemsets for the associated TIDs for each unique item as in *uniqueItemMap*.

The *cells* of a tree are forming a neighborhood that updates itself with each new itemset insertion. No strong grid neighborhood structure [35] like Von Neumann (4-neighbour) or Moore (8-neighbour) is followed here when an itemset is read. Here, the neighborhood of a *cell* can be thought of as an undirected graph. Hence, following the structure of an irregular cellular learning automata. Each tree can be updated simultaneously for an $\langle TID : itemset \rangle$ pair. The proximity among the items and their respective TIDs of occurrence can be visualized in trees. A few related terms have been introduced here, like *leaf*, *intermediate leaf*, and *FP Forest*. A *leaf* contains the last item of the itemset and therefore it has no successive *cell*. It stores the TID for the itemset in its tid field. In addition to a tid field that stores the TID, an *intermediate leaf* contains a link to the successive *cell* as well. In Figure 12.2, in the *final FP Forest* (highlighted in blue), O is an *intermediate leaf* with tid t1. Q is a *leaf* with tid t3, t7. As multiple FP tree-like structures are formed, we use the term, *FP Forest*. The steps for building *FP Forest*, as described in Algorithm 4, are discussed with an example below:

- i. (Algorithm 4: Line no. 2-5) *cells* are initialized for each unique item: M, N, and P.
- ii. (Algorithm 4: Line no. 6-24) As a *cell* is already created (*parent cell*) for the first item of an itemset, from the second item insertion and onwards, first, matching is performed to check whether the neighborhood already exists or not (Line no. 16-17). If exists, then it is followed, otherwise, a new *cell* is created as a new neighbor (Line no. 19-20). The item will be added in the value fields of the newly constructed consecutive *cells*. The last *cell* contains the corresponding TID (Line no. 12-13). For example, consider the unique item M, and its corresponding TIDs: 1, 3, 5, and 7. For TID 1, the itemset is M, O. For M, a *cell* is already created. For insertion of O, the existing neighborhood of M is checked. As no neighborhood exists, a new *cell* will be created as a neighbor of M, for item O. As O is the last item, its tid field contains the TID value of 1. For the *parent cell* M, the insertion of the second itemset for TID 3, i.e., M, O, Q, would first check the existing neighborhood of M, as O is already there, no new cell will be created. But, no neighborhood exists for O and Q. Hence, a new *cell* will be created for Q as a neighbor of O. The *cell* for Q will contain 3 in its tid field.

A pictorial view of the *FP Forest* is given in Figure 12.2. The final *FP Forest* is highlighted in blue.

Algorithm 3 CellBiClust: perform preprocessing

Input *TDS*: Transaction dataset

minItem: Minimum number of items in the cluster

minTID: Minimum number of supporting transactions for the items in the cluster

Output *PDS*: Pruned and ordered dataset; *uniqueItemMap*: a map comprising of unique items and supporting TIDs

```

1: begin
2:   count the support for each item in the dataset           ▷ List of 1-Itemsets
3:   for all items do
4:     if support > minTID then
5:       add the item in Freq_1_IS                           ▷ List of Freq_1_IS
6:     end if
7:   end for
8:   for each TID in TDS do                                 ▷ PDS creation
9:     remove infrequent items
10:    sort the frequent items by decreasing the support count
11:    add the itemset in the newly created PDS-row for the TID
12:  end for
13:  initialize a map uniqueItemMap                           ▷ uniqueItemMap creation
14:  for the first row in PDS do
15:    fItem ← first item in itemset; value ← TID
16:  end for
17:  for the second row onwards in PDS do
18:    for an entry in uniqueItemMap do
19:      if fItem is not a key then
20:        create an entry in uniqueItemMap where key = fItem and value = TID
21:      else
22:        set value = value ∪ TID
23:      end if
24:    end for
25:  end for
26: end

```

Algorithm 4 CellBiClust: Initialize CLA and construct FP Forest

Input *PDS* and *uniqueItemMap* as obtained from Algorithm 3

Output *FP Forest*

```

1: begin
2:   initialize an empty CLA
3:   for each item in uniqueItemMap do
4:     initialize a parent cell in CLA where cell.value  $\leftarrow$  item and cell.tid  $\leftarrow$  Null
5:   end for
6:   for each parent cell C in CLA do ▷ FP forest construction
7:     for each TID in TID list corresponding to C in uniqueItemMap do
8:       itemset  $\leftarrow$  read the itemset from PDS for the TID
9:       len  $\leftarrow$  |item.Set|
10:      curCell  $\leftarrow$  C
11:      for i = 0 to len-1 do
12:        if i=len-1 then
13:          Add TID to curCell.tid
14:          Exit loop
15:        end if
16:        if itemset[i+1] is in the neighbourhood of curCell then
17:          curCell  $\leftarrow$  the neighbour cell having itemset[i+1]
18:        else
19:          create a new cell with cell.value as itemset[i+1]
20:          curCell  $\leftarrow$  newly created cell
21:        end if
22:      end for
23:    end for
24:  end for
25: end

```

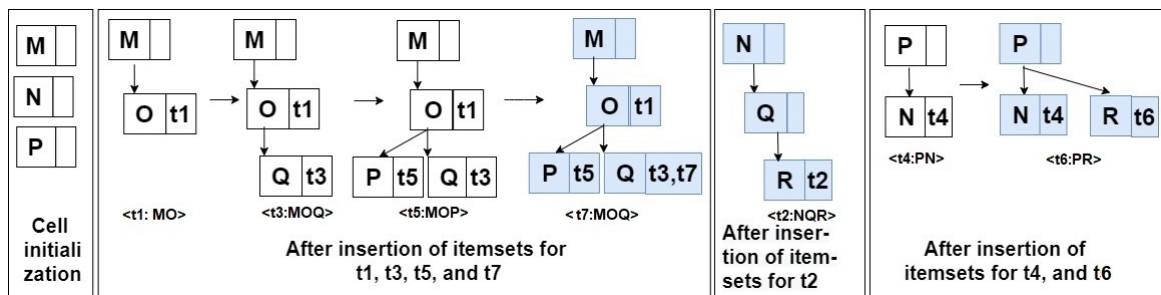


Figure 12.2: FP Forest building from the transaction dataset (The final FP-Forest is highlighted in blue)

Table 12.3: Generating *clusterMap* for the tree with *parent cell* M in final *FP Forest* shown in Figure 12.2

cur-Cell	vCells	if leaf	if intermediateLeaf	update clusterMap	Stack content	Comment
M	M	-	-	-	MO	O is the neighbor of M
O	MO	-	✓	$O \rightarrow \langle MO : t1 \rangle$	MOQP	Q, P are the neighbor of O
P	MOP	✓	-	$P \rightarrow \langle MOP : t5 \rangle$ $O \rightarrow \langle MO : t1, t5 \rangle$	MOQP	P will be removed from vCells and stack as it is leaf
Q	MOQ	✓	-	$Q \rightarrow \langle MOQ : t3, t7 \rangle$ $O \rightarrow \langle MO : t1, t3, t5, t7 \rangle$	MOQ	Q will be removed from vCells and stack as it is leaf
O will be removed from the stack followed by M as these are already in vCells						

Table 12.4: Generating *clusterMap* for the tree with *parent cell* N in final *FP Forest* as given in Figure 12.2

cur-Cell	vCells	if leaf	if intermediateLeaf	update clusterMap	Stack content	Comment
N	N	-	-	-	NQ	Q is the neighbor of N
Q	NQ	-	-	$Q \rightarrow \langle NQ : () \rangle$	NQR	R is the neighbor of Q
R	NQR	✓	-	$R \rightarrow \langle NQR : t2 \rangle$ $Q \rightarrow \langle NQ : t2 \rangle$	NQR	R will be removed from vCells and stack as it is the leaf
Q will be removed from the stack followed by N as these are already in vCells						

Extract initial biclusters: Algorithm 5 illustrates this step through *clusterMap* formation from the *FP Forest*. The term *clusterMap* has been introduced for storing the initial biclusters. The number of entries in the *clusterMap* will be equal to the number of frequent 1 items. It stores the initial biclusters in the form of $\langle \text{set of items: supporting TIDs} \rangle$ for each frequent 1 item separately.

Here, curCell is the current cell, vCells is the visited list of cells. The process for extracting initial biclusters will be performed for all the trees and will be updated on the same *clusterMap*. Table 12.3, 12.4, and 12.5 demonstrate the steps for generating *clusterMap* from the trees having parent cells M, N, and P, respectively. Considering Table 12.3, the following operations are performed as described in Algorithm 5,

- i. (Algorithm 5: Line no. 4-5) Let, the stack and vCells are initialized with the *parent cell* M. Let, the stack[top], i.e. M, is assigned to curCell. No entry is made for *parent cell* alone in clusterMap as we do not want any bicluster formed by a single item.
- ii. (Algorithm 5: Line no. 6) Now, repeat the following until the stack is empty.
- iii. The curCell M is not a *leaf/ intermediate leaf*, hence, no updation is done on the *clus-*

Table 12.5: Generating *clusterMap* for the tree with *parent cell* P in final *FP Forest* shown in Figure 12.2

cur-Cell	vCells	if leaf	if intermediateLeaf	update clusterMap	Stack content	Comment
P	P	-	-	-	PRN	R, N are the neighbors of P
N	PN	✓	-	$N \rightarrow \langle PN : t4 \rangle$	PRN	N will be removed from vCells and stack as it is leaf
R	PR	✓	-	$R \rightarrow \langle PR : t6 \rangle$	PR	R will be removed from vCells and stack as it is leaf
P will be removed from stack as it is already in vCells						

terMap.

(Algorithm 5: Line no. 23-25) Next, the neighbour of the curCell M will be pushed into the stack. So, O is pushed into the Stack.

(Algorithm 5: Line no. 26) Now, Stack[top] gives O, and we make it curCell.

(Algorithm 5: Line no. 27-29) So, for the curCell O, we append it to the vCells, and it will be MO.

(Algorithm 5: Line no. 7-15) As O is an *intermediate leaf*, the *clusterMap* will be updated for all the *cells* in vCells, except for the *parent cell*. Here, the *clusterMap* will be updated for O, only. So, <MO: t1> will be entered in the *clusterMap* corresponding to the frequent item O (Line no. 13). Here, MO is the only prefix of length greater than 1 of the current vCells MO.

(Algorithm 5: Line no. 23-25) Next, Q and P will be pushed into the stack as they are the neighbors of the curCell.

(Algorithm 5: Line no. 26-29) Stack[top] gives P, and we make it curCell. We append it to the vCells, and it will be MOP.

(Algorithm 5: Line no. 7-22) As P is a *leaf*, the *clusterMap* will be updated for all the *cells* in vCells, i.e. for P and O, except for the *parent cell*. *clusterMap* entry for P will be <MOP : t5> (Line no. 13), and for O, it will be <MO : t1, t5> (updated on the previous entry <MO : t1>) (Line no. 11). Here, MOP and MO, are the prefixes (of length greater than 1) of the current vCells MOP.

Being a *leaf*, the stack[top] and latest append of the vCells are checked for equality. If it returns true, then curCell will be removed from both the stack[top] and vCells, indicating that it has no more neighbor, and is already visited. So, P is removed from both (Line no. 16-21).

(Algorithm 5: Line no. 26-29) Now, stack[top] gives Q, and the latest append of the vCells is O, which are not the same. So, Q will be the curCell, and we append it to the vCells, and the vCells will be MOQ.

(Algorithm 5: Line no. 7-22) As Q is a *leaf*, the *clusterMap* will be updated for all the *cells* in vCells, i.e. for Q and O, except the *parent cell*. The *clusterMap* entry will be <MOQ:t3, t7> for Q (Line no. 13), and <MO:t1, t3, t5, t7> for O (updated on the previous entry <MO:t1, t5>) (Line no. 11). Here, MOQ and MO, are the prefixes (of length greater than 1) of the current vCells MOQ.

(Algorithm 5: Line no. 16-21) Being a *leaf*, the stack[top] and latest append of the vCells are checked for equality as before, and hence, Q will be removed from the vCells and the stack as well. Next, O and M will be removed sequentially after checking both the stack[top] and vCells as they are already visited.

(Algorithm 5: Line no. 30) Here, the stack is empty. Hence, the initial cluster extraction step for the tree with parent cell M is completed.

(Algorithm 5: Line no. 31) After all the tree branches are traversed in the *FP Forest*, the final *clusterMap* with initial biclusters has appeared in Table 12.6.

Derive final biclusters: The steps for generating maximal biclusters from the final *clusterMap* are given in Algorithm 6. The steps can be described as below:

i. (Algorithm 6, Line no. 2) An empty map named *biclusterMap* is initialized to store all

Algorithm 5 CellBiClust: Extract initial biclusters

Input *FP Forest* as obtained from Algorithm 4**Output** *clusterMap* : A map for storing initial biclusters of each item separately. Initial biclusters are in \langle set of items: supporting TIDs \rangle form

```

1: begin
2:   initialize clusterMap.
3:   for each parent cell P in CLA do
4:     initialize vCells and stack with P.value      ▷ vCells is a visited list of items
5:     curCell ← stack.top()
6:     while stack is not empty do
7:       if curCell is a leaf or intermediate leaf then      ▷ update clusterMap
8:         for each prefix S of vCells where  $|S| > 1$  do
9:           tempBiClust ← get entry for S[ $|S|-1$ ] in clusterMap
10:          if tempBiClust contains S then
11:            add curCell.tid to that entry
12:          else
13:            add (S,curCell.tid) to tempBiClust
14:          end if
15:        end for
16:        if curCell is leaf then      ▷ removing all already visited cells
17:          while vCells[ $|vCells| - 1$ ] == stack.top() do
18:            remove vCells[ $|vCells| - 1$ ] from vCells
19:            stack.pop()
20:          end while
21:        end if
22:      end if
23:      for each neighbor cell N of curCell do
24:        stack.push(N.value)
25:      end for
26:      curCell ← stack.top()
27:      if stack.top() not in vCells then
28:        append stack.top() to vCells
29:      end if
30:    end while
31:  end for
32: end

```

maximal biclusters.

ii. (Algorithm 6, Line no. 3-7) From the *clusterMap* (as shown in Table 12.6), for any pair <cluster of items: supporting TIDs>, it is first checked whether the number of items in the cluster satisfies the threshold (*minItem*), otherwise the pair is removed.

iii. (Algorithm 6, Line no. 8-13) Corresponding to any frequent 1 item, if an entry in the *clusterMap* contains more than one initial biclusters, then, taking any two at a time, the intersection between the items and union between the corresponding TIDs are performed. This will derive a new set of biclusters. For example, let us consider two initial biclusters: <*MPQR* : *t6* > && <*MNOPR* : *t3, t7, t10* >. The newly derived bicluster could be, < $MPQR \cap MNOPR$: $\{t6\} \cup \{t3, t7, t10\}$ >, i.e., <*MPR* : *t6, t3, t7, t10* >

iv. (Algorithm 6, Line no. 14-20) Finally, any cluster whose number of TIDs is less than *minTID* and the number of items is less than *minItem* (Line no. 15) will be removed.

v. The final set of maximal biclusters are found by removing redundancy (Line no. 17), and stored in *biclusterMap* (Table 12.6 is showing the final biclusters along with the initial biclusters).

Algorithm 6 CellBiClust: Derive Final biclusters

Input *clusterMap* as obtained from Algorithm 5

Output *biclusterMap* : a map for storing the final biclusters

```

1: begin
2:   initialize biclusterMap                                ▷ A map for storing final biclusters
3:   for each entry (k,v) in clusterMap do                ▷ removing infrequent entries from
   clusterMap
4:     if ( $|k.size| < minItem$ ) then
5:       remove (k,v)
6:     end if
7:   end for
8:   for all ((ki,vi) & (kj,vj) ∈ clusterMap) do          ▷ derive new clusters
9:      $k \leftarrow (ki \cap kj) \ \& \ v \leftarrow (vi \cup vj)$ 
10:    if ( $|k.size| > minItem$ ) then
11:      clusterMap.add(k,v)
12:    end if
13:  end for
14:  for all (k, v) pairs in clusterMap do
15:    if ( $k \geq minItem$ ) & ( $v \geq minTID$ ) then
16:      if  $(k,v) \not\subseteq (k',v')$  in biclusterMap then
17:        biclusterMap.add(k, v)
18:      end if
19:    end if
20:  end for
21: end

```

Table 12.6: Initial and final biclusters corresponding to the Figure 12.2

Frequent items	1- clusterMap (initial bicluster)	biclusterMap (final biclusters)
M	-	-
N	$\langle PN : t4 \rangle$	-
O	$\langle MO : t1, t3, t5, t7 \rangle$	$\langle MO : t1, t3, t5, t7 \rangle$
P	$\langle MOP : t5 \rangle$	-
Q	$\langle MOQ : t3, t7 \rangle; \langle NQ : t2 \rangle$	$\langle MOQ : t3, t7 \rangle$
R	$\langle NQR : t2 \rangle; \langle PR : t6 \rangle$	-

12.2.2 Analysis of Algorithmic Result:

In this section, the procedures for extracting novel occurrences from the obtained biclusters are discussed [92]. Novel occurrence can be found in the biclusters itself and from the rules obtained from the biclusters. We explore the analysis of species presence/ absence dataset in Section 12.4.

Occurrence prediction from biclusters:

The proposed algorithm finds the maximal biclusters. The occurrence can be predicted from those:

Occurrence prediction

Say, $I \times T$ represents a binary input transaction dataset where rows and columns correspond to the items and transactions, respectively. Let's assume, we obtain two biclusters, $BC1 \Leftarrow \langle T1, T2, T3, T4, T5 : I1, I2, I3 \rangle$, and $BC2 \Leftarrow \langle T1, T2, T3, T6, T7 : I1, I4, I5 \rangle$. The following steps could predict novel occurrences from these two biclusters.

1. Similarity between any two biclusters is computed by intersection operation. If the similarity is greater than the specified minimum similarity value (60% is taken here), then we attempt to predict new occurrences from remaining unmatched members. Here, $BC1 \cap BC2 \Rightarrow \langle T1, T2, T3 \rangle$, i.e. 60% similarity exhibits.
2. Next, the set difference operation is performed and new predictions would be alike, $\langle \text{L.H.S of } BC1 - \text{L.H.S of } BC2 \rangle : \langle \text{R.H.S of } BC2 - \text{R.H.S of } BC1 \rangle$ and vice versa i.e.

$$| \langle T1, T2, T3, T4, T5 \rangle - \langle T1, T2, T3, T6, T7 \rangle | : | \langle I1, I4, I5 \rangle - \langle I1, I2, I3 \rangle | \text{ i.e. } \mathbf{T4 \rightarrow I4, I5; T5 \rightarrow I4, I5;}$$

$$| \langle T1, T2, T3, T6, T7 \rangle - \langle T1, T2, T3, T4, T5 \rangle | : | \langle I1, I2, I3 \rangle - \langle I1, I4, I5 \rangle | \text{ i.e. } \mathbf{T6 \rightarrow I2, I3; T7 \rightarrow I2, I3;}$$

Occurrence prediction from association rules:

The proposed algorithm finds the maximal biclusters and we process those to generate association rules which in turn guide new occurrence prediction:

Occurrence rule generation

As mentioned in the previous section 12.2.2, say we obtain a bicluster BC1 from $I \times T$ and $BC1 \Leftarrow \langle T1, T2, T3, T4, T5 : I1, I2, I3 \rangle$.

1. Now for generating rules, we go after placing one item (from a clustered Itemset) at a time in the consequent part while keeping all the remaining items as antecedent. Thus, for all individual items in the Itemset, a rule will be generated. This drives us to generate non-redundant rules as our aim is to predict new interactions for individuals.

Table 12.7: Rule generation for bicluster $\langle T1, T2, T3, T4, T5 \rangle : \langle I1, I2, I3 \rangle$

Rule	Antecedent	Consequent	Support	Confidence	Object lists
1	[T1, T2, T3, T4]	$\rightarrow T5$	3	-	[I1, I2, I3]
2	[T1, T2, T3, T5]	$\rightarrow T4$	3	-	[I1, I2, I3]
3	[T1, T2, T4, T5]	$\rightarrow T3$	3	-	[I1, I2, I3]
4	[T1, T3, T4, T5]	$\rightarrow T2$	3	-	[I1, I2, I3]
5	[T2, T3, T4, T5]	$\rightarrow T1$	3	-	[I1, I2, I3]

Here, the confidence for a rule is computed by the ratio between the support of the rule and the support of the antecedent. Say, for the antecedent of rule 1, the supporting objects are $\langle I1, I2, I3, I4 \rangle$. Then, Confidence of the rule 1 will be $count\{I1, I2, I3\} / count\{I1, I2, I3, I4\} = 3/4$, i.e., 0.75. In this way, all the rules are found out.

2. These rules are filtered out based on the support and the confidence value preferred by the users.
3. After the first filtration, the rules are checked to remove redundancy. Let, a rule $R1$ exists in the form of $Antecedent_{R1} \rightarrow Consequent_{R1}$. Rule $R2$ is more general than $R1$, i.e., $R1 \preceq R2$, if $Antecedent_{R1} \sqsubseteq Antecedent_{R2}$ and the consequent is the same for both $R1$ and $R2$. Hence, $R1$ can be removed. In this particular approach, the consequent is containing one element for all the rules. So, only the above-mentioned checking would be able to find all the redundancy exists.

Occurrence rule prediction

Consider the bicluster: $\langle T1, T2, T3, T4, T5 : I1, I2, I3 \rangle$; and the rule obtained from the closed set of the bicluster: $\langle T1, T2, T3, T4 \rightarrow T5 \rangle$ with the supporting objects for the antecedent: I1, I2, I3, I4. From the bicluster and the rule, the prediction could be, Consequent of rule \rightarrow set difference between $|ObjectListsOfAntecedentOfRule - ObjectListOfBicluster|$, i.e. $[T5] \rightarrow [I4]$

12.3 Performance evaluation

The performance test is conducted to check the characteristics of the proposed approach in terms of its acceptability. We measure the behavior of CellBiClust with reference to Bimax [267] algorithm as it is a remarkably accepted reference for the biclustering algorithm in research community. Likewise Bimax, We also deal with binary input transaction dataset to extract maximal biclusters. Both are compared at similar Java Programming Environment, on a system with 8 GB memory and Intel i7-6500U @2.5GHZ CPU. The operating system is Windows 10 Home 64-bit.

We purposefully select the test datasets (Table 12.8) from multiple domains (e.g. bioinformatics, market-basket, biodiversity) where the datasets can be represented in binary form and the algorithm could be employed in deriving meaningful facts. Additionally, using our test datasets, the evaluation could be done on both rectangular datasets and square datasets. For rectangular datasets, we have taken the real datasets meeting 3 variations:

- where the number of rows and the number of columns are closer (Grocery Market dataset with lesser number of rows than columns, and Sundarban Mangrove dataset with greater number of rows than columns),
- where the number of rows and columns have a larger difference (Estuarine Fish Presence dataset), and
- where the number of rows and columns have a much larger difference (HIV-Human PPI dataset).

We consider the transposed forms also for observing the nature of the testing algorithms in varying workspaces. In addition to these, we also consider a synthetic square dataset of 100 by 100 with varying densities of 10%, 20%, and 50% that strengthen the performance evaluation in a wider aspect.

Parameters for performance measure The testing is performed by varying the minimum number of rows and the minimum number of columns. These two parameters are taken as input for reshaping the working space. Thus, based on these parameters, the execution time, memory usage, and the number of biclusters increases/ decreases. Hence, the comparison of execution time and memory consumption between these two algorithms could be considered as valid testing parameters. Since both the algorithms find the maximal number of biclusters, both are generating similar number of biclusters (Table 12.9). Huge number of biclusters are generated for the Estuarine Fish Presence dataset due to its highly sparse nature. We have also experimented by varying the dataset density with fixed number of rows and columns. Density is measured by the percentage of 1's presence in the dataset. We require this additional experimentation for explaining the superiority of our

Data source and characteristics of the test datasets

Dataset name	Dataset source	#Row	#Column	Data Density
Synthetic dataset	Generated by the authors	100	100	10%, 20%, 50%
HIV-Human PPI dataset	Article [92]	1432	19	~ 10%
Transposed PPI dataset	HIV-Human Article [92]	19	1432	~ 10%
Estuarine Fish Presence dataset	Book [110]	762	20	~ 5%
Transposed Estuarine Fish Presence dataset	Book [110]	20	762	~ 5%
Grocery Market dataset	Kaggle groceries data	50	63	~ 6%
Sundarban Mangrove dataset	Article [110]	32	15	~ 37%

Table 12.9: Number of biclusters generated for both CellBiClust and Bimax

Dataset name	Minimum support			
	2	3	4	5
HIV-Human PPI dataset	295	192	94	31
Transposed HIV-Human PPI dataset	86	48	9	1
Sundarban Mangrove dataset	81	65	50	36
Grocery Market dataset	50	4	1	1
Estuarine Fish Presence dataset	3277	3030	2502	1822
Transposed Estuarine Fish Presence dataset	3277	3030	2500	1821

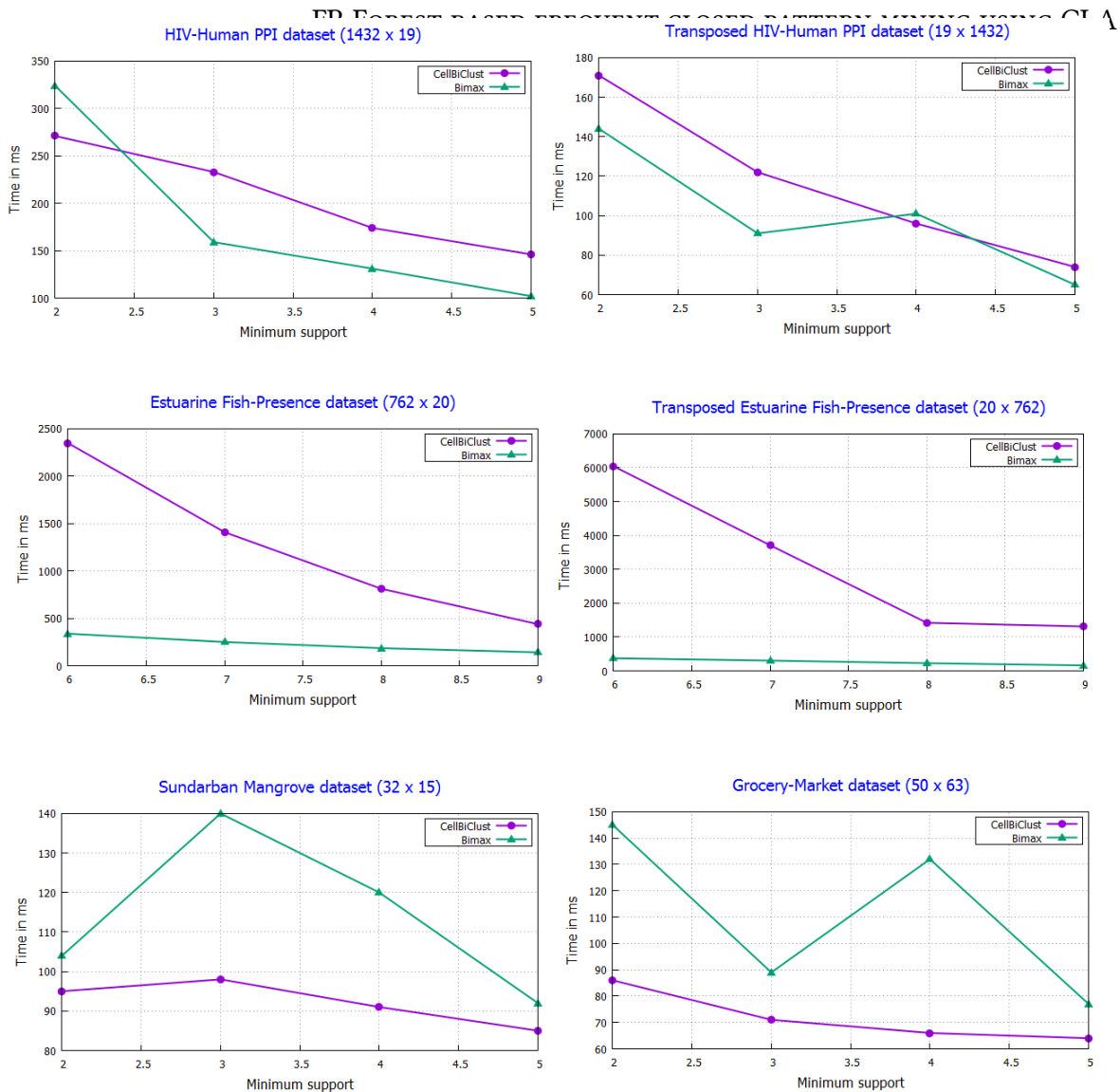


Figure 12.3: Comparison for time requirement

approach that cannot be explained by time and memory only.

Evaluation with respect to time Based on the distribution of 1's in different datasets, time is regulated. The time requirement has been shown in milliseconds in the graph (Figure 12.3). For smaller datasets (Sundarban Mangrove dataset, Grocery Market dataset) CellBiClust requires less time. But, for a larger number of rows/ columns (Estuarine Fish-Presence dataset and its transposed dataset), except for the minimum support with a very low value, our approach has the closer time requirement as the Bimax. Again, for a much larger dataset (HIV-Human PPI dataset and its transposed), the performance of both algorithms is almost similar and it can be explained through dataset density (discussed below in paragraph **Evaluation with respect to the dataset density**). So, it could be stated that we achieve an approximate time requirement like BiMax algorithm. However, this can be a trade-off, as memory consumption is reduced greatly in our approach.

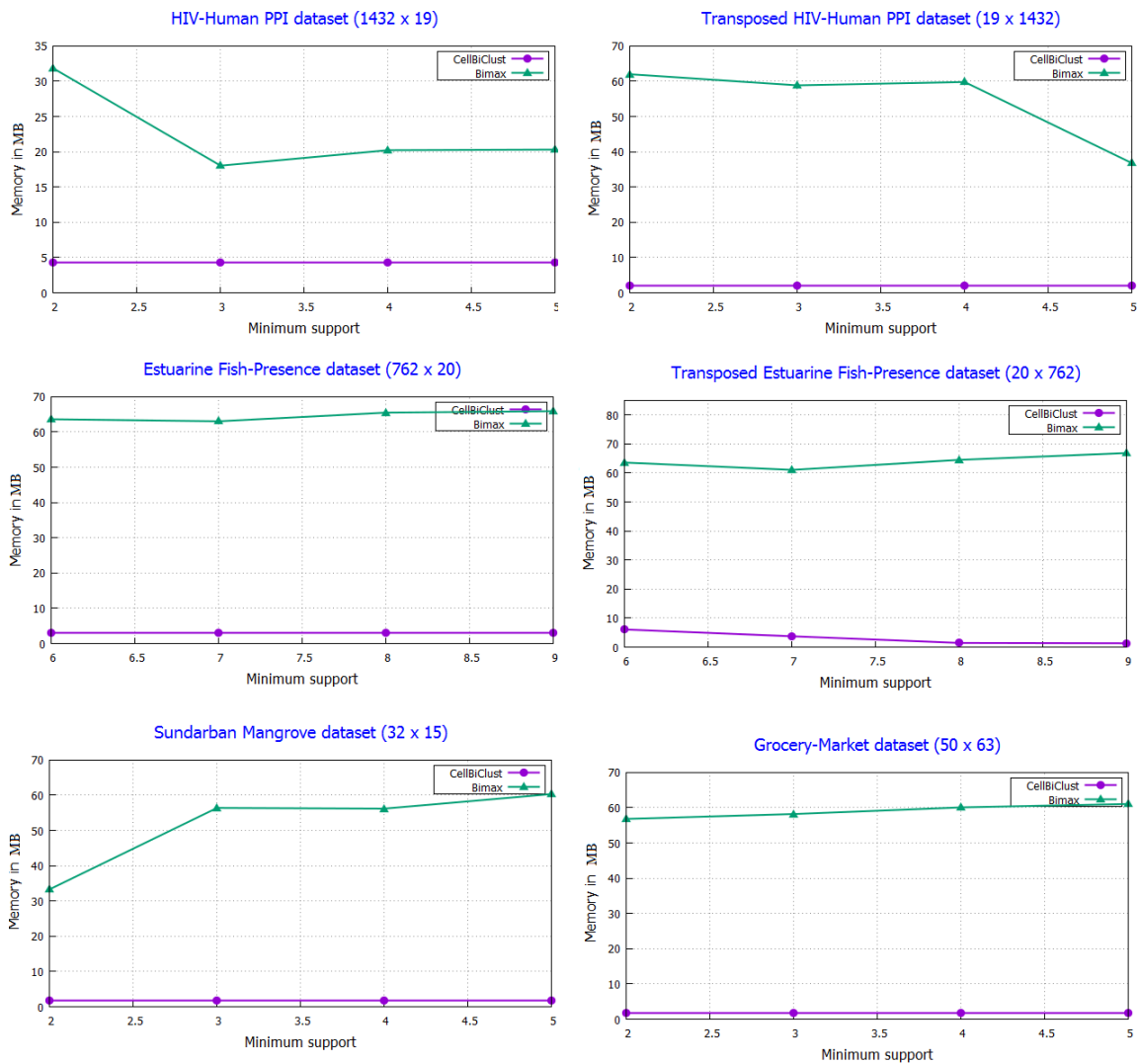


Figure 12.4: Comparison for memory usage

Evaluation with respect to memory Though execution time is comparable for both the algorithms, memory consumption is huge in the case of Bimax compared to CellBiClust (Figure 12.4). Here, memory is plotted in megabyte units. Bimax divides the whole matrix into multiple subproblems. Also, the columns having 1 value for a certain number of rows (depending on the dataset size) are forming a template for the subproblem formation. This leads to an unfair matrix splitting, particularly for sparse datasets. But, in the case of our approach, it reduces memory consumption at a higher rate. The reason is, that the ability of the tree data structure is efficient in the compact representation of the dataset. For all the datasets of varying sizes, CellBiClust consumes much less memory.

Evaluation with respect to the dataset density Performance is compared with an increasing density of 1s (Figure 12.5). Although evaluation with respect to memory shows the distinction of CellBiClust, evaluation with respect to time can be well-explained along with the dataset density. For smaller datasets, CellBiClust performs always better with

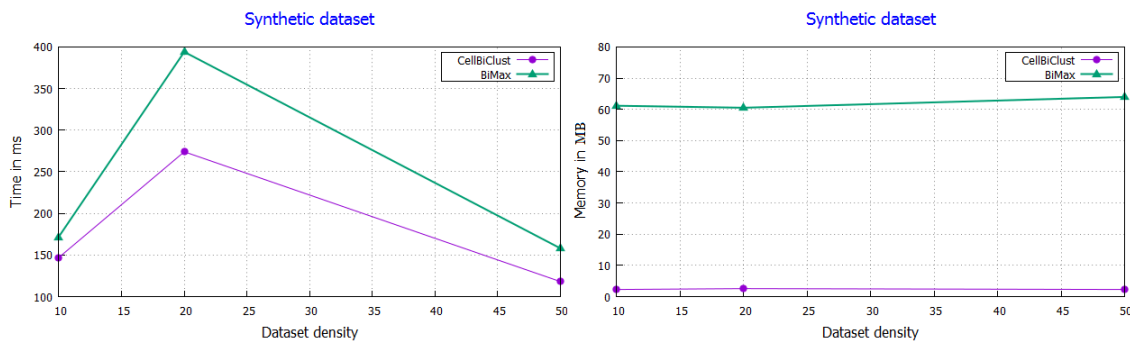


Figure 12.5: Time and memory comparison with varying data density

varying density, like Sundarban Mangrove dataset, Grocery Market dataset. But, for the larger datasets (Estuarine Fish Presence dataset) with very low density, for lower minimum support values, BiMax takes less time. Although, for larger minimum support values, the performance of both is comparable. If we consider an even larger dataset with comparatively larger density (HIV-Human PPI dataset), the time requirement of CellBiClust is closer to BiMax. So, it can be concluded that CellBiClust performs better or comparable with respect to BiMax in most of the scenarios with rectangular datasets. The performance of CellBiClust improves with dataset density and dataset size. If we consider a square dataset of moderate size, the performance of CellBiClust is always better with varying density for both the time requirement and memory usage (Figure 12.5).

12.4 Application on ecological forecasting

Although biodiversity and ecosystem informatics (BDEI) [275] is an interdisciplinary field that mitigates the research gap among computer scientists, biologists, conservationists, and others related to various aspects of biological diversity, literature needs more initiatives to extend the support of computer science in the context of ecological improvements. In this aspect, an efficient algorithmic solution could be made for predicting new associations among the species using frequent closed itemsets mining. Frequent closed patterns can identify association rules that in turn can further be used in ecological rule generation and prediction.

The species occurrence dataset could be mapped to a binary transaction dataset where the species are across the rows, and the observation sites are across the columns. The frequent co-occurrence pattern of species has become a focused issue in ecology for giving insight into the species distribution pattern. It needs substantial domain knowledge to identify those patterns. With the recent advancements of the data-science-based approach, examining frequently occurring species patterns could be trouble-free. We employ our proposed algorithm on a binary dataset of species occurrence. The analysis of the outcome (following the illustration as given in section 12.2.2) is justified by the domain expert. The intuitive knowledge and understanding of the domain expertise in the ecological domain play an essential role in guiding us in our addressed data mining application in ecological forecasting.

In this section, a detailed discussion is made on the knowledge extraction that could be helpful not only for the focused domain but also could establish the direction for other

Table 12.10: Statistics for the analytical results

1. Data set	2. Max-3 clusters	3. Unique Bi-prediction from clusters (Similarity threshold 85%)	4. Unique occurrence from bi-clusters (minimum confidence)	5. Generated rules (70% rules)	6. Unique occurrence from the prediction and the 3rd and 5th columns	7. Common prediction in count	8. Unique prediction combined unique prediction
FxE	1973	409	906	1004	259	1154	1277
ExF	2147	292	795	236	121	407	

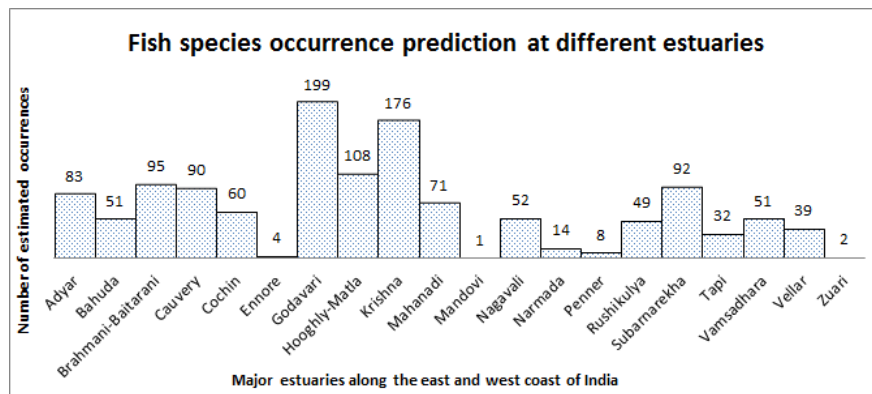


Figure 12.6: Representing histogram for fish occurrence prediction at 20 estuaries

research areas.

Preparation of dataset Although the proposed algorithm is applicable to any dataset in binary form, we have presented a biodiversity domain-specific usage of it. We use a curated presence-absence binary dataset of fish of Indian coastal regions [110]. 20 major estuaries are identified from the long coastal area of India and a total of 762 fish species occurrence data is portrayed. From the east coast of India, 15 estuaries are taken. These are Hooghly-Matla, Subarnarekha, Baitarani-Brahmani, Mahanadi, Rushikulya, Bahuda, Vamsadhara, Nagavali, Godavari, Krishna, Penner, Ennore, Adyar, Vellar, and Cauveri. Cochin, Zuari, Mandovi, Tapi, and Narmada are considered from the west coast of India. The input dataset is provided in the supplementary material.

Results on F×E dataset F×E denotes the dataset where rows are for the fish data and the columns are for the estuaries. Being a sparse dataset, the minimum support for bicluster generation is kept low. Experimentally we set it at 10. As a large number of biclusters are produced here, a larger similarity threshold has been fixed. It is found when the similarity percentage is 85, 409 unique occurrences could be predicted for individual species at individual estuaries. From the biclusters, association rules are also formed where we keep the minimum confidence level at 70% for filtering in the more significant ones. At this stage,

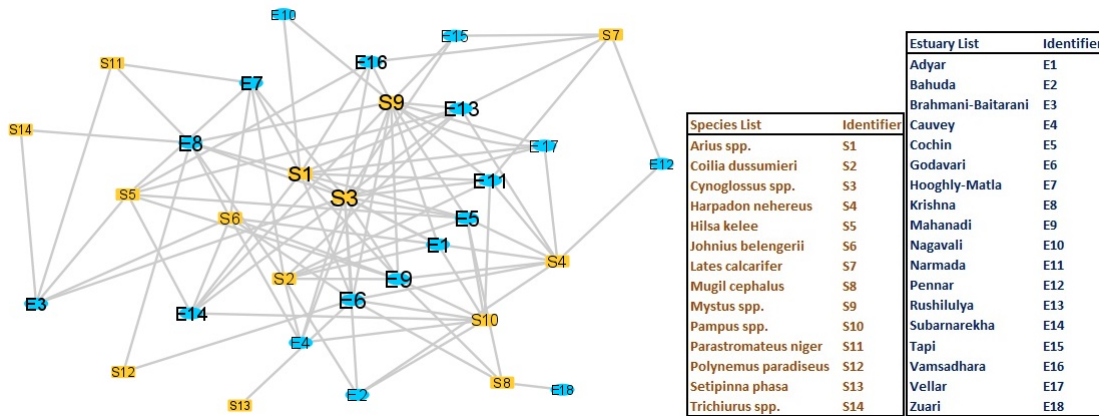


Figure 12.7: Network for predicted occurrence

1004 new occurrences have been predicted by the rules. Now, it has been observed that a significant part of the predicted occurrences (259 predictions in this case) are common for both of the cases. But in order to obtain a finer and complete prediction, both of the procedures should be followed. To this end, we would illustrate it considering the predicted species occurrence at Hooghly-Matla estuary. To examine more precisely, the similarity percentage of the biclusters is set to a higher value (85%) to get a lesser number of possible predictions as a large number of predictions are generated that may not be feasible to follow. On the other hand, the confidence of a rule is set at a minimum threshold of 70%, below which a rule would not be considered to have a strong impact.

With these specifications, 18 non-native fish to Hooghly-Matla estuary are listed from the predictions obtained from biclusters, whereas the predictions from the rules generate a list of 98 non-native species. Noticeably, *Synaptura commersonnii* is obtained only from the predictions from the biclusters. Though *Synaptura commersonnii* is not reported from Hooghly-Matla estuary in the input dataset, it has been found at Subarnarekha estuary and Brahmani-Baitarani estuary- that are adjacent to Hooghly-Matla estuary. Also, Mahanadi, Godavari, and Krishna - 3 other estuaries from the east coast of India, have recorded their presence. Besides, *Synaptura commersonnii* is also found at Cochin and Zuari - two estuaries from the west coast. Hence, its occurrence at Hooghly-Matla could be expected as it is one of the most diverse, species-rich estuaries. Therefore, it can be stated that both of the procedures for predictions, i.e., from the biclusters as well as from the rules, have importance in finding probable novel occurrence data.

Results on E × F dataset E × F is the transposed dataset of F × E. Here, the number of rows is much smaller representing the estuaries. Thus, experimentally we set minimum support to 5. As before, here also, we set the similarity threshold at 85% to predict the occurrences of the biclusters. In this case, 292 unique occurrences are predicted. Again, 236 number of occurrences have been predicted by the rules.

Concerning the Hooghly-Matla estuary, 21 and 10 unique occurrences are predicted from the biclusters themselves and from the rules, respectively. However, 5 are found to have common resulting in 26 unique novel occurrence predictions. Evidently, the prediction from the biclusters listed larger species than the prediction from the rules (opposite scenario

to the case of the F×E dataset), emphasizing the necessity of both of the procedures. The detailed information, as mentioned in Table 12.10, justifies the need for the transposed form of the dataset as it also contributes to the novel predictions.

The final merged predicted list obtained from both E×F and F×E From the two predicted sets of occurrences, 284 predictions are detected to have a common occurrence in both cases. The remaining occurrences are merged, generating 1277 newly predicted unique occurrences. Figure 12.6 presents a histogram for the fish species occurrence prediction made for individual estuaries. It can be observed that the estuary of Godavari has met the highest number of possibilities of occurrence of apparently not-occurring species. Following this, Krishna estuary and Hooghly-Matla estuary are found to have numerous non-native species occurrence probability. This kind of analysis would aid in diversity enhancement in terms of species richness in any ecological community of an estuary. It has been found that out of 762 species that we have considered, 382 species are highlighted in the predicted list based on the analysis followed.

Applicability of the predicted fish occurrence data-set on commercial fishes A study [276] has shown that almost 90% of the aquatic species are nourished at Hooghly-Matla estuary in Sundarban. Being the main hub for the coastal fishery in eastern India, fishing is one of the important means for securing the local peoples' lives. A list of commercial fishes from Hooghly-Matla estuary is given in [54]. We relate our predicted list to these commercial fishes and obtain Figure 12.7 showing the other probable estuaries for these commercial fishes to be harvested. S3, S1, and S9 are found to have wider possibilities and are represented in a larger font and a higher number of interactions with estuaries. Similarly, estuaries E9, E6, etc. have substantial prospects. It can be believed that this kind of knowledge derivation would aid in surveying for species restoration and re-sampling.

12.5 Summary

Mining biclusters is an important data mining task that has been addressed here. Exploiting the concept of cellular learning automata embedded in an FP Forest-like data structure makes it more efficient. Furthermore, predicting novel incidences or occurrences from the binary dataset is included to extend the use of biclustering. Experiments reveal that the proposed algorithm is efficient in terms of memory usage and execution time. The execution time can be reduced further by using parallelism in the cluster-finding step. Our implementation could work as a "black box" for forecasting new associations for any other binary dataset. Without a detailed understanding of the algorithmic implementation, we could easily generate the algorithmic outcome in terms of biclusters and the extended usage of the algorithmic result in terms of association rules, predicting lists with probabilities of association. Hence, this could be of interest to researchers from other domains, like biodiversity, bioinformatics, market basket analysis, or any other field with binary data, as they can make use of our approach with efficiency following the analysis we have shown.

CHAPTER **13**

FCA-BASED CONSTANT AND COHERENT-SIGNED BICLUSTER IDENTIFICATION

13.1	Introduction	192
13.1.1	Background and motivation	192
13.1.2	Related work	192
13.1.3	Contribution	193
13.2	Constant and coherent signed biclusters	193
13.3	Pattern structure for the signed partition & bicluster generation .	197
13.4	Summary	197

13.1 Introduction

The data mining task of finding coherent signed bicluster is not new in the field of gene expression data. It could also be applied in the area of computation-oriented biodiversity study with a significant impact on exposing domain-specific coherency. The present study considers a symbolic table filled with signs having meaning imposed by the users and proposes a novel signed biclustering methodology using formal concept analysis. The present work has the ability to identify both the constant and coherent signed biclusters. Moreover, aiming to reveal the usefulness of the proposed approach, we prepare a signed dataset corresponding to the spatio-temporal changes of abundance data of Sundarban mangroves, the vulnerable mangrove ecosystem. In this work, we explain our methodology theoretically with the help of a related but smaller synthetic dataset.

13.1.1 Background and motivation

The mangrove forest is identified as one of the most threatened ecosystems in the world [277]. The major reasons lying behind the world-wide mangrove loss are anthropogenic activities and the rapid rate of climate change [278]. Furthermore, along with the declining ecosystem, coastal livelihood is also substantially affected. This situation is turned into a crucial challenge where the identification of spatial and temporal changes of mangrove cover is deadly needed [279, 280] as it can lead towards the planning for long-term conservation of the mangrove forest [281, 61].

From the designing point of view, biclustering [36] is related to standard clustering in a matrix along both the row and column dimensions. To form a bicluster, a subset of objects along the rows get clustered down based upon a subset of attributes along with the columns. Whereas, formal concept analysis (FCA) [282] in mathematical theory, is becoming popular [283, 284], derives a concept hierarchy based upon the objects and their properties. Here the objects form a group by their common attributes and create a formal concept. Thus, a formal concept corresponds to a bicluster. In this aspect, it can be stated that FCA is homogeneous to biclustering as it brings out all the maximal rectangles from a binary matrix and arranges them in a hierarchical concept lattice [285].

Here we have proposed a signed biclustering algorithm to derive a specific type of bicluster based on the constant symbol and coherent symbolic changing data. Related to our addressed problem, this kind of bicluster would efficiently extract all the clustered regions having constant or coherent changes in mangrove cover considering all the aspects of the biodiversity.

13.1.2 Related work

The term biclustering is familiar to the researchers since the former studies in the field of gene expression data [286, 287, 92, 37]. Conceptually, in the gene expression dataset, any bicluster seems to have a subset of genes expressing similar behavior under a subset of conditions [288]. One important variation of such bicluster is identified in [288] and named coherent-sign-changes bicluster [36] where the gene expression values are either increasing or decreasing based upon the specific conditions under the submatrix forming the biclusters. Coming to the FCA-based bicluster formation, [289, 290], are the significant

research works where biclusters are discovered from the dense binary matrix. Instead of exact biclusters, here, it has been shown that a way to form approximate biclusters considering some empty cells as well. Bicluster formation from the numerical matrix is addressed in [291]. Triadic concept analysis, an extension of the formal concept analysis, is studied here. In addition to this, formal concept analysis and pattern structure-based biclustering methodology are studied in [292] where the symbolic matrix is used. Another variation, order-preserving bicluster is discussed in [285].

Regarding the study on the mangrove ecosystem, alpha diversity (diversity in a small area), beta diversity (comparison of diversity between two areas), and gamma diversity (diversity in a landscape) of Sundarban mangrove, are studied in [293]. These kinds of studies would be helpful in strategic planning for the conversion [294] as spatial and temporal changes in species maps can identify the regions that need proper protection policies. As per [295], Bangladesh Sundarban is divided into three ecological zones, hypo-saline, meso-saline, and hyper-saline (from lower to higher salinity). But the Farakka barrage in India (1975), causes a great reduction in freshwater supply [11]. Therefore, a major transformation in the salinity level is found in different zones [281]. Along with the zonal transformation due to the salinity, vegetation pattern is also changing. [293] reveals an explicit representation of the changing pattern of the geographic range and mangrove species abundance in the interval of 1980 to 2014. Our approach for extracting signed biclusters is appropriate to study this kind of dataset for deriving knowledge in conservation policy. Using the symbolic table, it is possible to cluster down whether a species is increasing in count or decreasing, whether a species is newly appeared at a site or completely disappeared, etc.

13.1.3 Contribution

The contribution made in this study is listed below:

- Here, our intention is to show the use of signed bicluster in analyzing data in the domain of biodiversity.
- For this, we have curated a dataset on the mangrove cover changes over the years for demonstrating the employment of signed bicluster.
- We have provided a formalization based upon the FCA and partition-based pattern structure by taking into account the direction of the symbolic changes and no changes as well.
- The addressed constant and coherent sign changing bicluster is novel as we are not restricted to binary symbols. We consider the change in symbolic direction instead of the magnitude of the attribute value. This kind of multi-symbolic sign-changing bicluster identification and its domain-specific study is new in the literature.
- We present a theoretical illustration of the hierarchical structure of all the biclusters and frame an interpretation of the derived clusters from the viewpoint of an expert.

13.2 Constant and coherent signed biclusters

This section gives examples of the constant and coherent signed biclusters with respect to our addressed biodiversity domain:

CHAPTER 13

Table 13.1: Symbolic representation for the changes in Sundarban Mangrove Species count data between 1986 to 2014 [293]

Sl No.	Species	Hyposaline	Mesosaline	Hypersaline
1	<i>Excoecaria agallocha</i>	~	+	~
2	<i>Heritiera fomes</i>	~	-	~
3	<i>Avicennia officinalis</i>	-	-1	-
4	<i>Sonneratia apetala</i>	-	0	-
5	<i>Amoora cucullata</i>	~	-	-
6	<i>Bruguiera sexangula</i>	+	~	~
7	<i>Xylocarpus moluccensis</i>	+	~	-
8	<i>Cynometra ramiflora</i>	-	-	0
9	<i>Cerbera manghas</i>	-	0	0
10	<i>Talipariti tiliaceum</i>	-	0	0
11	<i>Aegiceras corniculatum</i>	-	0	0
12	<i>Excoecaria indica</i>	-	0	0
13	<i>Tamarix dioica</i>	-	0	0
14	<i>Barringtonia racemosa</i>	-1	0	0
15	<i>Ceriops decandra</i>	1	1	+
16	<i>Sonneratia caseolaris</i>	-1	0	0
17	<i>Intsia bijuga</i>	+	0	0
18	<i>Lansea coromandelica</i>	-1	0	0
19	<i>Xylocarpus granatum</i>	+	-	-
20	<i>Pongamia pinnata</i>	+	0	0
21	<i>Syzygium fruticosum</i>	+	0	0
22	<i>Hypobathrum racemosum</i>	1	1	0
23	<i>Salacia chinensis</i>	0	-1	0
24	<i>Rhizophora mucronata</i>	0	1	0
25	<i>Lumnitzera racemosa</i>	0	-	0

Table 13.1 is derived from the data presented in a work [293] where Sundarban mangrove abundance data for 25 species are highlighted in hyposaline, mesosaline, and hypersaline zones. We represent the data symbolically to express the changes in species count data that is given in between 1986 to 2014. Each symbol is conveying a specific meaning. We have taken five such symbols, viz; 0, -, ~, +, and 1; where,

0: Historical absence (A species is absent in both 1986 and 2014)

-: Range contraction (Decrease in species count)

~: Unchanged (Same abundance data in both 1984 and 2014)

+: Range expansion (Increase in species count)

1: Introduced (Absent in 1986 but present in 2014)

-1: Disappeared (Present in 1986 but absent in 2014)

Our approach for finding out both the constant and coherent signed biclusters would be appropriate for this kind of dataset. As we are elaborating the theoretical framework here, for the convenience of the illustration, and to cover all the possible aspects, we would consider a smaller synthetic symbolic dataset.

Table 13.2: Synthetic example dataset for illustration

	n1	n2	n3	n4	n5
m1	+	-	+	-	-
m2	+	-	+	-	~
m3	+	0	+	0	~
m4	+	0	+	0	~
m5	+	1	+	1	~

Our example dataset $M \times N$ is given in Table 13.2. Say, M represents a list of species and N represents a list of zones. Each cell is representing the changes in total species count between an interval in years. As specified before, for Table 13.2, we have considered all the five symbols: 0, -, ~, +, and 1; representing their respective meanings.

All the left-hand-side symbols including \sim , i.e. 0, -, and \sim are forming the negative domain whereas all the right-hand-side symbols including \sim , i.e. \sim , +, 1 are forming the positive domain. The types of the biclusters are illustrated with examples below:

1. In Table 13.3, a positive constant bicluster is indicating an increasing species count (for species m1, m2). Or, in other words, the diversity increases in terms of species count in the regions numbered n1, n2, and n3.

Table 13.3: Constant bicluster

Condition 1: Constant bicluster i.e. bicluster with the same sign			
m1	$n1^+$	$n2^+$	$n3^+$
m2	$n1^+$	$n2^+$	$n3^+$

2. Table 13.4 is showing a few scenarios for condition 2. It extracts biclusters where a particular species diversity behaves homogeneously along with varying sites. At the same time, it also extracts a list of species that behave oppositely with respect to a particular site. The inversely signed elements for a particular site can be identified through this kind of biclusters. Thus the overall scenario for a site can be noted.

Table 13.4: Example for row constant and inverse column coherent biclusters

Condition2: Row constant + inverse column coherent values, i.e. row-wise same valued symbol and column-wise symbol from the opposite domains except considering \sim													
m1	$n1^-$	$n2^-$	$n3^-$	or	m1	$n1^1$	$n2^1$	$n3^1$	or	m1	$n1^+$	$n2^+$	$n3^+$
m2	$n1^+$	$n2^+$	$n3^+$		m2	$n1^0$	$n2^0$	$n3^0$		m2	$n1^0$	$n2^0$	$n3^0$

3. Table 13.5 inferences that the sites (n1, n2, n3) are exhibiting diminishing biodiversity for the species m1 and m2. More specifically, m2 has been disappearing, and m1 is shrinking (denoted by the left-hand-side bicluster).

Table 13.5: Example for row constant and negative column coherent biclusters

Condition 3: Row constant + negative column coherent values i.e. row wise same valued symbol and column-wise symbol from the negative domain									
m1	$n1^-$	$n2^-$	$n3^-$	or	m1	$n1^-$	$n2^-$	$n3^-$	
m2	$n1^0$	$n2^0$	$n3^0$		m2	$n1^\sim$	$n2^\sim$	$n3^\sim$	

4. Table 13.6 is indicating a positive response to biodiversity. As per the example shown here, m1 is expanding in nature at the sites n1, n2, and n3, whereas, m2 is newly introduced (for the left-sided bicluster) or keeping its count consistent (for the right-sided bicluster).

Table 13.6: Example for row constant and positive column coherent biclusters

Condition 4: Row constant + positive column coherent values, i.e. row wise same valued symbol and column-wise symbol from the positive domain									
m1	$n1^+$	$n2^+$	$n3^+$	or	m1	$n1^+$	$n2^+$	$n3^+$	
m2	$n1^1$	$n2^1$	$n3^1$		m2	$n1^\sim$	$n2^\sim$	$n3^\sim$	

CHAPTER 13

Table 13.8: Example for column constant and negative row coherent bicluster Table 13.9: Example for column constant and positive row coherent bicluster

Condition 6: Column constant + negative row coherent values, i.e. column wise same valued symbol and all the symbols are in the negative domain.			
m1	$n1^-$	$n2^0$	$n3^\sim$
m2	$n1^-$	$n2^0$	$n3^\sim$

Condition 7: Column constant + positive row coherent values, i.e. column wise same valued symbol and all symbols are from positive domain			
m1	$n1^1$	$n2^+$	$n3^1$
m2	$n1^1$	$n2^+$	$n3^1$

Table 13.10: Example for bicluster with coherent evolution on positive domain Table 13.11: Example for bicluster with coherent evolution on negative domain

Condition 8: Coherent evolution in positive domain		
m1	$n1^+$	$n2^+$
m2	$n1^+$	$n2^1$

Condition 9: Coherent evolution on negative domain		
m1	$n1^\sim$	$n2^-$
m2	$n1^-$	$n2^0$

Table 13.7: Example for column constant and inverse row coherent biclusters

Condition 5: Column Constant + inverse row coherent values, i.e. column-wise same valued symbol and the row-wise symbols are from the opposite domain except considering \sim					
m1	$n1^-$	$n2^+$	or	m1	$n1^-$ $n2^1$
m2	$n1^-$	$n2^+$		m2	$n1^-$ $n2^1$
				m1	$n1^+$ $n2^0$
				m2	$n1^+$ $n2^0$

5. The kind of biclusters shown in Table 13.7 reveals the site-wise similar changes towards species-biodiversity. This may lead to the identification of vulnerable sites where multiple species(m1, m2) counts are dropping, or they are gradually disappearing. Similarly, the positive scenario for any site can also be identified.
6. Table 13.8 is showing a bicluster where site-wise, multiple species (m1, m2) are clustered down based upon the common pessimistic scenario. As per the table, both m1 and m2 are losing their biodiversity across multiple sites.
7. The example shown in Table 13.9 identifies the optimistic scenario related to biodiversity. It highlights the positive circumstances of specific sites for some specific species. As per Table 13.9, site n2 has a list of species with increasing species count, whereas sites n1 and n3 have a newly introduced species list.
8. Table 13.10 identifies a bicluster with coherent evolution on the positive domain. It tends towards an overall optimistic evolution.
9. Table 13.11 identifies a bicluster with coherent evolution on the negative domain. It highlights the overall alarmist scenario.

13.3 Pattern structure for the signed partition & bicluster generation

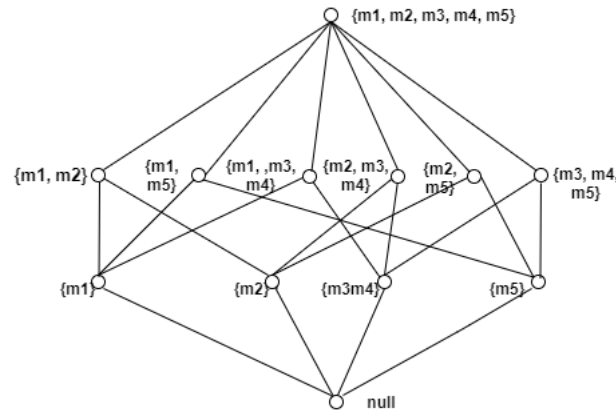


Figure 13.1: Signed partition pattern lattice for the pattern concept shown in Table 13.12: Only the Extents are shown within the diagram.

Table 13.12 has listed all the signed partition patterns and the corresponding biclusters that could be derived from Table 13.2. The extents for the concepts of Table 13.12 are hierarchically shown as a lattice in Figure 13.1. The corresponding intents can be found in Table 13.12

13.4 Summary

This work presents a novel approach for mining a multiple-signed dataset to identify both the constant and coherent signed biclusters. Bicluster has its importance in the recommendation system, along with the study related to commonly known bioinformatics. Here we have revealed another direction of using biclusters, i.e. in biodiversity study. We can conclude that this kind of signed bicluster retrieval from the spatio-temporal dataset of species vs region may help in identifying the vulnerable regions along with the unprotected or endangered species for biodiversity conservation. Not only that, it would help the conservationists for conserving or restoring the declining population of a species community, or even an ecosystem.

CHAPTER 13

Table 13.12: Signed partition concepts generated from Table 13.2 and the respective biclusters along with their types

Concept		Bicluster		Type
Extent	Intent	Objects	Attributes	
{m1}	$\{n1^+, n2^-, n3^+, n4^-, n5^-\}$	{m1}	$\{n1^+, n2^-, n3^+, n4^-, n5^-\}$	
{m2}	$\{n1^+, n2^-, n3^+, n4^-, n5^{\sim}\}$	{m2}	$\{n1^+, n2^-, n3^+, n4^-, n5^{\sim}\}$	
{m3m4}	$\{n1^+, n2^0, n3^+, n4^0, n5^{\sim}\}$	{m3m4}	$\{n1^+, n2^0, n3^+, n4^0, n5^{\sim}\}$	
{m5}	$\{n1^+, n2^1, n3^+, n4^1, n5^{\sim}\}$	{m5}	$\{n1^+, n2^1, n3^+, n4^1, n5^{\sim}\}$	
{m1, m2}	$\{\{n1^+, n2^-, n3^+, n4^-\}\{n5^{\sim} \sim\}\}$	{m1, m2}	$\{n1^+, n2^-, n3^+, n4^-\}$	Column constant & inverse row coherent;
{m1, m3, m4}	$\{\{n1^+, n3^+\}, \{n2^{-0}, n4^{-0}, n5^{\sim}\}\}$	{m1, m2}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
		{m1, m3, m4}	$\{n1^+, n3^+\}$	Constant;
{m1, m5}	$\{\{n1^+, n3^+\}, \{n2^{-1}, n4^{-1}\}, \{n5^{\sim} \sim\}\}$	{m1, m3, m4}	$\{n2^{-0}, n4^{-0}, n5^{\sim}\}$	Negative coherent evolution;
		{m1, m5}	$\{n1^+, n3^+\}$	Constant;
{m2, m3, m4}	$\{\{n1^+, n3^+, n5^{\sim}\}, \{n2^{-0}, n4^{-0}\}\}$	{m1, m5}	$\{n2^{-1}, n4^{-1}\}$	Row constant & inverse column coherent;
		{m1, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
{m2, m3, m4}	$\{\{n1^+, n3^+, n5^{\sim}\}, \{n2^{-0}, n4^{-0}\}\}$	{m2, m3, m4}	$\{n1^+, n3^+, n5^{\sim}\}$	Positive coherent evolution;
		{m2, m3, m4}	$\{n1^+, n3^+, n5^{\sim}\}$	Constant;
{m2, m5}	$\{\{n1^+, n3^+, n5^{\sim}\}, \{n2^{-1}, n4^{-1}\}\}$	{m2, m3, m4}	$\{n^{-0}, n4^{-0}\}$	Row constant & negative column coherent;
		{m2, m5}	$\{n1^+, n3^+, n5^{\sim}\}$	Positive coherent evolution;
{m3, m4, m5}	$\{\{n1^+, n3^+, n5^{\sim}\}, \{n2^{-1}, n4^{-1}\}\}$	{m2, m5}	$\{n2^{-1}, n4^{-1}\}$	Row constant & inverse column coherent evolution;
		{m2, m5}	$\{n2^{-1}, n4^{-1}\}$	Constant;
{m3, m4, m5}	$\{\{n1^+, n3^+, n5^{\sim}\}, \{n2^{01}, n4^{01}\}\}$	{m3, m4, m5}	$\{n1^+, n3^+, n5^{\sim}\}$	Positive coherent evolution;
		{m3, m4, m5}	$\{n2^{01}, n4^{01}\}$	Row constant & inverse column coherent;
{m1,m2, m3, m4}	$\{\{n1^+, n3^+\} \{n2^{-0}, n4^{-0}, n5^{\sim}\}\}$	{m1,m2, m3, m4}	$\{n1^+, n3^+\}$	Constant;
{m1, m2, m5}	$\{\{n1^+, n3^+\} \{n2^{-1}, n4^{-1}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4}	$\{n2^{-0}, n4^{-0}, n5^{\sim}\}$	Negative coherent evolution;
		{m1, m2, m5}	$\{n1^+, n3^+\}$	Constant;
{m1, m2, m5}	$\{\{n1^+, n3^+\} \{n2^{-1}, n4^{-1}\}, \{n5^{\sim} \sim\}\}$	{m1, m2, m5}	$\{n2^{-1}, n4^{-1}\}$	Row constant & inverse column coherent;
		{m1, m2, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
{m1,m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4, m5}	$\{n1^+, n3^+\}$	Constant;
		{m1,m2, m3, m4, m5}	$\{n2^{-01}, n4^{-01}\}$	Row constant & inverse column coherent;
{m1,m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
		{m1,m2, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Constant;
{m1, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4, m5}	$\{n1^+, n3^+\}$	Constant;
		{m1, m3, m4, m5}	$\{n2^{-01}, n4^{-01}\}$	Row constant & inverse column coherent;
{m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
		{m2, m3, m4, m5}	$\{n1^+, n3^+\}$	Constant;
{m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m2, m3, m4, m5}	$\{n2^{-01}, n4^{-01}\}$	Row constant & negative column coherent;
		{m2, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
{m1,m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4, m5}	$\{n1^+, n3^+\}$	Constant;
		{m1,m2, m3, m4, m5}	$\{n2^{-01}, n4^{-01}\}$	Row constant & inverse column coherent;
{m1,m2, m3, m4, m5}	$\{\{n1^+, n3^+\} \{n2^{-01}, n4^{-01}\}, \{n5^{\sim} \sim\}\}$	{m1,m2, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Negative coherent evolution;
		{m1,m2, m3, m4, m5}	$\{n5^{\sim} \sim\}$	Constant;

CHAPTER 14

A DATA-DRIVEN RULE FILTERING APPROACH
FOR FOREST RESTORATION

14.1	Introduction	200
14.2	Previous studies and scope of the data mining approach	202
14.3	Materials and methods	204
14.3.1	Dataset description	204
14.3.2	Proposed framework	205
14.4	Results	208
14.4.1	Analysis on frequency of occurrence	208
14.4.2	Analysis on objective and subjective measures	208
14.5	Discussion	219
14.5.1	Comparison to the earlier findings	219
14.5.2	Potential effects on conservationists	220
14.5.3	Comparison to concurrent research initiatives	221
14.5.4	Current Situation of Mangroves in India: Emerging Threats, Policy, and Future Suggestions	221
14.6	Summary	223

14.1 Introduction

Taxonomic distinctness Taxonomic distinctness [296, 297] is a biodiversity index that determines the average amount of distance between all species pairs in a sampled community, where this distance is measured as the length of the branch across a common phylogenetic [298] tree that connects these species. It considers differences and evolutionary relationships. Taxonomic distinctness is the theory of increasing species diversity by maximizing species differences. According to research [296, 299, 300], taxonomic markers of diversity and distinctness are sensitive indicators. A region's sustainable ecological condition is indicated by a higher taxonomic distinctness.

Role of taxonomic distinctness in the study of ecological restoration Taxonomy examines biological and ecological characteristics of species, going beyond the nomenclature and categorization of biological diversity [301, 300]. Therefore, it is essential for ecological restoration to include taxonomic and evolutionary linkages of the species as it can shed light on the potentiality of co-existence, cooperation, or competition. High taxonomic distinctness value in a region denotes that the species within a community differ greatly from one another in terms of their ecological roles, characteristics, and evolutionary histories. This implies that a large variety of ecological niches and adaptations are present within the community, increasing its overall diversity. Moreover, greater species divergence results in resource allocation or niche segregation (a natural selection procedure that excludes competition favoring cooperation for better survival) [302]. It could be a strategic action of co-existence.

Declining biodiversity of mangrove forest Mangroves act as ecological border security forces by preventing seawater infiltration and regulating a variety of climate catastrophes. Therefore the diminishing biodiversity of mangrove forests is a cause for serious concern [303, 304]. They also sustain a variety of terrestrial and marine life and offer essential services [305, 306]. However, both anthropogenic and natural forces pose threats to mangroves, which lowers their capacity for resilience and recovery from the adverse effects of environmental change [303, 307]. The loss of mangroves is mostly caused by the effects of climate change and rising sea levels, which overshadow other problems like excessive and illegal logging, prawn aquaculture, pollution, etc. [18]. In addition to offering ecological services, mangroves have significantly aided in the socioeconomic advancement of the area and the maintenance of local livelihoods [53, 52, 308]. Addressing the diminishing biodiversity of mangrove forests requires concerted efforts at all levels, including the international level. Implementing efficient management that might aid in the preservation and restoration of the biodiversity of mangrove forests should be part of conservation efforts. Thus, a data-driven study might be proposed in order to ensure the effective management of this priceless environment [309].

A Potential data-driven adaptive restoration strategy and the need for primary biodiversity data It is envisioned that closely related plants exhibit more frequent co-existence patterns than animals [49]. The homogeneity of characters in the process of evolution, such as adaptability, migration, resource uptake, and reproductivity, are the major influencers behind the co-existence pattern of different plant species for ecosystem re-

silience [50]. So, instead of using prevalent, widespread plant species for restoration purposes [51], exploration of co-occurrence data can suggest multiple other suitable species for ecosystem restoration. Hence, the use of association rule mining in finding frequent co-occurrence patterns can reduce the knowledge gap between data analyzers and restoration practitioners, and the advancement of knowledge sharing could assist domain researchers. However, substantial amounts of data are needed in the process of mining meaningful insights in problem-solving, trend prediction, and potential exploration. In the context of ecosystems, primary data on biodiversity is a fundamental requisite for strong decision-making in order to accomplish ecosystem conservation [225, 310, 280]. Therefore, the exploration of existing and prospective unpublished data and statistics needs to be strengthened to uncover the prevailing knowledge gap in biodiversity. Accordingly, it is understood that one of the potential approaches for confronting the growing crises in biodiversity is to obtain biodiversity-related data (species, genes, and ecosystems) and information in an appropriate form [225]. However, there are inconsistencies in the available data from multiple resources. Thus it becomes challenging to assemble when the integrated dataset is needed by the policymakers in formulating strategic action plans [227]. Moreover, mining the available data for formulating a restoration strategy, is another issue [280]. Therefore, the primary goal of this work would be to propose a novel data-driven framework for developing the ecological restoration strategy that makes use of domain expertise and compiles the data on taxonomy and occurrence of mangrove species to fulfill the research objective.

Research questions In this work, we try to follow the current research trend by highlighting the importance of data-driven analysis in biodiversity research. The study addressed biodiversity-related challenges and proposed a data-driven strategy that is less typically explored in the literature. In order to explain this issue, we have put forth a number of research questions, and we have attempted to address each of them across the entire manuscript.

- Addressed Question 1 (AQ1): What are the gaps between current experimental, theoretical, and computational approaches for mangrove rehabilitation and restoration and the presented data-driven approach?
- Addressed Question 2 (AQ2): How can data mining, which is an aspect of data science, be put forward in the study of species diversity?
- Addressed Question 3 (AQ3): What kind of species diversity data could be addressed by employing the data mining methodology of association rule mining?
- Addressed Question 4 (AQ4): How could the proposed approach be justified?
- Addressed Question 5 (AQ5): What is the significance of this study and how could the current approach be helpful for conservationists?
- Addressed Question 6 (AQ6): What recommendations and directions for the future may be drawn from this research?

The sections where each of the aforementioned AQs is answered are shown in Figure 14.1.

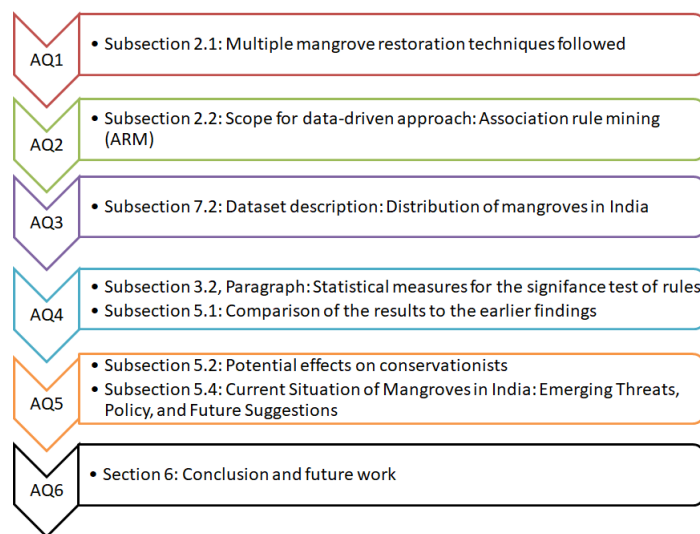


Figure 14.1: Addressed Questions in this research

Contribution Recapitulating the above discussion, the contribution made in this research towards study-based ecological restoration and rehabilitation is listed below:

1. Assembling information on Indian mangroves, including information on the Sundarban Mangroves, from both published and unpublished sources, and performing a preliminary analysis of how the different mangrove communities vary and are similar using biodiversity indexes.
2. Identifying the most pertinent set of mangrove associations from the huge quantity of frequent co-occurrence data on mangroves that may be delivered to the restoration practitioners. To achieve this, a novel domain-specific rule filtering metric based on taxonomic distinctness has been put forth. It would assist in study-based mangrove restoration through afforestation and reforestation while attaining the highest possible biodiversity richness.

14.2 Previous studies and scope of the data mining approach

The different mangrove restoration techniques that have been followed to date are precisely discussed here.

Direct planting First category can be named as direct planting [311]. The most frequently planted species in monogeneric planting were *Rhizophora apiculata* and *Rhizophora stylosa*; multi-species planting also falls under direct planting. In their persuasive argument, Lee and colleagues [312] argue with the idea that planting monogeneric stands of the widely utilized mangrove species *Rhizophora* (stilted mangroves) is the best way to restore lost mangrove cover. They contend that in order to mangrove recover, it is imperative to maintain the habitat that already exists and to rehabilitate the degraded environment, particularly on mangrove lands that have been used for farming and shrimp production.

Coastal engineering techniques The second category can be identified as the coastal engineering technique. It is used in both hard (different kinds of sea dykes and breakwaters) and soft engineering techniques (bamboo T-groins and fences, *Melaleuca* entrapping microsites before planting or to promote natural recruitment). In [313, 314], it has been observed that a planting phase may not be necessary for a successful mangrove restoration project. Natural regeneration mechanisms may be able to save mangroves from extinction, if stressful circumstances are removed and acceptable environmental conditions, such as adequate irrigation and quiet areas, are given, especially on exposed coasts. It has been experienced in the study of the west coast of Peninsular Malaysia.

Hydrological rehabilitation One of the primary causes of mangrove degradation has been identified as changes in hydrology. In the Celestun, Mexican mangrove rehabilitation area, secondary succession has taken place after a water reconnection was the only restoration measure implemented. Therefore, before planting, physical alterations need to be made to the site's hydrological conditions (taking into account surface elevation, tidal inundation, etc.) to promote natural regeneration. For example, in [315], the findings of this study demonstrate that the secondary succession and mangrove forests rehabilitation procedures were favored when reviving an aquatic connection with a coastline lagoon as the special restoration measure in a deteriorated mangrove region.

Data driven restoration In this field of research, data mining and machine learning algorithms have primarily served as tools to efficiently answer challenging analytical issues. Applications based on data mining and machine learning have been implemented concurrently with the other methodology of secondary data sources. Knowledge-based classifiers have been employed to categorize mangrove species [193, 194] and identify changes in mangrove cover over time [196]. Recently, we have proposed [63] a framework for recognizing the co-existence of salt marshes and mangroves for the regeneration of coastal forests. A proposition for determining dark diversity along with data mining methodology for combat biodiversity loss has been made in [62].

Summarized review on the existing mangrove rehabilitation and restoration approaches It has been seen that both rehabilitation and restoration have been followed by the mangrove restoration practitioners [316]. The primary objective of land use managers is rehabilitation, which aims at replacing any lost or impaired ecological function. Instead, ecologists emphasize restoration as the process of recovering an ecosystem to its original condition, as much as is feasible [316]. Gaps, homogeneity, and policy applicability are particularly problematic at the government level. Certain rehabilitation efforts have had varying degrees of success due to a variety of factors, including the use of improper strategies, a failure to involve local populations, or a failure to follow all the steps in the procedures that have been described in the literature. Therefore, mangrove restoration calls for an interdisciplinary approach. [317, 318, 319], which requires expertise from many different fields in the public sector, academia, and the community.

Scope for data-driven approach Previous studies have shown that several collaborative efforts have been conducted for the preservation and rehabilitation of mangroves [320]. Yet, a thorough study of the growth pattern of frequently co-occurring mangroves along the

Table 14.1: Multiple mangrove restoration techniques followed

Technique	Percentage of usage	Remarks/ Limitation
Direct planting [323, 187, 311, 312, 324]	- Used as the main restoration technique in all South East Asian countries (reported in 74% of the studies) [183]	- Yet monogeneric planting's efficacy has been questioned, at least in terms of habitat functionality [325] and coastal protection [326], as well as in terms of promoting faunal biodiversity [327, 328].
Coastal engineering techniques [329, 313, 314, 330]	- Engineering measures were incorporated with varied designs to facilitate restoration work (reported in 18% of the studies)[183]	- It disrupts natural processes, including sediment movement and tidal inundation, which can be harmful to the development and well-being of mangroves [331, 332].
Hydrological rehabilitation [333, 315, 315]	- In order to facilitate natural regeneration before planting, hydrological rehabilitation can be experimented (reported in 9% of the studies)[183]	- The process of hydrological rehabilitation might be costly. In regions where the mangrove ecosystem has been substantially damaged, hydrological rehabilitation may not be successful. Some species may gain when water flow is restored to a damaged mangrove habitat, but others that have adapted to the altered hydrology may suffer consequences [335, 336].
Data-driven approach [48, 63, 194, 196, 62]	- Despite the fact that very few researchers have used it, its effectiveness and influence have gradually become known.	- Encourage the use of natural restoration techniques. - It outlines a cutting-edge, data-driven management plan for rehabilitation. - The generated rules have been evaluated against the entirety of existing literature because rules depend on datasets being accurate. - Decisions based on research made before plantations save human resources.

various coastal ecosystems is still lacking. For example, a study in [321] has identified that adopting the wrong species and choosing the wrong site are the two main causes of poor survival. Instead of the native colonizers *Avicennia* and *Sonneratia*, the preferred but inappropriate *Rhizophora* are planted in exposed coastlines. ARM[322, 62] is an important data mining task used for finding close associations among the data items, predictions, recommendations, and other purposes. In this scenario, ARM can be exploited to find significant co-occurrences. So, a data-driven framework could be proposed based on interestingness measures of ARM to illustrate the retrieval of knowledge regarding species associativity. To this end, ARM extracts a large set of association rules. For obtaining the most significant rules, domain knowledge-based rule filtering measures can be used in addition to the conventional measures of ARM. Hence, a set of significant rules can be presented as useful for domain researchers in species-rich plantations and ecosystem restoration.

14.3 Materials and methods

14.3.1 Dataset description

Please refer to Table 5.2 and 5.3 in Chapter 5 for the occurrence information on the 34 mangrove species that have been identified, and for the information on the taxonomic details and distinctive taxonomic identification numbers. The metadata of the assembled datasets for the Indian mangroves are described in Table 14.2.

Along the east and west coast of India, 19 estuaries have been found along with Andaman. Estuaries have the most dynamic environments because saltwater from the sea and freshwater from rivers combine there. India's largest estuaries are primarily found along its eastern coast. The main estuaries considered in this study are: Hooghly-matla estuary, Subarnarekha estuary, Brahmani-baitarani estuary, Mahanadi estuary, Bhitarkanika, Godavari

Table 14.2: Description for the compiled datasets (Refer to Table 5.2 and 5.3 in Chapter 5)

Sl no	Description	Dimension	Rows	Columns
1	Indian mangroves occurrence data	34 x 19	34 Indian mangroves	19 estuaries along the east and west coasts
2	Taxonomic details for the mangroves in India	34 x 9	34 Indian mangroves	Unique identifier (WoRMS ID and ITIS TSN and GBIF taxon id), and taxonomic hierarchy

estuary, Vamsadhara estuary, Kakinada bay, Krisna delta, Pennar estuary, Pichavaram mangrove, Cauvery estuary, Vellar estuary, Ennore estuary, Andaman Island. Wandoor mangrove forest, Cochin estuary, Zuari estuary, Mandovi estuary, and Tapi estuary.

14.3.2 Proposed framework

In this study, a framework has been proposed using association rule mining (ARM) for filtering the association rule. Since ARM is able to extract every possible association, this results in an enormous collection of rules. Therefore, for a domain practitioner, it is not feasible to follow this substantial number of association rules. Two main procedures have been used in this framework, notably the use of both objective and subjective metrics, to extract significant association rules.

The whole framework for pruning important association rules is shown in Figure 14.2.

- The initial step would be to construct datasets. A method for mining association rules can be applied to a binary presence/absence dataset of species.
- Following that, the objective measures are used for association rule filtering.
- The next step performs a statistical validity assessment of the rules.
- Finally taxonomic distinctness is used as a subjective measure for generating the final set of statistically significant rules having maximum possible biodiversity richness.

The final set of filtered association rules is presented in the decreasing order of the total taxonomic distinctness.

Objective measures for rule filtering The objective measures for the rule filtering methodology primarily use the support-confidence framework, including filtering through lift value, eliminating redundancy, and the chi-squared test for statistical independence. All of these are measured as part of the development of association rules to evaluate the potency and significance of the relationships. These metrics are used to weed out false or weak associations and to identify those that are statistically significant and have practical applicability. For varied confidence levels, a small or large set of association rules is constructed depending on whether the support values are higher or lower. Most often, a lower support value is utilized to prevent data loss, creating a large number of rules in the process.

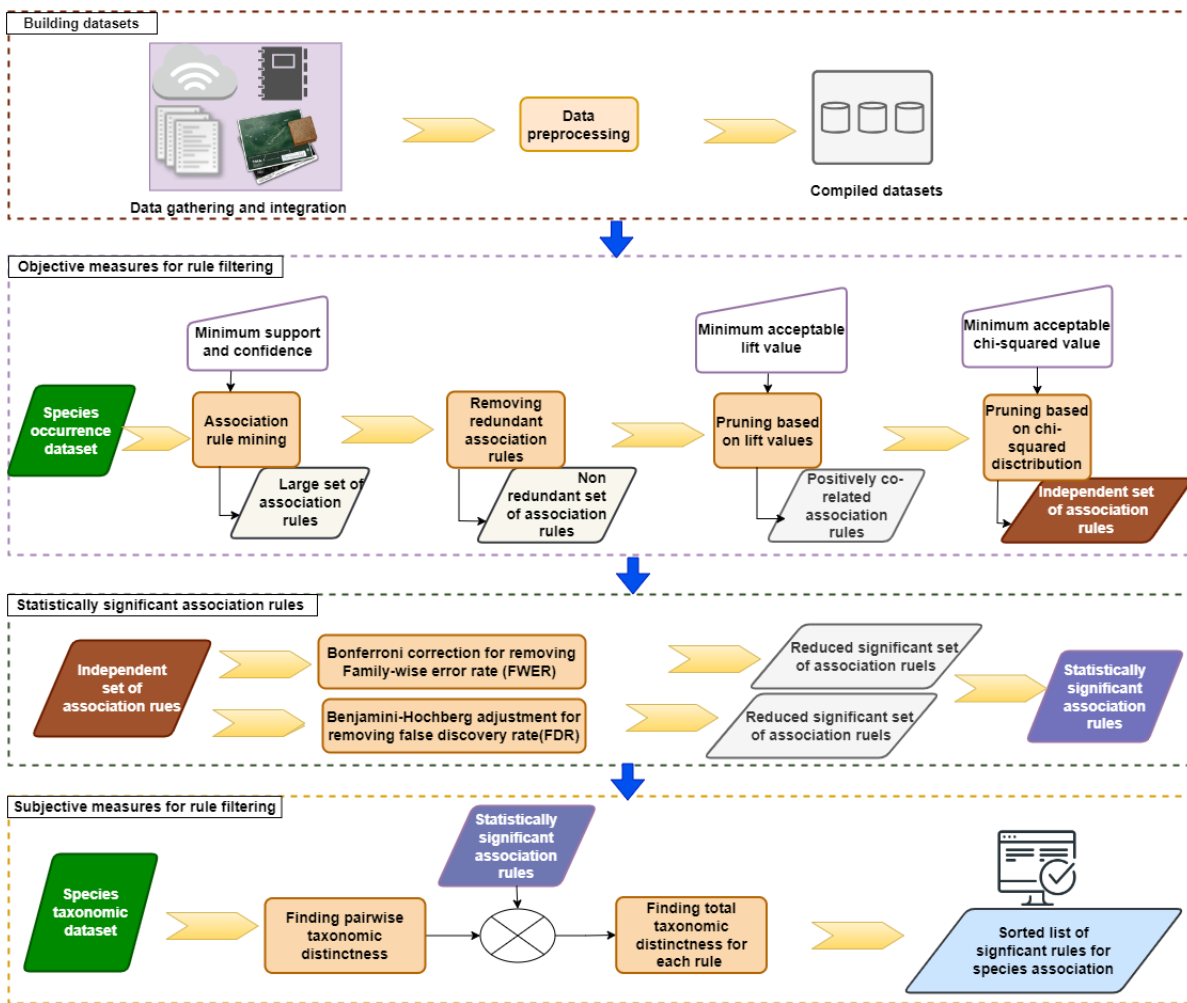


Figure 14.2: Proposed framework for rule filtering

Statistical measures for the significance test of rule The validity and significance of the relationships discovered using association rule mining can be evaluated using the chi-squared statistic. Even though the chi-squared statistic may be useful for assessing the significance of connections, it is essential to consider the particular characteristics of the data and its possible constraints. The proposed method is intended to be used with data on the co-occurrence of plant species. To achieve trustworthy and significant results in species co-occurrence analysis, it should be combined with statistical analysis with domain expertise, and take into account different validation methods as necessary. The following considerations specific to species co-occurrence data have been kept in mind:

1. Species Interactions: In species co-occurrence data, species interactions, and ecological factors play a crucial role. The chi-squared statistic assumes independence between items, which may not hold true for species co-occurrence. Ecological factors, such as competition or facilitation between species, can lead to non-independence. Therefore, it's important to interpret the results of the chi-squared statistic with caution and consider domain knowledge to account for potential confounding factors. Chi-squared statistics have been used to reduce the enormous number of generated rules after using small support and confidence thresholds. However, we do not rely completely on the

chi-squared test. Domain knowledge of taxonomic distinctness has been employed in a later stage to get the final deliverables.

2. **Multiple Comparisons:** In species co-occurrence analysis, multiple association rules may be generated and tested simultaneously. This increases the risk of obtaining false-positive associations due to multiple comparisons. Applying appropriate multiple testing corrections, such as Bonferroni correction or false discovery rate control, can help mitigate this issue and provide more reliable validation. We have applied both Bonferroni and Benjamini corrections in the statistical validation test.

Therefore, although association rule generation is not a rigorous hypothesis test, it does involve multiple hypothesis testing, and it is important to limit the false discovery rate (FDR) or family-wise error rate (FWER) to ensure that the associations found are not simply the outcome of chance. We use these two methods to find a statistically significant set of rules. In this regard, FWER [337] is an assessment of the likelihood of getting one or more false positives (Type I errors) when doing several hypothesis tests concurrently. One often employed method to adjust the FWER is the Bonferroni correction [338], which modifies the significant threshold of each test depending on the number of tests run. With more tests performed, there is a higher chance of at least one false positive, which can lead to incorrect conclusions.

Since FWER is conservative, it is more likely to make a Type II error—failing to recognize an intriguing rule. Benjamini and Hochberg [337] suggested an approach to FDR control. The main goal of Benjamini and Hochberg’s work is to reduce the false discovery rate (FDR) when evaluating numerous hypotheses.

Subjective measures for rule filtering using taxonomic distinctness: We could now propose the ranking of a collection of association rules based on the values of the rules’ total taxonomic distinctness, which could be ranked in sorted order. The rules with more total taxonomic distinctness could suggest more species richness in an area. This is how we apply domain-specific knowledge in association rule filtering along with existing statistical interestingness measures.

An association rule is understood to be derived through frequent relationships between items. Let’s say a rule has the following syntax: $\langle A, B, C \rangle \rightarrow \langle D \rangle$ where D represents the focal species. Therefore, it is possible to classify A , B , C , and D as closely related species with shared traits. We know that, for any two species, the taxonomic distinctness can be calculated. On the basis of the taxonomic distinctness of each individual pair, the total taxonomic distinctness for a rule might then be determined. The total taxonomic distinctness for the rule can be expressed as below,

$$TD = (u + v + w + x + y + z), \tag{14.1}$$

where, u , v , w , x , y , z are the pairwise taxonomic distinctness between A - B , A - C , B - C , A - D , B - D , C - D , respectively, as depicted in Table 14.3. Hence, we sort and rank all the rules. Rules with higher taxonomic distinctness will be of higher priority to the policymakers.

Table 14.3: Taxonomic distinctness between individual pair

	<i>A</i>	<i>B</i>	<i>C</i>
<i>B</i>	u		
<i>C</i>	v	w	
<i>D</i>	x	y	z

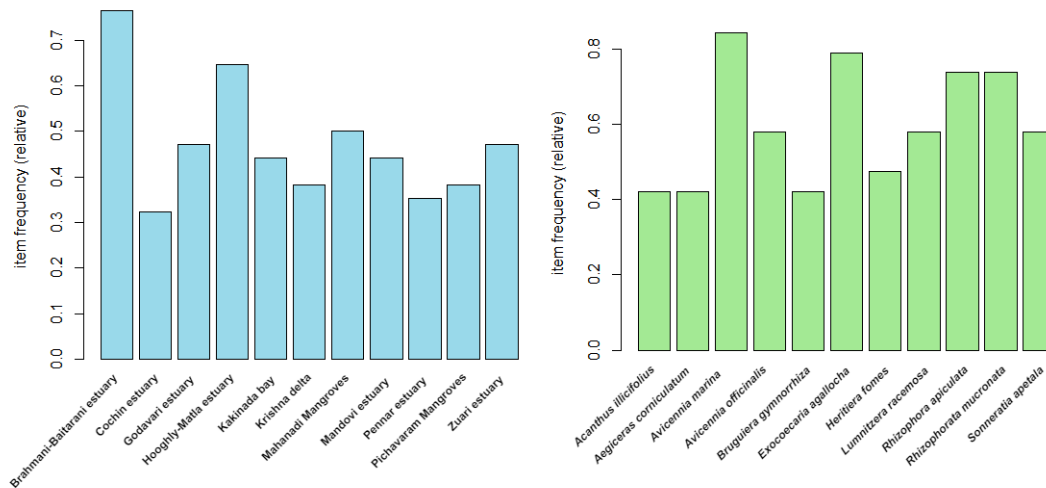


Figure 14.3: Frequency plot: Diversity of mangroves along Estuaries (LHS) and Species spread over the estuaries (RHS)

14.4 Results

14.4.1 Analysis on frequency of occurrence

Finding frequently-occurring itemsets and extracting powerful association rules from those itemsets are the two key processes in the rule mining task. The Indian Mangrove dataset (Table 5.2) is used for studying frequent occurrences. Two frequency plots (Figure 14.3) from the Indian Mangrove dataset have been produced as the initial step in the process of finding the association rule.

The most diversified areas for Indian mangroves have been identified as the Hooghly-Matla estuary, the Mahanadi mangroves, and the Brahmani-Baitarani estuary (Figure 14.3, L.H.S. plot). The most frequent species in the study regions, however, are *Rhizophora epiculata*, *Exocoecaria agallocha*, *Rhizophora mucronata*, and *Avicennia marina* (Figure 14.3, R.H.S. plot).

14.4.2 Analysis on objective and subjective measures

Results on objective measures Among the Indian mangroves, the most crucial IUCN status is found for *Heritiera fomes*: Endangered, *Sonneratia griffithii*: Critically endangered, *Ceriops decandra*: Near threatened, *Phoenix paludosa*: Near threatened. Following the objective and subjective measures for rule filtering, as stated in Figure 14.2, first the statistically significant useful rules, and, subsequently, the ecologically significant frequent associations, are obtained. So, primarily, for each rule, the support, confidence, lift, and chi-squared

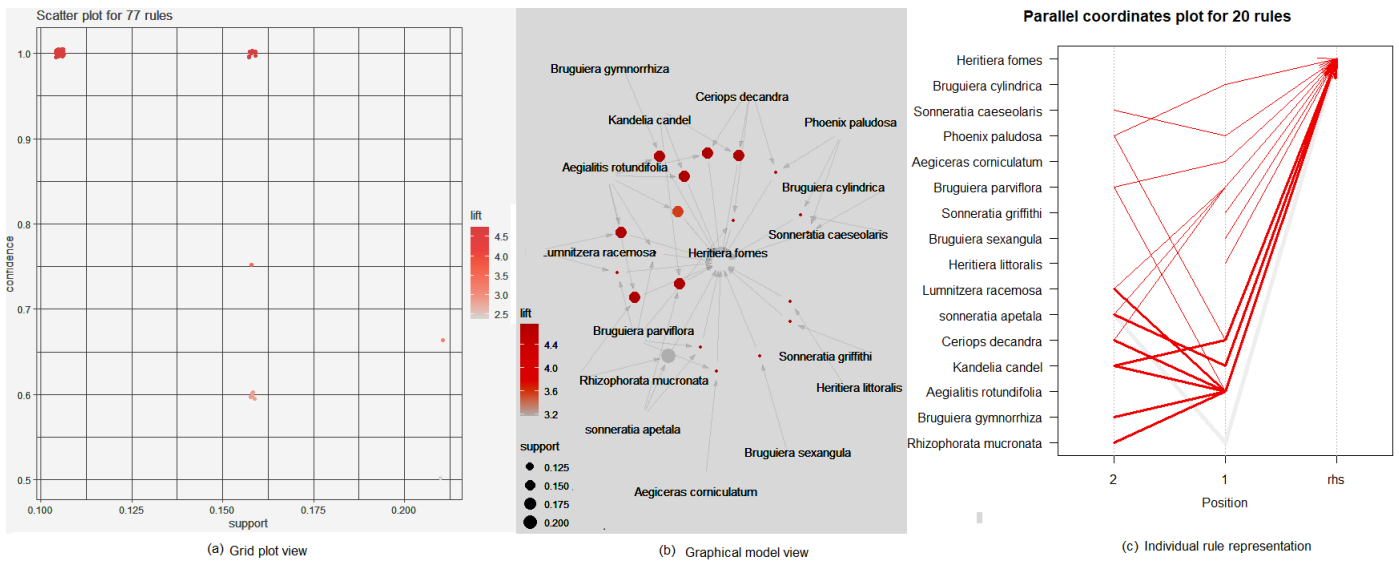


Figure 14.4: Visualization of the generated association rules for *Heritiera fomes* after objective measure

statistics are obtained. The filtering values for the measures are shown in Table 14.4. The support values have been chosen to obtain a feasible set of association rules. The confidence level for each rule has been considered to be more than 60%, although, in most of the cases, it is 1, specifying a stronger association rule. Lift values are also taken to be greater than 1 in all cases, indicating a stronger co-occurrence relationship among the species. The obtained chi-squared value for each rule shows the statistical independence. Figure 14.4 represents the rules obtained for *Heritiera fomes* from objective measures of filtering. In the R environment, plot 14.4(a) has been generated. Each rule has been plotted using the support and confidence measures along the X and Y axis, respectively, and it will be shaded according to the lift measure, demonstrating how strongly the items are associated with one another. It could be seen that the filtered rules have confidence 1 and lift ratios are reasonably high to show a strong association. As we proceed with the visualization, Figure 14.4(b) uses a graph to show the lifts and supports of different species, but mostly to show which species are associated with which in the growing habitat. Support levels determine the size and lift ratios define the color of graph nodes. For a rule, the antecedent, or LHS, is displayed in the entering lines, while the consequent, or RHS, is displayed in the arrowhead. Here, the rules are shown for *Heritiera fomes*. strong and frequent associations are found with *Ceriops decandra*, *Kandelia candel*, *Aegialitis rotundifolia* etc. A parallel coordinate plot in Figure 14.4(c) presents a parallel visual coordinate system. It would help us clearly to see which species along with which ones, form frequent associations with *Heritiera fomes*. As mentioned above, the RHS or consequent is the species under study (i.e. *Heritiera fomes*). The positions are in the LHS where 1 and 2 axis denotes the frequent species combination found along with *Heritiera fomes*. Similar kinds of visualization of the filtered rules generated after objective measures for *Sonneratia griffithii*, *Ceriops decandra* and *Phoenix paludosa* are shown in Figure 14.5, 14.6, and 14.7, respectively.

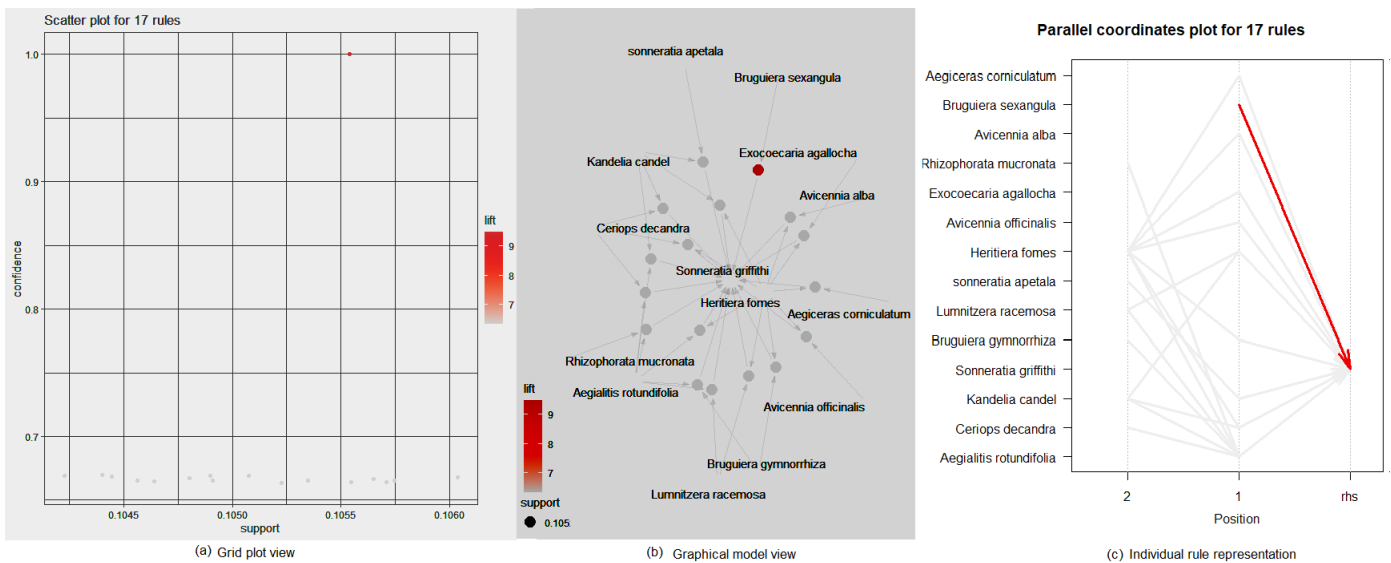


Figure 14.5: Visualization of the generated association rules for *Sonneratia griffithii* after objective measure

Results on significance test For multiple comparison problems, a statistical validity assessment has been carried out after the filtration by the objective measures of support, confidence, lift, and chi-squared values. Bonferroni correction has been observed for the method to adjust for multiple comparisons. The required significance level of alpha was set to 0.01. The significance level of 0.05 was used for identifying a statistically significant rule for the Benjamini-Hochberg test. The details are listed in Table 14.4.

Table 14.4: Useful set of rules after validation

Generated rules on Species	<i>Heritiera fomes</i>	<i>Sonneratia griffithii</i>	<i>Phoenix Paludosa</i>	<i>Ceriops Decandra</i>
Characteristics				
# Rules	2289783	2289783	2289783	2289783
# Covering rate	100	100	100	100
Support threshold	0.1	0.1	0.1	0.1
Confidence threshold	0.5	0.5	0.5	0.5
Lift	> 0.2 of maximum lift	> 0.5 of maximum lift	> 0.2 of maximum lift	> 0.15 of maximum lift
Chi-squared test	> 6	>6	>6	>6
Results				
# Association rule after objective measure	77	17	38	17
# After significance test by Bonferroni correction	0	17	3	0
# After significance test by Benjamini-Hochberg correction	7	17	38	17

Multiple comparison techniques like the Bonferroni correction and Benjamini-Hochberg adjustment are frequently employed in statistical hypothesis testing to reduce FWER or FDR. Nevertheless, depending on the dataset and the particular test being run, they operate differently and can produce various results. The family-wise error rate (FWER), which is the likelihood of creating at least one type I error (false positive) among all the hypotheses tested, is controlled using the Bonferroni correction, a very rigorous technique. The Bonferroni adjustment minimizes the likelihood of a false positive for each individual test by dividing the significance level (alpha) by the number of tests being run. As a result, compared to other procedures, the Bonferroni correction has a tendency to be stricter and may produce a larger rate of false negatives (missing real positives).

A DATA-DRIVEN RULE FILTERING APPROACH FOR FOREST RESTORATION

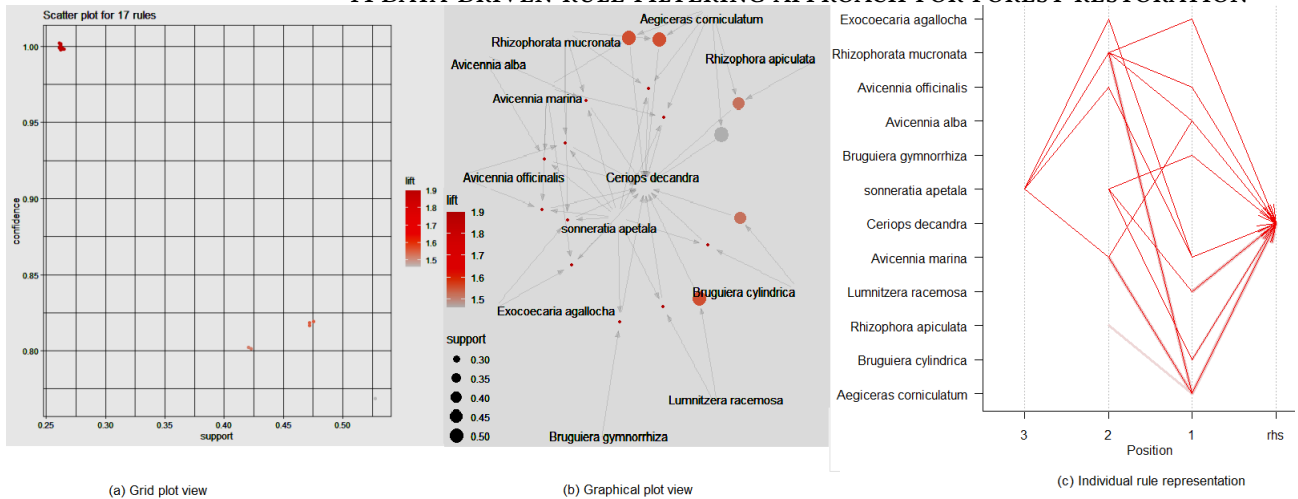


Figure 14.6: Visualization of the generated association rules for *Ceriops decandra* after objective measure

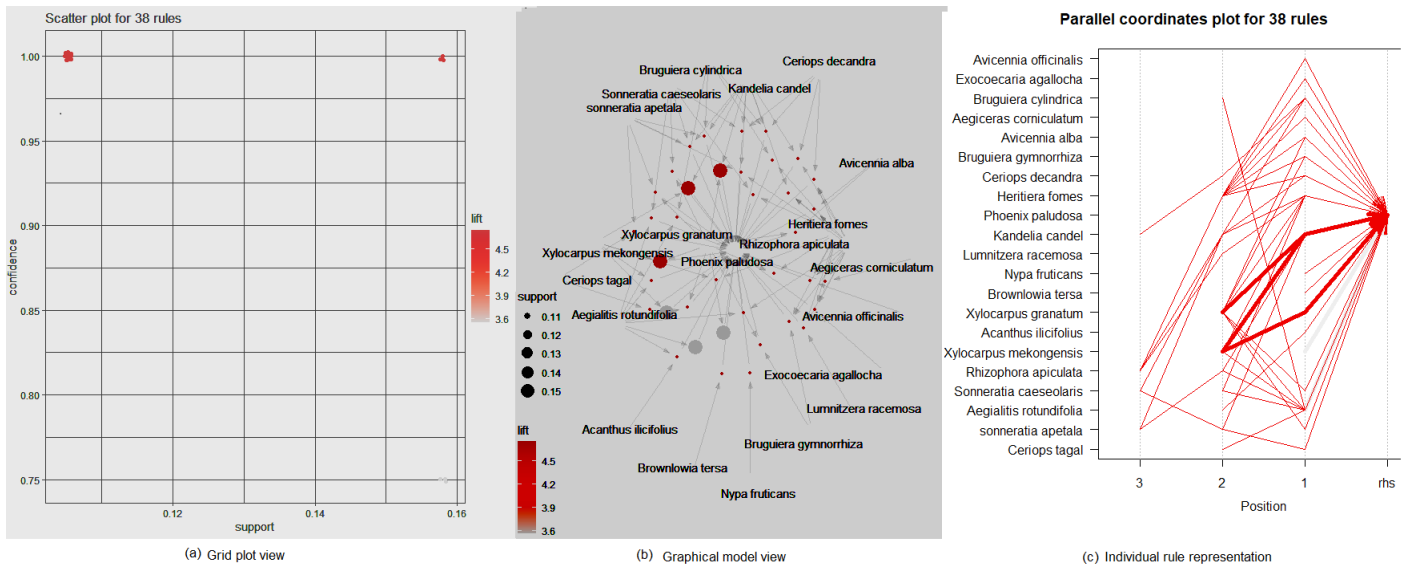


Figure 14.7: Visualization of the generated association rules for *Phoenix paludosa* after objective measure

The FDR, or false positive rate (FPR), is the ratio of false positives to all rejected hypotheses. The Benjamini-Hochberg correction, on the other hand, is a less conservative approach that regulates the FDR. The FDR and the number of tests being run both affect the critical value used in the Benjamini-Hochberg correction, which ranks the p-values from lowest to highest and sets a cut-off based on it. The rate of true positives and false positives may both be higher with this procedure than with the Bonferroni adjustment due to its less exact nature.

Therefore, depending on the significance level, the number of tests, and the magnitude of the effects being evaluated, it is feasible that the Bonferroni adjustment produces a null set while the Benjamini-Hochberg correction provides a not-null set for the same dataset, in order to achieve the required trade-off between type I and type II errors. As we have seen for *Heritiera fomes* and *Ceriops decandra*, the Bonferroni correction returns a null set (as shown in Table 14.4). Table 14.5, 14.6, 14.7, and 14.8 list the pruned association rules for *Heritiera fomes*, *Sonneratia griffithii*, *Ceriops decandra*, and *Phoenix paludosa*, respectively,

Table 14.5: Association rules for *Heritiera fomes*: After significance test

Rules
Set of rules after Bonferroni correction
Null Set
Subset of rules after Benjamini correction
<i>Aegialitis rotundifolia</i> , <i>Kandelia candel</i> => <i>Heritiera fomes</i>
<i>Aegialitis rotundifolia</i> , <i>Ceriops decandra</i> => <i>Heritiera fomes</i>
<i>Aegialitis rotundifolia</i> , <i>Bruguiera gymnorrhiza</i> => <i>Heritiera fomes</i>
<i>Aegialitis rotundifolia</i> , <i>Lumnitzera racemosa</i> => <i>Heritiera fomes</i>
<i>Aegialitis rotundifolia</i> , <i>Rhizophorata mucronata</i> => <i>Heritiera fomes</i>
<i>Kandelia candel</i> , <i>sonneratia apetala</i> => <i>Heritiera fomes</i>
<i>Ceriops decandra</i> , <i>Kandelia candel</i> => <i>Heritiera fomes</i>

Table 14.6: Association rules for *Sonneratia griffithii*: After significance test

Rules
Set of rules after Bonferroni correction
<i>Bruguiera sexangula</i> => <i>Sonneratia griffithii</i>
<i>Aegialitis rotundifolia</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Heritiera fomes</i> , <i>Kandelia candel</i> => <i>Sonneratia griffithii</i>
<i>Ceriops decandra</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Bruguiera gymnorrhiza</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Heritiera fomes</i> , <i>Lumnitzera racemosa</i> => <i>Sonneratia griffithii</i>
<i>Avicennia alba</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Aegiceras corniculatum</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Avicennia officinalis</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Exocoecaria agallocha</i> , <i>Heritiera fomes</i> => <i>Sonneratia griffithii</i>
<i>Aegialitis rotundifolia</i> , <i>Ceriops decandra</i> => <i>Sonneratia griffithii</i>
<i>Aegialitis rotundifolia</i> , <i>Bruguiera gymnorrhiza</i> => <i>Sonneratia griffithii</i>
<i>Aegialitis rotundifolia</i> , <i>Lumnitzera racemosa</i> => <i>Sonneratia griffithii</i>
<i>Aegialitis rotundifolia</i> , <i>Rhizophorata mucronata</i> => <i>Sonneratia griffithii</i>
<i>Kandelia candel</i> , <i>sonneratia apetala</i> => <i>Sonneratia griffithii</i>
<i>Ceriops decandra</i> , <i>Kandelia candel</i> => <i>Sonneratia griffithii</i>
Set of rules after Benjamini correction
The same set of rules is generated

after statistical validation.

Results on subjective measures The taxonomic distinctness TD between any two species can be derived as discussed in the paragraph named subjective measure of rule filtering in subsection 14.3.2. Figure 14.8 presents a heatmap with dendrograms showing the hierarchical clustering of species under study based on their taxonomic distinctness. We consider the most frequently co-occurring species (identified after statistical validation) in building the heatmap. The TD values in the co-relation matrix are obtained by using the function `taxa2dist()` in *vegan package*. Considering the taxonomic rank, *Ceriops decandra* and *Kandelia candel* are in differ in genus level, therefore having minimum TD, whereas, *Ceriops decandra* and *Phoenix paludosa*, differs in class level in itself, having maximum TD. These TD values are considered in computing the total TD.

The obtained set of statistically significant frequent associations is sorted starting from the most significant ecological value, i.e., having the highest total taxonomic distinctness among the species in the collection, during the filtering process using subjective measures. Figure 14.9 depicts the hierarchical cluster for the frequent co-occurred species list of *Heritiera fomes* discovered following the objective filtration and statistical validation. The pre-

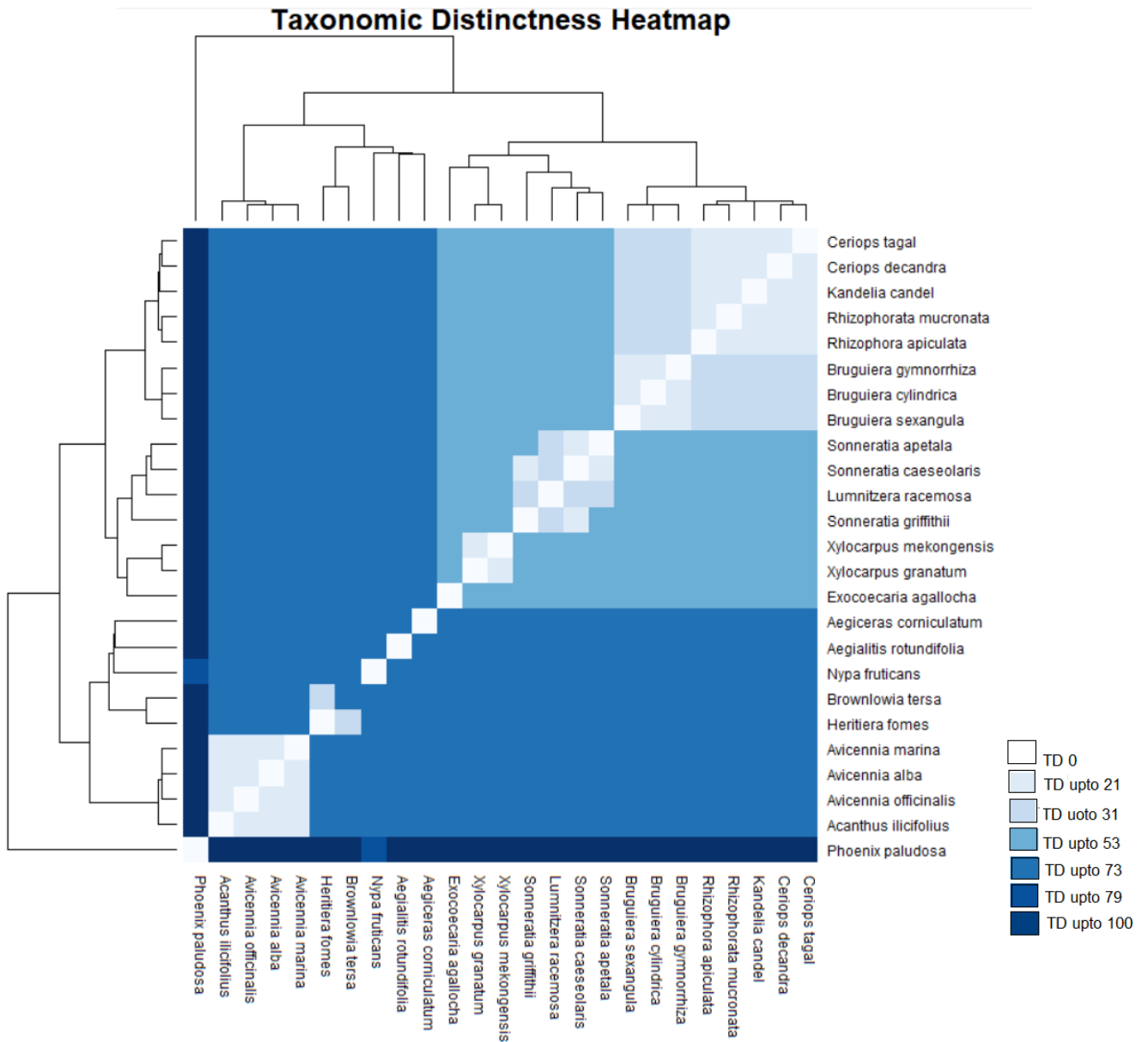


Figure 14.8: A heatmap with dendrograms showing hierarchical clustering of species under study based on their taxonomic distinctness(TD). The values are obtained using *taxatodist()* representing difference along taxonomic hierarchy

Table 14.7: Association rules for *Ceriops decandra*: After Significance test

Rules
Set of rules after Bonferroni correction
Null set
Set of rules after Benjamini-Hochberg correction
<i>Aegiceras corniculatum</i> => <i>Ceriops decandra</i>
<i>Lumnitzera racemosa</i> => <i>Ceriops decandra</i>
<i>Aegiceras corniculatum</i> , <i>Avicennia marina</i> => <i>Ceriops decandra</i>
<i>Aegiceras corniculatum</i> , <i>Rhizophorata mucronata</i> => <i>Ceriops decandra</i>
<i>Bruguiera cylindrica</i> => <i>Ceriops decandra</i>
<i>Aegiceras corniculatum</i> , <i>Rhizophora apiculata</i> => <i>Ceriops decandra</i>
<i>Bruguiera cylindrica</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Bruguiera gymnorrhiza</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Lumnitzera racemosa</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Avicennia alba</i> , <i>Avicennia marina</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Avicennia alba</i> , <i>Rhizophorata mucronata</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Aegiceras corniculatum</i> , <i>Avicennia marina</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Avicennia marina</i> , <i>Avicennia officinalis</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Avicennia marina</i> , <i>Exocoecaria agallocha</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Aegiceras corniculatum</i> , <i>Rhizophorata mucronata</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Avicennia officinalis</i> , <i>Rhizophorata mucronata</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>
<i>Exocoecaria agallocha</i> , <i>Rhizophorata mucronata</i> , <i>sonneratia apetala</i> => <i>Ceriops decandra</i>

sensation uses the popular R function [339], which employs the agglomeration approach for hierarchical clustering, with each species beginning in its own cluster and the two most related species being clustered below. *Rhizophora mucronata*, *Kandelia candel*, and *Ceriops decandra*, for example, are all members of the *Rhizophoraceae* family but differ in genus (Refer to Table 5.3). As a result, their taxonomic distinctness is lower, causing them to cluster together. *Sonneratia apetala* and *Lumnitzera racemosa* are members of the same order *Myrtales*, which is distinct from *Rhizophora mucronata*, *Kandelia candel*, and *Ceriops decandra*. *Heritiera fomes* and *Aegialitis rotundifolia*, on the other hand, belong from the same class *Magnoliopsida*, although they diverge further down the taxonomic tree. As a result, three significant clusters (shown by colors) are discovered.

Figure 14.10 lists the ranked association rules after employing subjective measure of filtering for *Heritiera fomes*. The ranking is based upon the total TD (refer to paragraph 14.3.2). The range of total TD, particularly for the rules generated for *Heritiera fomes* has been shown on the RHS in the color gradient (169 to 220). The values are obtained in the R environment using *vegan Package*. The final list of species association is shown in LHS. The priority of each rule has been shown in the color gradient. For example, *Aegialitis rotundifolia*, *Kandelia candel*, and *Heritiera fomes* are forming frequent associations, and their total TD is summed up to the highest value. They belong to different clusters (Figure 14.9), therefore maximizing the total TD. Therefore, it can be concluded that *Aegialitis rotundifolia*, *Kandelia candel*, and *Heritiera fomes* are forming the frequently co-occurred species list while enriching the biological diversity of the ecosystem (reported previously in literature [340]). The final set of significant and ranked frequent associations for *Sonneratia griffithii* is shown in Figure 14.12, for *Ceriops decandra* is shown in Figure 14.14, for *Phoenix paludosa* is shown in Table 14.16.

Table 14.8: Association rules for *Phoenix paludosa*: After Significance test

Rules
Set of rules after Bonferroni correction
<i>Xylocarpus granatum</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
Set of rules after Benjamini-Hochberg correction
<i>Xylocarpus granatum</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> => <i>Phoenix paludosa</i>
<i>Nypa fruticans</i> => <i>Phoenix paludosa</i>
<i>Brownlowia tersa</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Ceriops tagal</i> => <i>Phoenix paludosa</i>
<i>Ceriops tagal</i> , <i>sonneratia apetala</i> => <i>Phoenix paludosa</i>
<i>Heritiera fomes</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Sonneratia caeseolaris</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>sonneratia apetala</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Bruguiera gymnorrhiza</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Avicennia alba</i> , <i>Xylocarpus granatum</i> => <i>Phoenix paludosa</i>
<i>Heritiera fomes</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Xylocarpus mekongensis</i> => <i>Phoenix paludosa</i>
<i>Heritiera fomes</i> , <i>Sonneratia caeseolaris</i> => <i>Phoenix paludosa</i>
<i>Bruguiera cylindrica</i> , <i>Heritiera fomes</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Sonneratia caeseolaris</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Bruguiera cylindrica</i> => <i>Phoenix paludosa</i>
<i>Acanthus ilicifolius</i> , <i>Aegialitis rotundifolia</i> => <i>Phoenix paludosa</i>
<i>Aegialitis rotundifolia</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Heritiera fomes</i> , <i>Kandelia candel</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Ceriops decandra</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Bruguiera gymnorrhiza</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Heritiera fomes</i> , <i>Lumnitzera racemosa</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Avicennia alba</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Aegiceras corniculatum</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Avicennia officinalis</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Exocoecaria agallocha</i> , <i>Heritiera fomes</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>sonneratia apetala</i> , <i>Sonneratia caeseolaris</i> => <i>Phoenix paludosa</i>
<i>Ceriops decandra</i> , <i>Kandelia candel</i> , <i>Sonneratia caeseolaris</i> => <i>Phoenix paludosa</i>
<i>Bruguiera cylindrica</i> , <i>Kandelia candel</i> , <i>sonneratia apetala</i> => <i>Phoenix paludosa</i>
<i>Kandelia candel</i> , <i>Rhizophora apiculata</i> , <i>sonneratia apetala</i> => <i>Phoenix paludosa</i>
<i>Bruguiera cylindrica</i> , <i>Ceriops decandra</i> , <i>Kandelia candel</i> => <i>Phoenix paludosa</i>
<i>Ceriops decandra</i> , <i>Kandelia candel</i> , <i>Rhizophora apiculata</i> => <i>Phoenix paludosa</i>

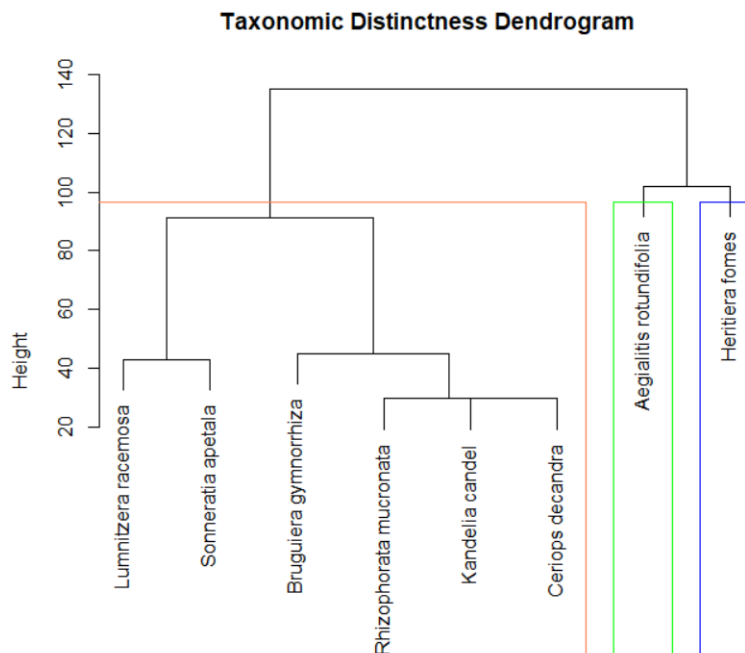


Figure 14.9: Hierarchical cluster of co-occurred species with *Heritiera fomes* based on taxonomic distinctness

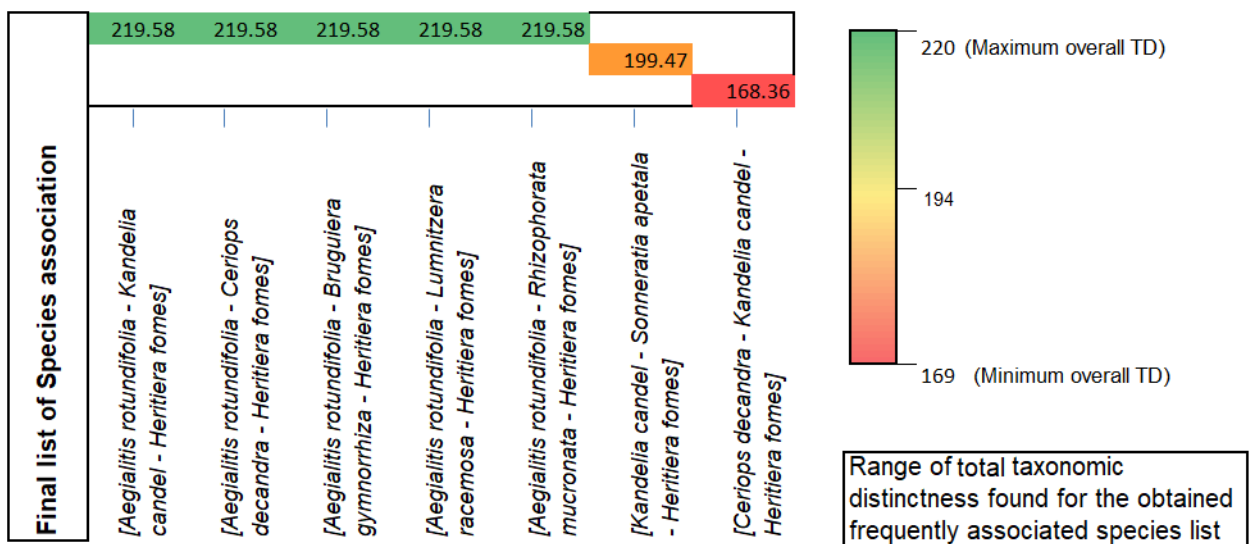


Figure 14.10: List of ranked associations for *Heritiera fomes* obtained by subjective measure of total taxonomic distinctness for an association

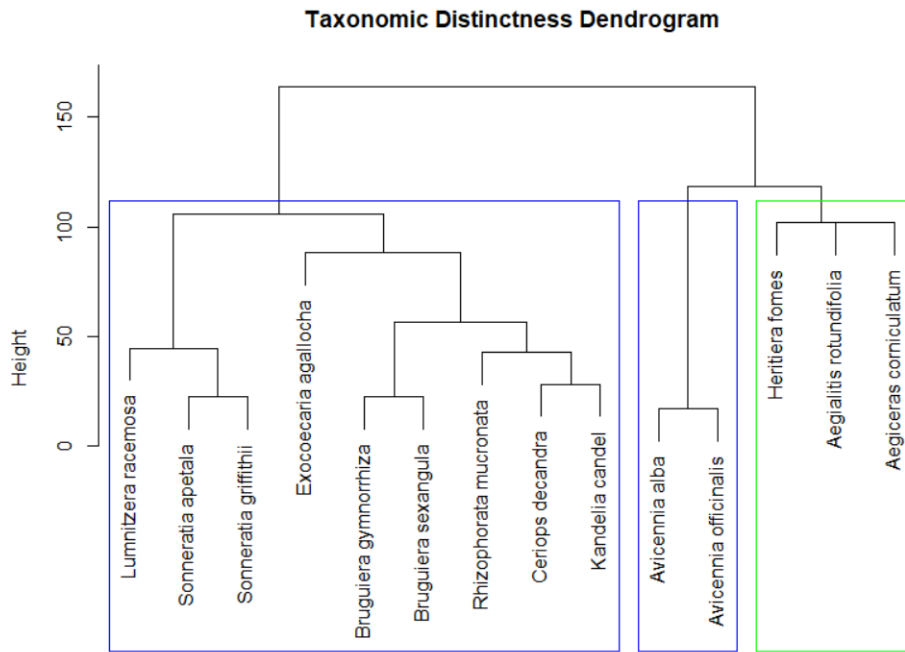


Figure 14.11: Hierarchical cluster of co-occurred species with *Sonneratia griffithii* based on taxonomic distinctness

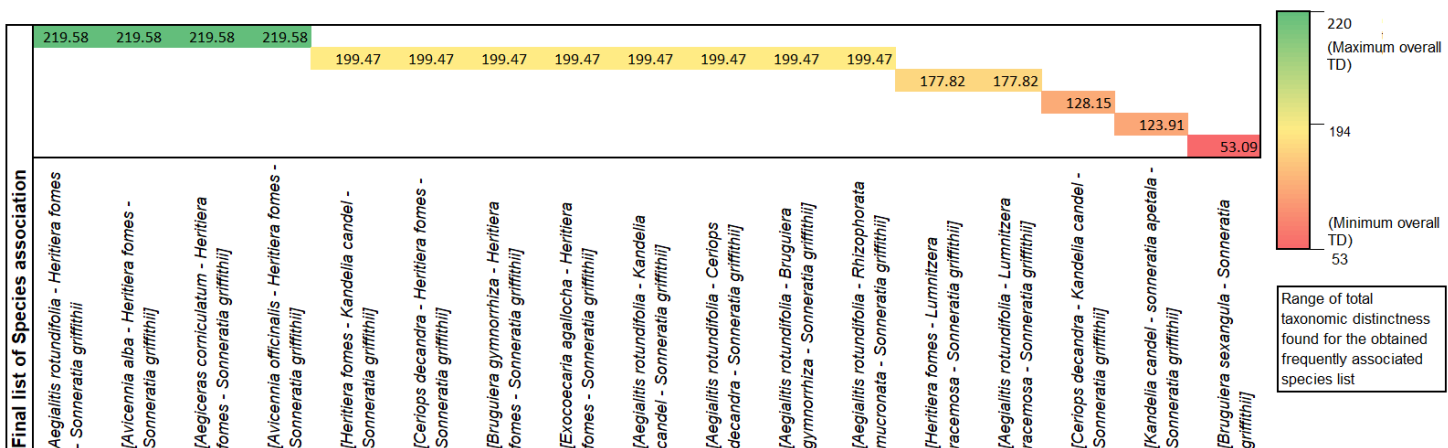


Figure 14.12: List of ranked associations for *Sonneratia griffithii* obtained by subjective measure of total taxonomic distinctness for an association

Taxonomic Distinctness Dendrogram

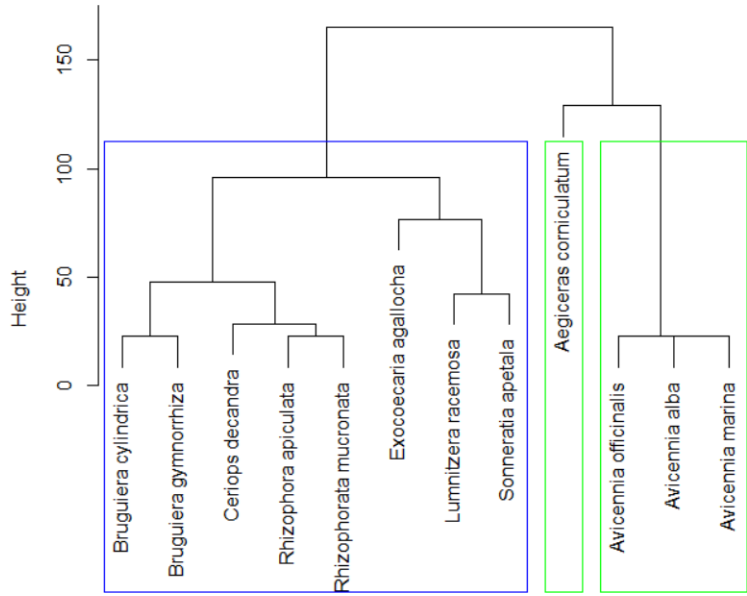


Figure 14.13: Hierarchical cluster of co-occurred species with *Ceriops decandra* based on taxonomic distinctness

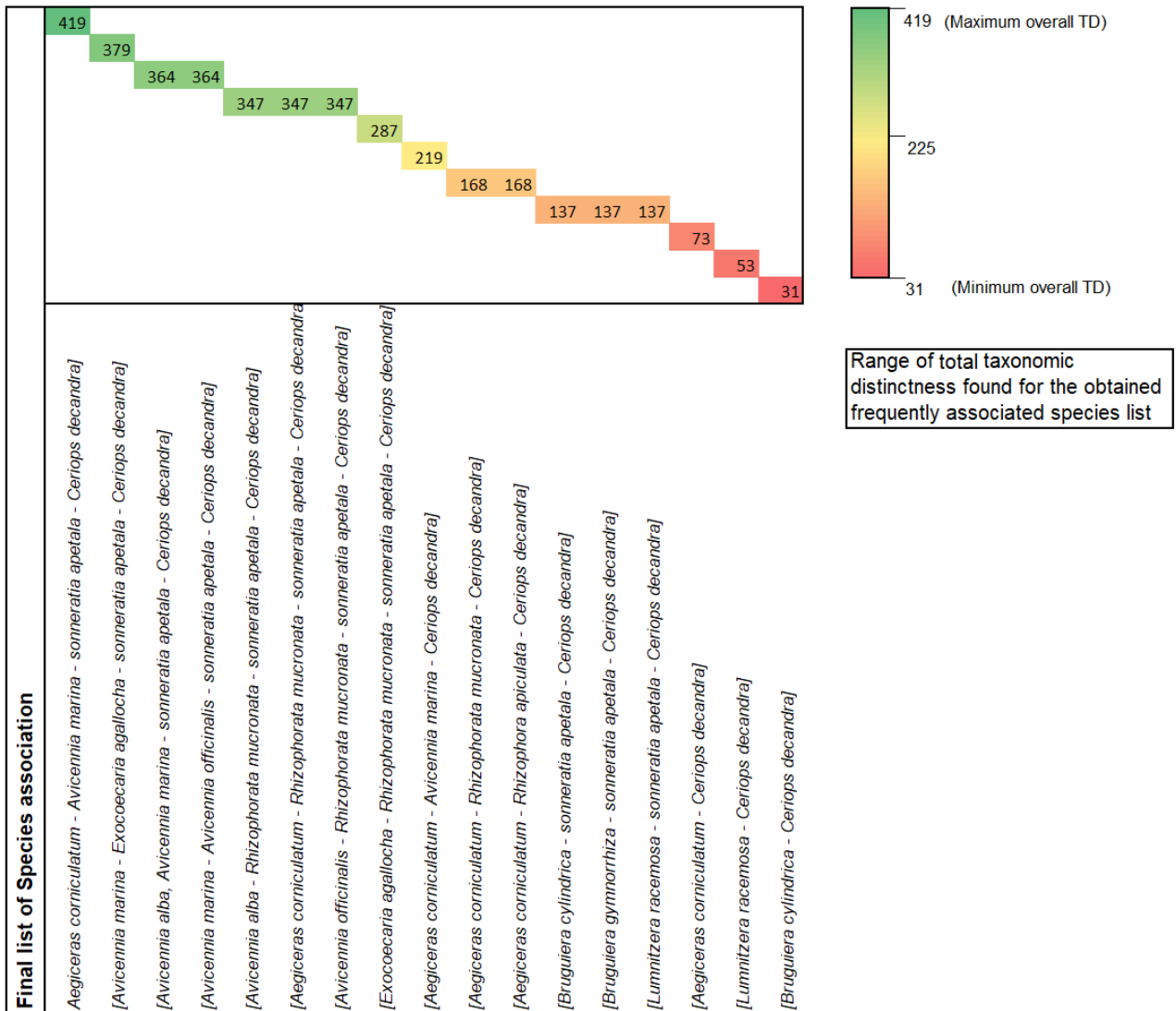


Figure 14.14: List of ranked associations for *Ceriops decandra* obtained by subjective measure of total taxonomic distinctness for an association

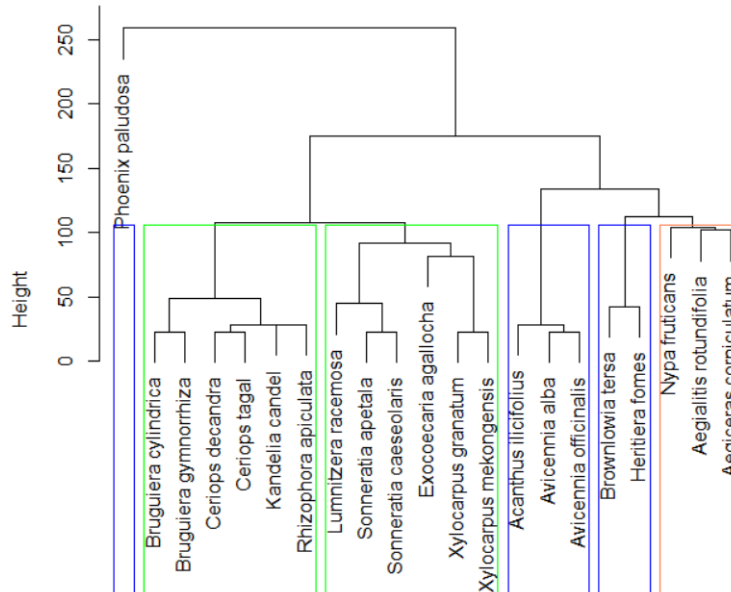


Figure 14.15: Hierarchical cluster of co-occurred species with *Phoenix paludosa* based on taxonomic distinctness

14.5 Discussion

The study regarding the frequent co-occurrence pattern of multiple mangrove species can be undoubtedly significant to conservationists for ecosystem restoration. In this work, we provide a useful way for obtaining important data on frequent co-occurrences across species that can maintain an elevated degree of species richness while achieving the highest levels of taxonomic distinctness. The suggested process is based on association rule mining, which relies on establishing frequent item sets and corresponding rules. For the purpose of filtering association rules from the perspective of conservationists, one domain-specific metric has been proposed in addition to several already existing interestingness measures.

14.5.1 Comparison to the earlier findings

The degree to which mangrove species can tolerate salinity and their preferred habitats have been studied in [341]. Every type of mangrove has a particular salinity range where it may thrive. As the environmental regime shifts, this tendency and its level of tolerance may also vary. On the basis of this concept, a few indicator species have been discovered according to their preferred salinity.

In low salinity areas, *Heritiera* is thought of as an indicator species [341]. Previous studies [342, 343] have found *Sonneratia* and *Lumnitzera* as the common species with *Heritiera*. Our research, which is summarised in Figure 14.10, has indicated that all of these have frequent correlations and when combined, can yield better taxonomic distinctness. A rich colony of *Sonneratia* has been identified as abundant in a wide range of salinity zones. A large variety of species, *H. fomes*, *A. rotundifolia*, *A. corniculata* have been observed to form frequent associations. *H. fomes* has a preference for low saline area [341], whereas increasing salinity hinders the plant growth of *A. corniculata* [344]. Similarly, higher salinity indicates less suitability of *A. rotundifolia*. The list of frequently associated species exhibiting high species richness with *Sonneratia* has been shown in Figure 14.11.

subsequent step of subjective measure, only significant rules are considered. The ranking shows the chronological order as per the total taxonomic distinctness calculated for all the associated species in a rule.

Throughout the span of the salinity gradient, the adaptation of Sundarban mangrove species has been found to vary significantly [347]. A common group of species can therefore be recognized based on environmental conditions. Hence, the restoration specialists might select mangroves from the frequently occurring assemblage of groups.

Here, a list of these frequent associations has been compiled using the suggested methodology. Any information related to a species that is being studied can be gleaned from the list. For instance, if growing the endangered species *Heritiera fomes* is the objective, our findings (Figure 14.10) could suggest that *Aegialitis rotundifolia*, *Kandelia candel* along with *Heritiera fomes* could achieve high taxonomic distinctness. Similar to this, *Aegialitis rotundifolia* and *Ceriops decandra* also form a rich and diverse species association. Multiple other species, viz, *Bruguiera gymnorrhiza*, *Lumnitzera racemosa*, *Rhizophora mucronata*, *Kandelia candel* form a similar taxonomic distinctness with *Aegialitis rotundifolia* and *Heritiera fomes*. So, practitioners can make an informed decision about the composition of species.

14.5.3 Comparison to concurrent research initiatives

This section discusses comparisons of concurrent research initiatives in terms of methodology, outcomes, and restrictions. Mangrove restoration has been addressed by multiple researchers in many parallel and ongoing projects, and those have been reviewed and listed in [63]. Several ways of restoration include the way of natural restoration, like, afforestation, reforestation [221, 189], study-based plantation for systematic mangrove habitat restoration projects using coastal protection, salt marsh plantation [63, 192, 222], a knowledge-based decision for restoration activities via classification of mangrove species, forest structure assessment, and the temporal and geographic distribution of mangroves and salt marsh [195, 194, 193, 196].

Here the authors have studied the approach to find the frequent co-occurrences of mangroves. A similar approach has been used by very few studies in the literature. A species conditional-occurrence algorithm [348] has been proposed for measuring conditional co-occurrence probability. The role of indicator species for conservationists is undoubtedly significant for ecosystem restoration. The suggested strategy offers a cutting-edge way of managing conservation areas, assessing the benefits of restoration, and monitoring habitat quality. In our previous studies [64, 58] we have used the idea of frequent co-occurrence, and association rule mining on species assemblage datasets. A huge number of association rule generation was the main drawback of this approach. As a result, filtration of this massive set of rules is required so that restoration practitioners can use it.

14.5.4 Current Situation of Mangroves in India: Emerging Threats, Policy, and Future Suggestions

Mangroves and Climate Change in India Multiple studies [349, 350] find the consequences of mangrove loss on poverty and climate change in India (specifically in the region of Sundarban which is dwelling to one of the biggest patches of mangrove forests in the world). The authors argue that the loss of mangroves has led to increased vulnerability to

climate change, and propose a range of policy interventions to support the conservation of mangroves and promote sustainable development. Mangrove forest health, crop quality, and soil quality are all at risk because of climate change and rising sea levels. Serious changes in fishing patterns and hydrological characteristics have also occurred, which have had severe effects on fishermen. Frequent cyclones and unpredictable monsoon patterns adversely affect the ecology of the area and local livelihood as well.

Mangroves and livelihood in India Mangroves are key sources of livelihood for local inhabitants in the Sundarban region. The degradation of mangrove habitats can result in declines in fish and crab populations, lowering the income and food security of these communities. Fishing and crabbing are major sources of income generation for many households. According to the findings of the econometric research [52, 53], the region's diminishing trends in the harvesting of prawn seeds, honey, and other non-timber forest items, pose the greatest risks to the lives of the locals in Sundarban. The results also show that seasonal interstate migration is more common in Sundarban households with family incomes of less than \$50, higher dependency rates (>5), and fewer people with land ownership. The main reason for such migration was stated by more than 70% of those surveyed as livelihood risks brought on primarily by the consequences of climate change. Overall, the loss of mangroves in the Indian Sundarban region can have a major effect on poverty levels by lowering the availability of livelihood possibilities and exposing inhabitants nearby to greater risk of natural catastrophes and climate change effects. A few studies [308, 351] have attempted to emphasize the possibilities of a few non-traditional alternative vocations that may improve the economic standing of the offshore dwellers of the Indian Sundarban, many of whom are surviving Below the Poverty Line.

Mangroves and Indian laws and regulation A recent report [352] on the conservation and protection of mangroves states that the government has put in place both regulatory and promotional initiatives to safeguard, maintain, and expand the forest area across the country. Through the Coastal Regulation Zone (CRZ) Notification (2019), various regulatory measures are being carried out in accordance with the Environment (Protection) Act of 1986, the Wild Life (Protection) Act of 1972, the Indian Forest Act of 1927, the Biological Diversity Act of 2002, and rules and amendments are framed with these acts. Every two years, the Forest Survey of India (FSI) evaluates the mangrove cover of the country and categorizes it into three density classes: very dense, moderately dense, and open mangrove cover. The results are published accordingly in the India State of Forest Report (ISFR). According to ISFR 2021, as compared to the mangrove cover estimated in year 2019, the country's mangrove cover increased by 17 sq. km. India government policies and regulations for restoration of mangroves both at central and state levels, have been the subject of numerous studies [350]. In addition to general environmental protection regulations, India has established organizations at the Central and State levels to explicitly address the consequences of climate change on Sundarban. Local saline-resistant crops need to be encouraged by the government. The need to establish flood relief centers and rapid action response teams for cyclones and storms is urgent. The state government has a separate department on Sundarban affairs and Sundarban Development Board for implementing different developmental activities in Sundarban area. Moreover, the Department of Disaster Management and Civil Defence, State, and National Disaster Management Authority

are performing several activities for disaster management in Sundarbans. The mangrove ecosystem is incredibly productive and offers millions of people a variety of goods and services. However, the quality of the mangrove environment is in a declining trend during the past few decades. Mangrove forests are the best carbon sinks and have the highest ecosystem service value(ESV). Due to severe natural and human-caused challenges, the total loss of ESVs in the Sundarban Biosphere Reserve (SBR) during the past 45 years has been assessed at 3310.79 million USD [41]. Unauthorized encroachment to the Sundarban forest is strictly prohibited through legislation and regulation to mitigate climate change disasters. It is essential to restore mangroves by data-driven planting, reforestation, and afforestation [42, 43]. Therefore, the suggested framework might be seen as a step in the right direction for restoration professionals, stakeholders, and researchers to safeguard mangroves holistically.

14.6 Summary

This study displays the objective and subjective measures for association rule filtering on species data that would aid in uncovering the most important and effective rules. This approach would be particularly beneficial because the subjective measure of the total taxonomic distinctness is domain-specific and aids in preserving the highest level of biodiversity richness.

Filtration would reduce the number of rules generated and only retain important rules. As a result, a useful collection of frequent associations between species is obtained, which conservation ecologists can utilize. This might be a practical method of enhancing biological richness via the maintenance of the ecosystem.

Part V

Conclusions

CHAPTER **15**

CONCLUSION AND FUTURE SCOPE

15.1	Conclusions	227
15.2	Future Scope	230

15.1 Conclusions

This section gives a concise summary of the findings demonstrating contributions made in this field of research. The contribution made in this thesis holds three essential aspects of the work done toward computational biodiversity. These are the performed case studies, proposed frameworks, and proposed algorithms for data mining as well as examples of their use in newly assembled datasets. Figure 15.1 is depicting the complete visualization of the contributions made in terms of case studies performed, proposed frameworks, and proposed algorithms along with the methodologies used and datasets studied. Below, the contributions are listed:

Contribution 1: Coined the term "Computational Biodiversity"

The phrase "Computational Biodiversity" may be used to describe the use of several computational approaches in biodiversity. We coined this phrase, conducted a thorough analysis of recent advancements in biodiversity protection, and highlighted the significance of collaborative efforts between ecologists and computer science experts. A detailed explanation has been provided in Chapter 2 and published in [56].

Contribution 2: A case study on Sundarban Mangrove: A littoral mangrove forest of West Bengal

A case study is performed for knowledge discovery of mangrove species in Sundarban using existing data mining methodology in addition to exploratory data analysis. The findings include the likelihood of occurrence for species to a degraded site for member introduction or restoration objectives, pointing the way for effective management via survey or resurvey. A detailed explanation is provided in Chapter 4 and this proposal is published in [58].

Contribution 3: Analysis and compilation of Indian estuarine data of flora and fauna

Here, we develop a set of guidelines that ecologists can offer in order to summarize closely occurring member lists, anticipated lists of sites for member expansion, etc. Therefore, this study would help to strengthen the estuarine diversity that might pave the way for region-based future research. This study has been published in [59] and described in Chapter 6.

Contribution 4: Data compilation and exploratory data analysis for knowledge discovery from species presence/ absence data

The occurrence data and taxonomical data of mangroves of both India and Sundarban have been curated and analyzed. The analysis is carried out using alpha and beta diversity indices. PCA analysis is done using Jaccard dissimilarity index on Indian estuarine mangroves. Shannon's H' , richness, and Pielou's evenness indices are studied on blockwise Sundarban mangrove data. A detailed explanation has been made in Chapter 5 and the corresponding article is published in [57].

Contribution 5: A case study on the application of tri-cluster on Indian statewide forest cover along time dimension

Here, we were able to successfully combine the tri-clustering concept with an informative structure, allowing us to track changes in the amount of forest cover and

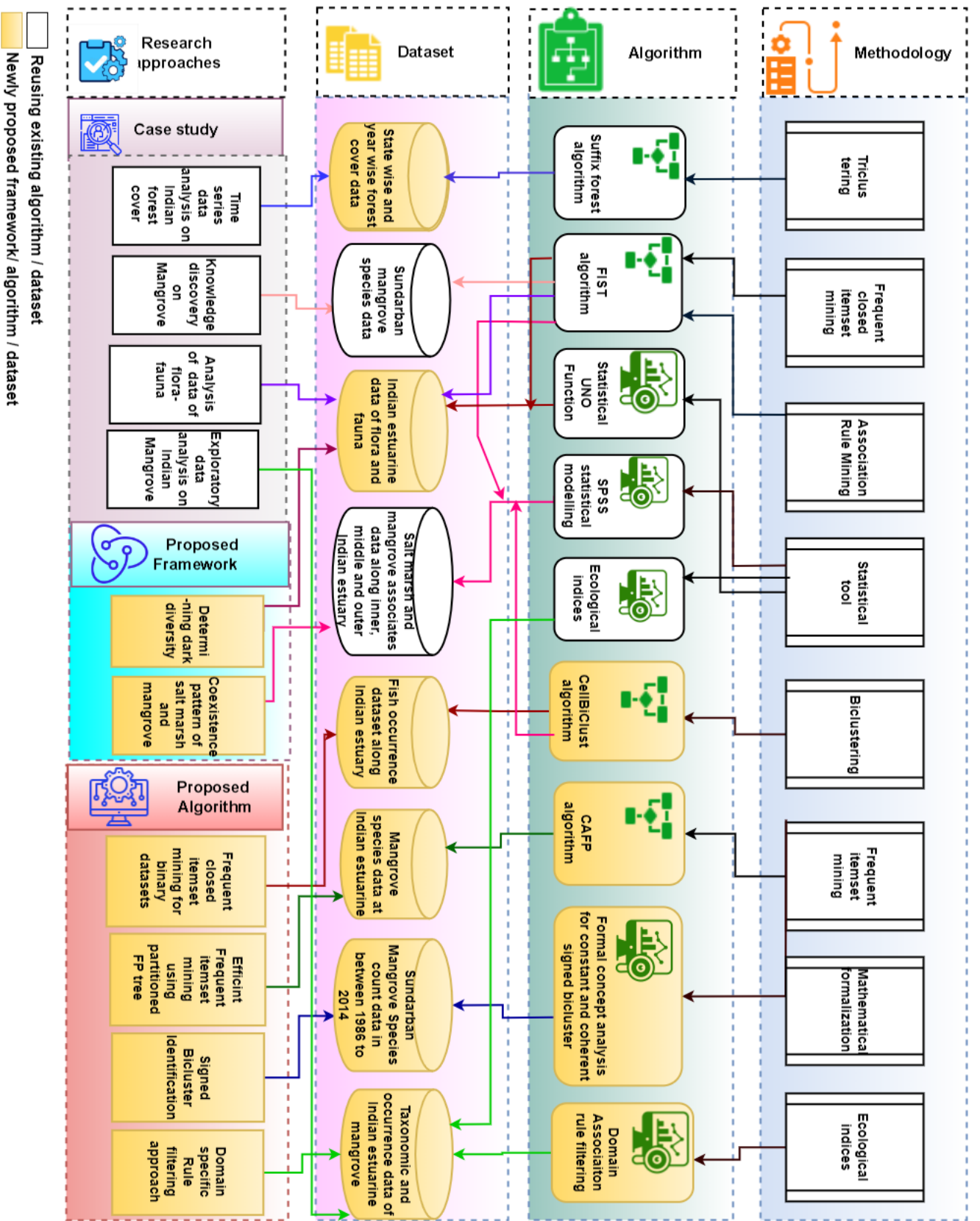


Figure 15.1: Complete visualization of the contribution made: in terms of case studies performed, proposed frameworks, and proposed algorithms along with the methodologies used and datasets studied

mangrove cover over time in various states and union territories in India. The detailed description is made in Chapter 7 and the related article is [60].

Contribution 6: Proposal of a combined approach of data mining and statistical analysis for computing dark diversity: A case study on Indian fauna

This study's premise was that, instead of relying solely on presence data, the probabilistic projection of member distribution, which includes the region of non-occurrence, can neutralize the loss of biodiversity by integrating potential ecosystems. Here we find a competent solution for biodiversity restoration by introducing the proposition of applying the dark diversity function to the presence-absence species dataset before the process of rule mining methodology. A detailed explanation has been made in Chapter 8 and the corresponding article is published in [62].

Contribution 7: Novel data mining framework for revealing coexistence pattern of mangroves, their associates, and salt marsh

This study aims to identify a unique restoration approach by evaluating the frequent co-existence status of salt marshes, with the mangroves, and mangrove associates in various zones of degraded mangrove patches for species-rich plantation. The knowledge gained from this study about the characteristics of co-occurring species and their possible interactions may be useful for both immediate and long-term restoration efforts. The development and long-term operation of salt marsh and mangrove ecosystems must thus be studied using a thorough, interdisciplinary methodology that considers both ecological and physical thresholds and bottlenecks. It has been published in [63] and described here in Chapter 9.

Contribution 8: Proposing novel algorithm for frequent itemset mining and its usage in rare mangrove identification

In this research, we propose a novel technique for frequent itemset mining that is efficient in terms of time and memory needs by combining various FP tree structures and cellular learning automata (CLA). Using publicly accessible actual and synthetic datasets made specifically for pattern mining algorithms, extensive experimentation has been carried out by comparing the performance of the proposed method with the top algorithms. The potential domain-specific use of the novel frequent itemset mining methodology in the study of species biodiversity data has also been discussed in addition to the proposal of the methodology. The detailed explanation is made in Chapter 11 and the corresponding article is published in [64].

Contribution 9: Implementation of a novel FP-Tree based biclustering algorithm and its extended usage in predicting species co-occurrence

We conceptualize, develop, and assess a new, effective algorithm called CellBi-Clust that uses the tree data structure and takes advantage of the idea of irregular cellular learning automata to add parallelism to the process of finding biclusters. We also demonstrate the usefulness of CellBiClust, which may be expanded to generate association rules. In turn, the biclusters and association rules produce predictions for fresh connections with specific probabilities. We demonstrate it on the species occurrence dataset in ecology, along with validation from the end of domain expertise. A detailed description is provided in Chapter 12 and the related article is published in [65].

Contribution 10: Proposal of novel constant and coherent signed bicluster in biodiversity study

Here, a novel signed biclustering methodology is proposed using formal concept analysis. The capacity to recognize both constant and coherent signed biclusters is a strength of the current study. With the aid of an analogous but smaller synthetic dataset, the method has been explained in Chapter 13, and the corresponding article is published in [66].

Contribution 11: Proposal of Domain-specific rule filtering technique: Applied on IUCN red-listed species

To get rid of a huge number of generated association rules, a domain-specific rule filtering methodology has been proposed that derives only statistically significant association rules. Then the rules are ranked depending on domain knowledge. The proposed rule filtering framework could be used to catalog and activate plantations with a wide variety of species in order to increase ecological sustainability. A detailed explanation has been made in Chapter 14 and the corresponding article is published in [57].

15.2 Future Scope

The future scopes are listed below:

1. We initially performed a comprehensive review of some recent initiatives toward computational biodiversity. The urgent need for computational techniques in biodiversity study has been experienced through this study. In the future, system modeling could be attempted using a computational framework for building holistic solutions for complex environmental and ecological issues, even incorporating big data. Henceforth, automation in a built-in model would assist ecologists in finding feasible solutions with minimum human intervention.
2. The only source of data for our study was different published and unpublished reports of field surveys. Obtaining real-time data from inaccessible areas will significantly aid in the compilation of datasets. Remote sensing techniques may be useful in acquiring more accurate data analysis in the future. In the future, we would like to focus on accumulating data from satellite imagery and employing a data-driven learning algorithm to extract more accurate facts.
3. Here, we have compiled multiple datasets manually, from online and offline resources. Automating the procedure for data integration can help streamline the process and reduce manual effort. It can save time and effort while ensuring consistency and accuracy. Our next aim would be to automate the procedure for data integration.
4. The aim of a case study in ecology is to investigate a specific ecological phenomenon or research question in a particular context. Case studies in ecology typically involve detailed examination and analysis of a specific ecosystem, species, population, or ecological process. They aim to deepen our understanding of ecological systems, contribute to scientific knowledge, and provide insights into ecological patterns, interactions, and

dynamics. The proposed frameworks could be incorporated in the future for the generation or restoration of biodiversity through various case studies.

5. The probable future research directions from algorithmic improvements could be: use the concept of cellular learning automata as employed in proposed methodology, can be used in finding
 - (1) Frequent itemsets over uncertain datastream
 - (2) Closed frequent itemsets for generating association rules for recommendation systems
 - (3) Frequent weighted itemsets mining
 - (4) top-k rank frequent itemsets
 - (5) frequent itemset mining on hadoop
6. It is found that data mining has huge potential to solve issues related to identifying forest loss, predicting potential sites for reforestation, predicting species behavior with respect to time, and other aspects. Although the prospect of data mining is not widely applied by the biodiversity domain researchers, we have tried to draw focus on this aspect of species biodiversity. Further study is required on genetic biodiversity and ecosystem biodiversity, as well. Along with data mining, for huge data, the deep learning-based methodology has a valuable prospect in application towards biodiversity.

REFERENCES

- [1] Thomas Elmqvist, Wayne C Zipperer, and Burak Güneralp. “Urbanization, habitat loss and biodiversity decline: Solution pathways to break the cycle”. In: *The Routledge handbook of urbanization and global environmental change*. Routledge, 2015, pp. 163–175.
- [2] Padma Sharma and Daizy R Batish. “Reasons of Biodiversity Loss in India”. In: *Biodiversity in India: Status, Issues and Challenges*. Springer, 2022, pp. 555–567.
- [3] Nicholas SG Williams et al. “A conceptual framework for predicting the effects of urban environments on floras”. In: *Journal of ecology* 97.1 (2009), pp. 4–9.
- [4] James EM Watson et al. “The exceptional value of intact forest ecosystems”. In: *Nature ecology & evolution* 2.4 (2018), pp. 599–610.
- [5] Tovah Siegel. “Evaluating the Impacts of Land Cover and Climate Change on Biodiversity at Local, Regional, and Global Scales”. PhD thesis. George Mason University, 2023.
- [6] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning*. John Wiley & Sons, 2003.
- [7] Rima Kumari et al. “Deforestation in India: Consequences and Sustainable Solutions”. In: *Forest Degradation Around the World* (2019).
- [8] Chandra Giri et al. “Monitoring mangrove forest dynamics of the Sundarbans in Bangladesh and India using multi-temporal satellite data from 1973 to 2000”. In: *Estuarine, coastal and shelf science* 73.1-2 (2007), pp. 91–100.
- [9] David J Curnick et al. “The value of small mangrove patches”. In: *Science* 363.6424 (2019), pp. 239–239.
- [10] Md Jamal Faruque et al. “Monitoring of land use and land cover changes by using remote sensing and GIS techniques at human-induced mangrove forests areas in Bangladesh”. In: *Remote Sensing Applications: Society and Environment* 25 (2022), p. 100699.
- [11] Abdul Aziz and Ashit Ranjan Paul. “Bangladesh Sundarbans: present status of the environment and biota”. In: *Diversity* 7.3 (2015), pp. 242–269.
- [12] Rajojit Chowdhury et al. “Effects of nutrient limitation, salinity increase, and associated stressors on mangrove forest cover, structure, and zonation across Indian Sundarbans”. In: *Hydrobiologia* 842.1 (2019), pp. 191–217.
- [13] Md Rahman et al. “Impact of increased salinity on the plant community of the Sundarbans Mangrove of Bangladesh”. In: *Community Ecology* 21.3 (2020), pp. 273–284.
- [14] Abhiroop Chowdhury et al. “Dynamics of salinity intrusion in the surface and ground water of Sundarban Biosphere Reserve, India”. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 944. IOP Publishing. 2021, p. 012061.

- [15] Nabonita Pal et al. "Impact of Aquatic Salinity on Mangrove Seedlings: A Case Study on *Heritiera fomes* (Common Name: Sundari)". In: *Biomedical Journal* (2017).
- [16] Banani Mandal, Arundhati Ganguly, and Arunava Mukherjee. "A Review for Understanding the Reasons of Vanishing Sundari Tree *Heritiera fomes* Buchanan-Hamilton from Sundarban Mangroves". In: *Environment and Ecology* 39.4 (2021), pp. 813–817.
- [17] Anirban Mukhopadhyay et al. "Aquatic salinization and mangrove species in a changing climate: impact in the Indian Sundarbans". In: *World Bank Policy Research Working Paper* 1.8532 (2018).
- [18] Kaberi Samanta and Sugata Hazra. "Mangrove forest cover changes in Indian Sundarban (1986–2012) using remote sensing and GIS". In: *Environment and Earth Observation*. Springer, 2017, pp. 97–108.
- [19] Laura J. Sonter, Saleem H. Ali, and James E. Watson. "Mining and biodiversity: Key issues and research needs in conservation science". In: *Proceedings of the Royal Society B: Biological Sciences* 285.1892 (2018). doi: 10.1098/rspb.2018.1926.
- [20] X Shan et al. "Species richness promotes ecosystem carbon storage: evidence from biodiversity-ecosystem functioning experiments". In: *Proceedings of the Royal Society* (2020). doi: 10.1098/rspb.2020.2063.
- [21] Susmita Dasgupta, Istiak Sobhan, and David Wheeler. "The impact of climate change and aquatic salinization on mangrove species in the Bangladesh Sundarbans". In: *Ambio* 46.6 (2017). doi: 10.1007/s13280-017-0911-0, pp. 680–694.
- [22] M. T. Naidu and O. A. Kumar. "Tree diversity, stand structure, and community composition of tropical forests in Eastern Ghats of Andhra Pradesh, India". In: *Journal of Asia-Pacific Biodiversity* 9.3 (2016), pp. 328–334.
- [23] FM Qamar et al. "Distribution and habitat mapping of key fauna species in selected areas of western Himalaya, Pakistan". In: *The Journal of Animal and Plant Sciences* 21.2 (2011), pp. 396–399.
- [24] S. Inthasone et al. "Biodiversity and Environment Data Mining". In: *Scientific Journal of National University of Laos* 9 (2015), pp. 116–128.
- [25] U. K. Sarkar et al. "Length weight relationship and condition factor of selected freshwater fish species found in River Ganga, Gomti and Rapti, India". In: *Journal of Environmental Biology* 34.5 (2013), p. 951.
- [26] R Sani et al. "Length–weight relationships of 14 Indian freshwater fish species from the Betwa (Yamuna River tributary) and Gomti (Ganga River tributary) rivers". In: *Journal of Applied Ichthyology* 26.3 (2010), pp. 456–459.
- [27] J. Pearce and S. Ferrier. "An evaluation of alternative algorithms for fitting species distribution models using logistic regression". In: *Ecological Modelling* 128.2 (2000), pp. 127–147.

REFERENCES

- [28] M. Ekström et al. “Logistic regression for clustered data from environmental monitoring programs”. In: *Ecological Informatics* 43 (2018), pp. 165–173.
- [29] H. Liao and W. Sun. “Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method”. In: *Procedia Environmental Sciences* 2 (2010), pp. 970–979.
- [30] F. A. Gougeon et al. “Automatic individual tree crown delineation using a valley-following algorithm and rule-based system”. In: *Proc. International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry, Victoria, British Columbia, Canada*. 1998, pp. 11–23.
- [31] L. A. E. Silva et al. “Applying data mining techniques for spatial distribution analysis of plant species co-occurrences”. In: *Expert Systems with Applications* 43 (2016), pp. 250–260.
- [32] Michel de A Silva et al. “Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon”. In: *International Conference on Advanced Data Mining and Applications*. Springer. 2013, pp. 384–395.
- [33] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. “Mining association rules between sets of items in large databases”. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. <http://doi.org/10.1145/170036.170072>. 1993, pp. 207–216.
- [34] Jiawei Han, Jian Pei, and Yiwen Yin. “Mining frequent patterns without candidate generation”. In: *ACM sigmod record* 29.2 (2000), pp. 1–12.
- [35] Mohammad Karim Sohrabi and Reza Roshani. “Frequent itemset mining using cellular learning automata”. In: *Computers in human behavior* 68 (2017), pp. 244–253.
- [36] Sara C Madeira and Arlindo L Oliveira. “Biclustering algorithms for biological data analysis: a survey”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 1.1 (2004), pp. 24–45.
- [37] Kartick Chandra Mondal. “Algorithms for data mining and bio-informatics”. PhD thesis. Université Nice Sophia Antipolis, 2013.
- [38] Evan M. Adams. “Using migration monitoring data to assess bird population status and behavior in a changing environment”. PhD thesis. The Graduate School, The University of Maine, 2014, p. 123. ISBN: 978-1-321-63451-8. URL: <https://www.proquest.com/dissertations-theses/using-migration-monitoring-data-assess-bird/docview/1664842393/se-2?accountid=16284>.
- [39] Rui Henriques and Sara C Madeira. “Triclustering algorithms for three-dimensional data analysis: a comprehensive survey”. In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–43.

- [40] Walaa K Mousa and Manish N Raizada. “Biodiversity of genes encoding anti-microbial traits within plant associated microbes”. In: *Frontiers in plant science* 6 (2015), p. 231.
- [41] Biswajit Bera et al. “Significant reduction of carbon stocks and changes of ecosystem service valuation of Indian Sundarban”. In: *Scientific Reports* 12.1 (2022), pp. 1–17.
- [42] Adam Irwansyah Fauzi et al. “Assessing potential climatic and human pressures in Indonesian coastal ecosystems using a spatial data-driven approach”. In: *ISPRS International Journal of Geo-Information* 10.11 (2021), p. 778.
- [43] Zhonghua Yu et al. “Mapping the Mangrove Forest Restoration Potential and Conservation Gaps in China Based on Random Forest Model”. In: *Journal of People Plants Environment (JPPE)* 25.4 (2022), pp. 425–446.
- [44] GA Thivakaran. “Mangrove restoration: an overview of coastal afforestation in India”. In: *Wetland Science: Perspectives From South Asia* (2017), pp. 501–512.
- [45] Aaron M Ellison. “Mangrove restoration: do we know enough?” In: *Restoration ecology* 8.3 (2000), pp. 219–229.
- [46] Thomas Worthington and Mark Spalding. “Mangrove restoration potential: A global map highlighting a critical opportunity”. In: (2018).
- [47] Ram Ranjan. “Optimal mangrove restoration through community engagement on coastal lands facing climatic risks: The case of Sundarbans region in India”. In: *Land Use Policy* 81 (2019), pp. 736–749.
- [48] Shan Wei et al. “Developing a grid-based association rules mining approach to quantify the impacts of urbanization on the spatial extent of mangroves in China”. In: *International Journal of Applied Earth Observation and Geoinformation* 102 (2021), p. 102431.
- [49] Daehyun Kim and Sewon Ohr. “Coexistence of plant species under harsh environmental conditions: an evaluation of niche differentiation and stochasticity along salt marsh creeks”. In: *Journal of Ecology and Environment* 44.1 (2020), pp. 1–16.
- [50] J. L. Harper et al. “The evolution and ecology of closely related species living in the same area”. In: *Evolution* 15.2 (1961), pp. 209–227.
- [51] Taylor E Shaw. “Species diversity in restoration plantings: Important factors for increasing the diversity of threatened tree species in the restoration of the Araucaria forest ecosystem”. In: *Plant diversity* 41.2 (2019). <https://doi.org/10.1016/j.pld.2018.08.002>, pp. 84–93.
- [52] Malay Pramanik et al. “Population health risks in multi-hazard environments: action needed in the Cyclone Amphan and COVID-19–hit Sundarbans region, India”. In: *Climate and Development* 14.2 (2022), pp. 99–104.
- [53] Malay Pramanik et al. “Climate Change-Livelihood-Migration Nexus: A Case Study from Sundarbans, India”. In: *EGU General Assembly Conference Abstracts*. 2021, EGU21–14256.

REFERENCES

- [54] AA Danda et al. *State of art report on biodiversity in Indian Sundarbans*. WWF-India, 2017.
- [55] L. K. Banerjee. “Biodiversity of the Sundarbans: Status and Challenges”. In: *Biological Diversity-Origin, Evolotion and Conservation*. Ed. by A. K. Sharma, D. Roy, and S. N. Ghosh. New Delhi: Viva Books, 2012, pp. 327–375. ISBN: 978-81-309-2113-6.
- [56] Moumita Ghosh and Kartick Chandra Mondal. “Computational Biodiversity”. In: *Proceedings of International Conference on Advanced Computing Applications: ICACA 2021*. Springer. 2022, pp. 739–750.
- [57] Moumita. Ghosh, Anirban. Roy, and Kartick. Chandra. Mondal. “Exploratory data analysis of Indian littoral forest: A novel approach for domain rule filtering”. In: xxx (2022).
- [58] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “Knowledge discovery of Sundarban Mangrove species: a way forward for managing species biodiversity”. In: *SN Computer Science* 3.1 (2022), pp. 1–14.
- [59] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “Analysis of Indian estuarine data of flora & fauna”. In: *Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 2*. Springer. 2022, pp. 393–410.
- [60] Kartick Chandra Mondal et al. “Introducing suffix forest for mining tri-clusters from time-series data”. In: *Innovations in Systems and Software Engineering* (2022), pp. 1–23.
- [61] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “Knowledge Discovery of Sundarban Mangrove Species: A Way Forward for Managing Species Biodiversity. In Press”. In: *Proceedings of Computational Intelligence in Communications and Business Analytics (CICBA-2021)*. 2021, pp. 1–15.
- [62] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “Determining dark diversity of different faunal groups in Indian estuarine ecosystem: a new approach with computational biodiversity”. In: *International Conference on Emerging Applications of Information Technology*. Springer. 2021, pp. 147–158.
- [63] Moumita Ghosh, Kartick Chandra Mondal, and Anirban Roy. “Recognition of co-existence pattern of salt marshes and mangroves for littoral forest restoration”. In: *Ecological Informatics* 71 (2022), p. 101769.
- [64] Moumita Ghosh et al. “Frequent itemset mining using FP-tree: a CLA-based approach and its extended application in biodiversity data”. In: *Innovations in Systems and Software Engineering* (2022), pp. 1–19.
- [65] Moumita. Ghosh et al. “CellBiClust: A Novel ICLA based Distributed FP-Tree Approach for Inclusion Maximal Biclusters: Submitted in journal”. In: xxx (2022).

- [66] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “FCA-Based Constant and Coherent-Signed Bicluster Identification and Its Application in Biodiversity Study”. In: *Proceedings of International Conference on Advanced Computing Applications: ICACA 2021*. Springer. 2022, pp. 679–691.
- [67] Moumita Ghosh et al. “Knowledge discovery in biodiversity:https://knowledgedb.ml”. In: *xxx* (2023).
- [68] R. S. Shrivastava. “Biodiversity of river Ganga (India): An Environmental Economist Perspective”. In: (2008).
- [69] H. Kabir et al. “Conservation of a river for biodiversity and ecosystem services: the case of the Halda—the unique river of Chittagong, Bangladesh”. In: *International Journal of River Basin Management* 13.3 (2015), pp. 333–342.
- [70] C. S. Reddy et al. “Threat evaluation for biodiversity conservation of forest ecosystems using geospatial techniques: a case study of Odisha, India”. In: 69 (2014), pp. 287–303.
- [71] R. Raghavan et al. “The conservation status of decapod crustaceans in the Western Ghats of India: an exceptional region of freshwater biodiversity”. In: *Aquatic Conservation: Marine and Freshwater Ecosystems* 25.2 (2015), pp. 259–275.
- [72] M. K. Gautam, R. K. Manhas, and A. K. Tripathi. “Patterns of diversity and regeneration in unmanaged moist deciduous forests in response to disturbance in Shiwalik Himalayas, India”. In: *Journal of Asia-Pacific Biodiversity* 9.2 (2016), pp. 144–151.
- [73] P. Boets et al. “Evaluation and comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the habitat suitability for alien macroinvertebrates”. In: *Environmental Modelling & Software* 74 (2015), pp. 92–103.
- [74] A Abdollahnejad, D Panagiotidis, P Surov, et al. “Investigation of a possibility of spatial modelling of tree diversity using environmental and data mining algorithms”. In: *Journal of FOR. SCI* 62.12 (2016), pp. 562–570.
- [75] R. Pouteau et al. “Support vector machines to map rare and endangered native plants in Pacific islands forests”. In: *Ecological Informatics* 9 (2012), pp. 37–46.
- [76] M. A. Acevedo et al. “Automated classification of bird and amphibian calls using machine learning: A comparison of methods”. In: *Ecological Informatics* 4.4 (2009), pp. 206–214.
- [77] Francisco Ronaldo Alves de Oliveira et al. “Tree legumes with fertilizer potential: a multivariate approach”. In: *Revista Ciência Agronômica* 52.1 (2021), pp. 1–10.
- [78] S. Gobeyn et al. “Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador”. In: *Environmental Modelling & Software* 92 (2017), pp. 269–316.
- [79] K. López-de Ipiña et al. “Automatic acoustic analysis for biodiversity preservation: A multi-environmental approach”. In: *Bioinspired Intelligence (IWOB), 2015 4th International Work Conference on*. IEEE. 2015, pp. 43–48.

REFERENCES

- [80] S. A. Siddiqui et al. “Deep Learning for Microalgae Classification”. In: (2017).
- [81] Janne Mäyrä et al. “Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks”. In: *Remote Sensing of Environment* 256 (2021), p. 112322.
- [82] W. P. Tsai, F. J. Chang, and E. E. Herricks. “Exploring the ecological response of fish to flow regime by soft computing techniques”. In: *Ecological Engineering* 87 (2016), pp. 9–19.
- [83] Kavitha Mandhir Sandhya et al. “Fish assemblage structure and spatial gradients of diversity in a large tropical reservoir, Panchet in the Ganges basin, India”. In: *Environmental Science and Pollution Research* 26.18 (2019), pp. 18804–18813.
- [84] M. Pfeifer et al. “BIOFRAG—a new database for analyzing BIODiversity responses to forest FRAGmentation”. In: 4.9 (2014), pp. 1524–1537.
- [85] S. Inthasone, N. Pasquier, et al. “The BioKET Biodiversity Data Warehouse: Data and Knowledge Integration and Extraction”. In: *International Symposium on Intelligent Data Analysis*. Springer. 2014, pp. 131–142.
- [86] D. Nemitz et al. “Mining the Himalayan uplands plant database for a conservation baseline using the public GMBA webportal”. In: *Protection of the Three Poles*. Springer, 2012, pp. 135–158.
- [87] T. V. Ramachandra and A. Suja. “Sahyadri: Western Qhats Biodiversity Information System <http://ces. Use, erne t. in/biodiversity>”. In: *Biodiversity in Indian Scenarios* (2006), p. 1.
- [88] P. Genovesi et al. “Alien mammals in Europe: updated numbers and trends, and assessment of the effects on biodiversity”. In: *Integrative zoology* 7.3 (2012), pp. 247–253.
- [89] Elena Baralis et al. “CAS-Mine: providing personalized services in context-aware applications by means of generalized rules”. In: *Knowledge and information systems* 28.2 (2011), pp. 283–310.
- [90] Sergio A Alvarez. “Chi-squared computation for association rules: preliminary results”. In: *Boston, MA: Boston College* 13 (2003).
- [91] Henry Oliver Lancaster and Eugene Seneta. “Chi-square distribution”. In: *Encyclopedia of biostatistics* 2 (2005).
- [92] K. C. Mondal et al. “A new approach for association rule mining and bi-clustering using formal concept analysis”. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2012, pp. 86–101.
- [93] Mohammed J Zaki. “Generating non-redundant association rules”. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2000, pp. 34–43.
- [94] Garrett Birkhoff. *Lattice theory*. Vol. 25. American Mathematical Soc., 1940.

- [95] Lizhuang Zhao and Mohammed J Zaki. “Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data”. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. New York, NY, USA: Association for Computing Machinery, 2005, pp. 694–705. doi: <https://doi.org/10.1145/1066157.1066236>.
- [96] Hamid Beigy and Mohammad Reza Meybodi. “A mathematical framework for cellular learning automata”. In: *Advances in Complex Systems* 7.03n04 (2004). <http://doi.org/10.1142/S0219525904000202>, pp. 295–319.
- [97] Mehdi Esnaashari and Mohammad Reza Meybodi. “Irregular cellular learning automata”. In: *IEEE transactions on cybernetics* 45.8 (2014). <http://doi.org/10.1109/TCYB.2014.2356591>, pp. 1622–1632.
- [98] Amir Hosein Fathy Navid and Amir Bagheri Aghababa. “Cellular learning automata and its applications”. In: *Emerging Applications of Cellular Automata* (2013), pp. 85–111. doi: [10.5772/52953](https://doi.org/10.5772/52953).
- [99] Sandeep Thakur et al. “Assessment of changes in land use, land cover, and land surface temperature in the mangrove forest of Sundarbans, northeast coast of India”. In: *Environment, Development and Sustainability* (2020). doi: [10.1007/s10668-020-00656-7](https://doi.org/10.1007/s10668-020-00656-7), pp. 1–27.
- [100] Andres Payo et al. “Projected changes in area of the Sundarban mangrove forest in Bangladesh due to SLR by 2100”. In: *Climatic Change* 139.2 (2016), pp. 279–291.
- [101] Michael Sievers et al. “Indian Sundarbans mangrove forest considered endangered under Red List of Ecosystems, but there is cause for optimism”. In: *Biological Conservation* 251 (2020). doi: [10.1016/j.biocon.2020.108751](https://doi.org/10.1016/j.biocon.2020.108751), p. 108751.
- [102] S. Singh, Z. A. Malik, and C. M. Sharma. “Tree species richness, diversity, and regeneration status in different oak (*Quercus* spp.) dominated forests of Garhwal Himalaya, India”. In: *Journal of Asia-Pacific Biodiversity* 9.3 (2016), pp. 293–300.
- [103] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. doi: [10.1007/3-540-31190-4](https://doi.org/10.1007/3-540-31190-4). Springer Science & Business Media, 2006.
- [104] Mohammad Asad Hussain et al. “Changes of the seasonal salinity distribution at the Sundarbans coast due to impact of climate change”. In: *4th International Conference on Water & Flood Management (ICWFM)*. 2013, pp. 637–648.
- [105] S Harun Rashid et al. “Undergrowth species diversity of Sundarban mangrove forest Bangladesh in relation to salinity”. In: *Ber. Inst. Landschafts-Pflanzenökologie Univ. Hohenheim* 17 (2008). doi: [10.3329/bjb.v40i2.9778](https://doi.org/10.3329/bjb.v40i2.9778), pp. 41–56.
- [106] Mark O Hill. “Diversity and evenness: a unifying notation and its consequences”. In: *Ecology* 54.2 (1973). doi: [10.2307/1934352](https://doi.org/10.2307/1934352), pp. 427–432.
- [107] Edward H Simpson. “Measurement of diversity”. In: *Nature* 163.4148 (1949), p. 688.

REFERENCES

- [108] Lesley D Clarke and Nola J Hannon. “The mangrove swamp and salt marsh communities of the Sydney district: III. Plant growth in relation to salinity and water-logging”. In: *The Journal of Ecology* (1970). doi: 10 . 2307 / 2258276, pp. 351–369.
- [109] Bushra Nishat. *Landscape narrative of the Sundarban: towards collaborative management by Bangladesh and India*. Tech. rep. The World Bank, 2019.
- [110] Kailash. Chandra, C. Raghunathan, and Swetapadma. Dash. *Current status of Estuarine Biodiversity in India: 1-575*. <http://doi.org/10.13140/RG.2.1.3781.2088>. the Director, Zool. Surv. India, Kolkata, 2018.
- [111] Georgina M Mace. “The role of taxonomy in species conservation”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359.1444 (2004), pp. 711–719.
- [112] Courtney Ann Shaw. “ITIS (The Integrated Taxonomic Information System)”. In: *Navigating the Shoals: Evolving User Services in Aquatic and Marine Science Libraries: Proceedings of the 29th Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC)*. Vol. 29. IAMSLIC. 2004, p. 17.
- [113] Mark J Costello et al. “Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases”. In: *PloS one* 8.1 (2013), e51629.
- [114] Tim Robertson et al. “The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet”. In: *PloS one* 9.8 (2014), e102623.
- [115] Jari Oksanen et al. “The vegan package”. In: *Community ecology package* 10.631-637 (2007), p. 719.
- [116] A Ghosh et al. “Floral diversity of mangroves and mangrove associated species in the India Sundarbans with special reference to distribution and abundance”. In: *Journal of the Indian Society of Coastal Agricultural Research* 21.1 (2003), pp. 53–58.
- [117] J John Sepkoski Jr. “Alpha, beta, or gamma: where does all the diversity go?” In: *Paleobiology* (1988), pp. 221–234.
- [118] Raimundo Real and Juan M Vargas. “The probabilistic basis of Jaccard’s index of similarity”. In: *Systematic biology* 45.3 (1996), pp. 380–385.
- [119] Leah L Bremer and Kathleen A Farley. “Does plantation forestry restore biodiversity or create green deserts? A synthesis of the effects of land-use transitions on plant species richness”. In: *Biodiversity and Conservation* 19.14 (2010), pp. 3893–3915.
- [120] Carlo Ricotta and Giancarlo Avena. “On the relationship between Pielou’s evenness and landscape dominance within the context of Hill’s diversity profiles”. In: *Ecological Indicators* 2.4 (2003), pp. 361–365.

- [121] Ian F Spellerberg and Peter J Fedor. “A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the ‘Shannon–Wiener’ Index”. In: *Global ecology and biogeography* 12.3 (2003), pp. 177–179.
- [122] BC Jha et al. “Estuarine fisheries management options and strategies”. In: *CIFRI Policy Papers* 3 (2008), pp. 1–23.
- [123] J.R.B. Alfred, A.K. Das, and A.K. Sanyal. *Ecosystems of India:1-410*. ENV15-Zool. Surv. India, Kolkata, 2001.
- [124] David M Nelson and Mark E Monaco. “National overview and evolution of NOAA’s estuarine living marine resources (ELMR) Program”. In: (2000). URL: <https://www.biodiversitylibrary.org/page/3817819>.
- [125] S. Dzeroski. “Environmental Applications of Data Mining”. In: *Lecture Notes of Knowledge Technologies, University of Trento* (2003).
- [126] I. S. Sitanggang et al. “APPLICATION OF CLASSIFICATION ALGORITHMS IN DATA MINING FOR HOTSPOTS OCCURRENCE PREDICTION IN RIAU PROVINCE INDONESIA.” In: *Journal of Theoretical & Applied Information Technology* 43.2 (2012).
- [127] S. E. Sesnie et al. “The multispectral separability of Costa Rican rainforest types with support vector machines and Random Forest decision trees”. In: *International Journal of Remote Sensing* 31.11 (2010), pp. 2885–2909.
- [128] K. Liu et al. “Monitoring mangrove forest changes using remote sensing and GIS data with decision-tree learning”. In: *Wetlands* 28.2 (2008), pp. 336–346.
- [129] E. Guirado et al. “Deep-Learning Convolutional Neural Networks for scattered shrub detection with Google Earth Imagery”. In: *arXiv preprint arXiv:1706.00917* (2017).
- [130] I. Heredia. “Large-Scale Plant Classification with Deep Neural Networks”. In: *Proceedings of the Computing Frontiers Conference*. ACM, 2017, pp. 259–262.
- [131] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. “An apriori-based algorithm for mining frequent substructures from graph data”. In: *European conference on principles of data mining and knowledge discovery*. Berlin, Heidelberg: Springer-Verlag, 2000, pp. 13–23. DOI: https://doi.org/10.1007/3-540-45372-5_2.
- [132] Kartick Chandra Mondal et al. “Prediction of protein interactions on HIV-1-human PPI data using a novel closure-based integrated approach”. In: *International Conference on Bioinformatics Models, Methods and Algorithms*. SciTePress, 2012, pp. 164–173.
- [133] Kartick Chandra Mondal and Nicolas Pasquier. “Galois closure based association rule mining from biological data”. In: *Biological Knowledge Discovery Handbook* (2013), pp. 761–802. DOI: [10.1002/9781118617151.ch35](https://doi.org/10.1002/9781118617151.ch35).

REFERENCES

- [134] Bradley J Cardinale et al. “Biodiversity loss and its impact on humanity”. In: *Nature* 486.7401 (2012), p. 59.
- [135] D. Roe, N. Seddon, and J. Elliott. “Biodiversity Loss is a Development Issue: A Rapid Review of Evidence”. In: *Issue Paper, International Institute for Environment and Development (IIED), London* 798 (2019), pp. 678–683.
- [136] A. E. Magurran. “Biological diversity”. In: *Current Biology* 15.4 (2005), R116–R118.
- [137] T. H. Booth et al. “BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies”. In: *Diversity and Distributions* 20.1 (2014), pp. 1–9.
- [138] H. Reiss et al. “Species distribution modelling of marine benthos: a North Sea case study”. In: *Marine Ecology Progress Series* 442 (2011), pp. 71–86.
- [139] S. Harrison and H. Cornell. “Toward a better understanding of the regional causes of local community richness”. In: *Ecology letters* 11.9 (2008), pp. 969–979.
- [140] M. Kamruzzaman et al. “Regeneration status of mangrove species under mature stands in the oligohaline zone of the Sundarbans, Bangladesh”. In: *Regional Studies in Marine Science* 16 (2017), pp. 15–20.
- [141] R. J. Dube. “Stand Structure and Diversity Analysis of Timber Tree Species at Lupane State University Paddock, Matabeleland North, Zimbabwe”. PhD thesis. Lupane State University, 2018.
- [142] S Sreelekshmi et al. “Mangrove species diversity, stand structure and zonation pattern in relation to environmental factors—A case study at Sundarban delta, east coast of India”. In: *Regional Studies in Marine Science* 35 (2020), p. 101111.
- [143] N. Myers et al. “Biodiversity hotspots for conservation priorities”. In: *Nature* 403.6772 (2000), p. 853.
- [144] F. E. Zachos and J. C. Habel. *Biodiversity hotspots: distribution and protection of conservation priority areas*. Springer Science & Business Media, 2011.
- [145] M. Zobel et al. “The formation of species pools: historical habitat abundance affects current local diversity”. In: *Global Ecology and Biogeography* 20.2 (2011), pp. 251–259.
- [146] R. J. Lewis, R. Szava-Kovats, and M. Pärtel. “Estimating dark diversity and species pools: an empirical assessment of two methods”. In: *Methods in Ecology and Evolution* 7.1 (2016), pp. 104–113.
- [147] F. De Bello et al. “Measuring size and composition of species pools: a comparison of dark diversity estimates”. In: *Ecology and Evolution* 6.12 (2016), pp. 4088–4101.
- [148] R. J. Lewis et al. “Applying the dark diversity concept to nature conservation”. In: *Conservation Biology* 31.1 (2017), pp. 40–47.
- [149] C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

- [150] J. J. Brown et al. “A novel method to predict dark diversity using unconstrained ordination analysis”. In: *Journal of Vegetation Science* (2019).
- [151] E. W. Beals. “Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data”. In: *Advances in ecological research*. Vol. 14. Elsevier, 1984, pp. 1–55.
- [152] R. H. okland. “Are ordination and constrained ordination alternative or complementary strategies in general ecological studies?” In: *Journal of Vegetation Science* 7.2 (1996), pp. 289–292.
- [153] P. Legendre and E. D. Gallagher. “Ecologically meaningful transformations for ordination of species data”. In: *Oecologia* 129.2 (2001), pp. 271–280.
- [154] Biswajit Mondal, Ashis Kumar Saha, and Anirban Roy. “Mapping mangroves using LISS-IV and Hyperion data in part of the Indian Sundarban”. In: *International Journal of Remote Sensing* 40.24 (2019), pp. 9380–9400.
- [155] Yaping Chen and Yong Ye. “Effects of salinity and nutrient addition on mangrove *Excoecaria agallocha*”. In: *PloS one* 9.4 (2014), e93337.
- [156] Jashimuddin Karim and Ansarul Karim. “Effect of salinity on the growth of some mangrove plants in Bangladesh”. In: *Towards the rational use of high salinity tolerant plants*. Springer, 1993, pp. 187–192.
- [157] BF Clough. “Growth and salt balance of the mangroves *Avicennia marina* (Forsk.) Vierh. and *Rhizophora stylosa* Griff. in relation to salinity”. In: *Functional Plant Biology* 11.5 (1984), pp. 419–430.
- [158] N Suárez and E Medina. “Salinity effect on plant growth and leaf demography of the mangrove, *Avicennia germinans* L.” In: *Trees* 19.6 (2005), pp. 722–728.
- [159] Kakoli Banerjee, Roberto Cazzolla Gatti, and Abhijit Mitra. “Climate change-induced salinity variation impacts on a stenoeious mangrove species in the Indian Sundarbans”. In: *Ambio* 46.4 (2017), pp. 492–499.
- [160] Beth A Polidoro et al. “The loss of species: mangrove extinction risk and geographic areas of global concern”. In: *PloS one* 5.4 (2010), e10095.
- [161] Mst Begam et al. “Native salt-tolerant grass species for habitat restoration, their acclimation and contribution to improving edaphic conditions: a study from a degraded mangrove in the Indian Sundarbans”. In: *Hydrobiologia* 803.1 (2017), pp. 373–387.
- [162] Abhijit Mitra et al. “Impact of salinity on mangroves”. In: *Jour. Coast. Env* 1.1 (2010).
- [163] Daniel A Friess et al. “Are all intertidal wetlands naturally created equal? Bottlenecks, thresholds and knowledge gaps to mangrove and saltmarsh ecosystems”. In: *Biological Reviews* 87.2 (2012), pp. 346–366.
- [164] Michael J Osland et al. “Winter climate change and coastal wetland foundation species: salt marshes vs. mangrove forests in the southeastern United States”. In: *Global change biology* 19.5 (2013), pp. 1482–1494.

REFERENCES

- [165] Daniel Alongi. *The energetics of mangrove forests*. Springer Science & Business Media, 2009.
- [166] John RL Allen. “Morphodynamics of Holocene salt marshes: a review sketch from the Atlantic and Southern North Sea coasts of Europe”. In: *Quaternary Science Reviews* 19.12 (2000), pp. 1155–1231.
- [167] Samuel C Snedaker and Enrique J Lahmann. “Mangrove understorey absence: a consequence of evolution?” In: *Journal of Tropical Ecology* 4.3 (1988), pp. 311–314.
- [168] Michael J Osland et al. “Ecosystem development after mangrove wetland creation: plant–soil change across a 20-year chronosequence”. In: *Ecosystems* 15.5 (2012), pp. 848–866.
- [169] NP Yensen et al. “Halophytes to manage oilfield salt water: disposal by irrigation/evapotranspiration and remediation of spills”. In: *Proceedings of the Sixth International Petroleum Environmental Conference. Environmental Issue and Solutions in Petroleum Exploration, Production and Refining, Houston, TX, November*. Vol. 1618. 1999.
- [170] Mirza Hasanuzzaman et al. “Potential use of halophytes to remediate saline soils”. In: *BioMed research international* 2014 (2014).
- [171] S Mbense et al. “Rapid colonization of degraded mangrove habitat by succulent salt marsh”. In: *South African journal of botany* 107 (2016), pp. 129–136.
- [172] Dinesh C Sharma. *How bio-restoration is helping revive degraded mangroves in Sunderbans*. 2019. URL: <https://www.downtoearth.org.in/news/wildlife-biodiversity/how-bio-restoration-is-helping-revive-degraded-mangroves-in-sunderbans-66782>.
- [173] Donna Marie Bilkovic et al. “The role of living shorelines as estuarine habitat conservation strategies”. In: *Coastal Management* 44.3 (2016), pp. 161–174.
- [174] Rachel K Gittman et al. “Marshes with and without sills protect estuarine shorelines from erosion better than bulkheads during a Category 1 hurricane”. In: *Ocean & Coastal Management* 102 (2014), pp. 94–102.
- [175] Philip W Stevens, Clay L Montague, and Kenneth J Sulak. “Fate of fish production in a seasonally flooded saltmarsh”. In: *Marine Ecology Progress Series* 327 (2006), pp. 267–277.
- [176] Karen L McKee, Donald R Cahoon, and Ilka C Feller. “Caribbean mangroves adjust to rising sea level through biotic controls on change in soil elevation”. In: *Global Ecology and Biogeography* 16.5 (2007), pp. 545–556.
- [177] Erik S Yando, Michael J Osland, and MW Hester. “Microspatial ecotone dynamics at a shifting range limit: plant–soil variation across salt marsh–mangrove interfaces”. In: *Oecologia* 187.1 (2018), pp. 319–331.

- [178] Daniel R Richards and Daniel A Friess. “Rates and drivers of mangrove deforestation in Southeast Asia, 2000–2012”. In: *Proceedings of the National Academy of Sciences* 113.2 (2016), pp. 344–349.
- [179] LK Banerjee. “Sundarbans biosphere reserve”. In: *Floristic diversity and conservation strategies in India* 5 (2002), pp. 2801–2829.
- [180] Aaron M Ellison, Elizabeth J Farnsworth, and Rachel E Merkt. “Origins of mangrove ecosystems and the mangrove biodiversity anomaly”. In: *Global Ecology and Biogeography* 8.2 (1999), pp. 95–115.
- [181] Roy R Lewis III, Benjamin M Brown, and Laura L Flynn. “Methods and criteria for successful mangrove forest rehabilitation”. In: *Coastal wetlands*. Elsevier, 2019, pp. 863–887.
- [182] Clint Cameron et al. “Community structure dynamics and carbon stock change of rehabilitated mangrove forests in Sulawesi, Indonesia”. In: *Ecological Applications* 29.1 (2019), e01810.
- [183] Aaron M Ellison, Alexander J Felson, and Daniel A Friess. “Mangrove rehabilitation and restoration as experimental adaptive management”. In: *Frontiers in Marine Science* 7 (2020), p. 327.
- [184] Jeffrey Chow. “Determinants of household fuelwood collection from mangrove plantations in coastal Bangladesh”. In: *Forest Policy and Economics* 96 (2018), pp. 83–92.
- [185] Naohiro Matsui et al. “Ten year evaluation of carbon stock in mangrove plantation reforested from an abandoned shrimp pond”. In: *Forests* 3.2 (2012), pp. 431–444.
- [186] D. Wodehouse and D. Skeffington. *Mangrove Action Project (2019). “CBEMR” Mangrove Restoration*. <https://mangroveactionproject.org/mangrove-restoration/>. Accessed: 04-10-2021. 2019.
- [187] Shing Yip Lee et al. “Better restoration policies are needed to conserve mangrove ecosystems”. In: *Nature ecology & evolution* 3.6 (2019), pp. 870–872.
- [188] Daniel O Suman. “Mangrove management: challenges and guidelines”. In: *Coastal Wetlands*. Elsevier, 2019, pp. 1055–1079.
- [189] Mohammad Main Uddin et al. “Ecological development of mangrove plantations in the Bangladesh Delta”. In: *Forest Ecology and Management* 517 (2022), p. 120269.
- [190] Samantha Lai et al. “The effects of urbanisation on coastal habitats and the potential for ecological engineering: a Singapore case study”. In: *Ocean & Coastal Management* 103 (2015), pp. 78–85.
- [191] Mariana Mayer-Pinto et al. “Building ‘blue’: an eco-engineering framework for fore-shore developments”. In: *Journal of Environmental Management* 189 (2017), pp. 109–114.

REFERENCES

- [192] Rebecca L Morris et al. "From grey to green: Efficacy of eco-engineering solutions for nature-based coastal defence". In: *Global change biology* 24.5 (2018), pp. 1827–1842.
- [193] Florent Taureau et al. "Mapping the mangrove forest canopy using spectral unmixing of very high spatial resolution satellite images". In: *Remote Sensing* 11.3 (2019), p. 367.
- [194] Kasturi Devi Kanniah et al. "Remote sensing to study mangrove fragmentation and its impacts on Leaf Area Index and gross primary productivity in the South of Peninsular Malaysia". In: *Remote Sensing* 13.8 (2021), p. 1427.
- [195] Jingjing Cao et al. "Object-based mangrove species classification using unmanned aerial vehicle hyperspectral images and digital surface models". In: *Remote Sensing* 10.1 (2018), p. 89.
- [196] Alejandro Navarro et al. "Mangrove and saltmarsh distribution mapping and land cover change assessment for South-eastern Australia from 1991 to 2015". In: *Remote Sensing* 13.8 (2021), p. 1450.
- [197] Hammad Gilani et al. "Evaluating mangrove conservation and sustainability through spatiotemporal (1990–2020) mangrove cover change analysis in Pakistan". In: *Estuarine, Coastal and Shelf Science* 249 (2021), pp. 107–128.
- [198] Jeffrey J Kelleway et al. "Review of the ecosystem service implications of mangrove encroachment into salt marshes". In: *Global Change Biology* 23.10 (2017), pp. 3967–3983.
- [199] Steven C Pennings and Mark D Bertness. "Salt marsh communities". In: *Marine community ecology* 11 (2001), pp. 289–316.
- [200] Manpreet Kaur and Shivani Kang. "Market Basket Analysis: Identify the changing trends of market data using association rule mining". In: *Procedia computer science* 85 (2016), pp. 78–85.
- [201] Moumita Ghosh et al. "Finding Prediction of Interaction Between SARS-CoV-2 and Human Protein: A Data-Driven Approach". In: *Journal of The Institution of Engineers (India): Series B* (2021), pp. 1–10.
- [202] Malay Kumar Pramanik. "Assessment the impact of sea level rise on mangrove dynamics of Ganges delta in India using remote sensing and GIS". In: *Assessment* 4.21 (2014).
- [203] AKM Nazrul Islam. "Germination Eco-Physiology and Plant Diversity in Halophytes of Sundarban Mangrove Forest in Bangladesh". In: *Halophytes for Food Security in Dry Lands*. Elsevier, 2016, pp. 277–289.
- [204] Blaine Hanson, Stephen R Grattan, and Allan Fulton. *Agricultural salinity and drainage*. University of California Irrigation Program, University of California, Davis, 1999.

- [205] Ran Zhao. “Theory and operation of capacitive deionization systems”. PhD thesis. Wageningen University, 2013. ISBN: ISBN 978-94-6173-639-0.
- [206] Mostafa A Shirazi and Larry Boersma. “A unifying quantitative analysis of soil texture”. In: *Soil Science Society of America Journal* 48.1 (1984), pp. 142–147.
- [207] Derek G Groenendyk et al. “Hydrologic-process-based soil texture classifications for improved visualization of landscape function”. In: *PloS one* 10.6 (2015), e0131299.
- [208] ESF Office of Communications. *Soil pH: What it Means*. Accessed: 04-10-2021. 2020. URL: <https://www.esf.edu/pubprog/brochure/soilph/soilph.htm>.
- [209] Michael AA Cox and Trevor F Cox. “Multidimensional scaling”. In: *Handbook of data visualization*. Springer, 2008, pp. 315–347.
- [210] Joseph B Kruskal. *Multidimensional scaling*. Sage, 1978. ISBN: 0-8039-0940-3.
- [211] Kenneth Sturrock and Jorge Rocha. “A multidimensional scaling stress evaluation table”. In: *Field methods* 12.1 (2000), pp. 49–60.
- [212] P Ragavan et al. “Critical notes on the identity and distribution of *Sonneratia griffithii* Kurz (Lythraceae) in India—a critically endangered mangrove species”. In: *Nordic Journal of Botany* 37.1 (2019), e02119.
- [213] MA Hoque et al. “Present status of salinity rise in Sundarbans area and its effect on Sundari (*Heritiera fomes*) species”. In: *Research Journal of Agriculture and Biological Sciences* 2.3 (2006), pp. 115–121.
- [214] Abhiroop Chowdhury, Pranabes Sanyal, and Subodh Kumar Maiti. “Dynamics of mangrove diversity influenced by climate change and consequent accelerated sea level rise at Indian Sundarbans”. In: *International Journal of Global Warming* 9.4 (2016), pp. 486–506.
- [215] Sauren Das, Monoranjan Ghose, Robert Spooner-Hart, et al. “Effects of salinity on photosynthesis, leaf anatomy, ion accumulation and photosynthetic nitrogen use efficiency in five Indian mangroves”. In: *Wetlands Ecology and Management* 15.4 (2007), pp. 347–357.
- [216] Swapan K Sarker et al. “Are we failing to protect threatened mangroves in the Sundarbans world heritage ecosystem?” In: *Scientific reports* 6.1 (2016), pp. 1–12.
- [217] Jacob Solomon Raju Aluri and Henry Jonathan Karyamsetty. “Reproductive ecology of *Aegialitis rotundifolia* Roxb., A crypto-viviparous mangrove plant species in Krishna mangrove forest, Andhra Pradesh”. In: *Transylvanian Review of Systematical and Ecological Research* 20.1 (2018), pp. 17–30.
- [218] PB Tomlinson. *the Botany of Mangroves Cambridge University Press London*. 1986.
- [219] Sanit Aksornkoae et al. “Plants in Mangroves, Chalongrat Co”. In: *Ltd., Thailand* 120 (1992).

REFERENCES

- [220] Erik S Yando et al. “Jump-starting coastal wetland restoration: a comparison of marsh and mangrove foundation species”. In: *Restoration Ecology* 27.5 (2019), pp. 1145–1154.
- [221] Nibedita Mukherjee et al. “An interdisciplinary framework to evaluate bioshield plantations: Insights from peninsular India”. In: *Acta Oecologica* 63 (2015), pp. 91–100.
- [222] Johanna N Tachas et al. “Eco-engineered mangroves provide complex but functionally divergent niches for estuarine species compared to natural mangroves”. In: *Ecological Engineering* 170 (2021), p. 106355.
- [223] John L Schnase, Judy Cushing, and James A Smith. “Biodiversity and ecosystem informatics”. In: *Journal of intelligent information systems* 29.1 (2007). <https://doi.org/10.1007/s10844-006-0027-7>, pp. 1–6.
- [224] Simon Barkow et al. “BicAT: a biclustering analysis toolbox”. In: *Bioinformatics* 22.10 (2006), pp. 1282–1283.
- [225] Joan E Ball-Damerow et al. “Research applications of primary biodiversity databases in the digital age”. In: *PloS one* 14.9 (2019), e0215794.
- [226] Vijay Barve. “Discovering and developing primary biodiversity data from social networking sites: A novel approach”. In: *Ecological Informatics* 24 (2014), pp. 194–199.
- [227] Christian König et al. “Biodiversity data integration—the significance of data resolution and domain”. In: *PLoS biology* 17.3 (2019), e3000183.
- [228] Alan Jovic, Karla Brkic, and Nikola Bogunovic. “An overview of free software tools for general data mining”. In: *2014 37th International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2014, pp. 1112–1117.
- [229] Moumita Ghosh, Anirban Roy, and Kartick Chandra Mondal. “Determining dark diversity of different faunal groups in indian estuarine ecosystem: a new approach with computational biodiversity”. In: *Advanced Techniques for IoT Applications: Proceedings of EAIT 2020*. Springer, 2022, pp. 147–158.
- [230] Shamila Nasreen et al. “Frequent pattern mining algorithms for finding associated frequent patterns for data streams: a survey”. In: *Procedia Computer Science* 37 (2014), pp. 109–116. DOI: 10.1016/j.procs.2014.08.019.
- [231] Jiawei Han et al. “Frequent pattern mining: current status and future directions”. In: *Data mining and knowledge discovery* 15.1 (2007). <https://doi.org/10.1007/s10618-006-0059-1>, pp. 55–86.
- [232] Yen-Liang Chen et al. “Market basket analysis in a multiple store environment”. In: *Decision support systems* 40.2 (2005), pp. 339–354. DOI: 10.1016/j.dss.2004.04.009.

- [233] Nicolas Pasquier et al. “Efficient mining of association rules using closed itemset lattices”. In: *Information systems* 24.1 (1999), pp. 25–46. DOI: 10 . 1016 / S0306 - 4379 (99) 00003 - 4.
- [234] Ya-Han Hu and Tzu-Wei Yeh. “Discovering valuable frequent patterns based on RFM analysis without customer identification information”. In: *Knowledge-Based Systems* 61 (2014), pp. 76–88. DOI: 10 . 1016 / j . knosys . 2014 . 02 . 009.
- [235] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. “Effective pattern discovery for text mining”. In: *IEEE transactions on knowledge and data engineering* 24.1 (2010), pp. 30–44. DOI: 10 . 1109 / TKDE . 2010 . 211.
- [236] Mohammad Sohrabi and Soodeh Akbari. “A comprehensive study on the effects of using data mining techniques to predict tie strength”. In: *Computers in Human Behavior* 60 (2016), pp. 534–541. DOI: 10 . 1016 / j . chb . 2016 . 02 . 092.
- [237] Stefan Naulaerts et al. “A primer to frequent itemset mining for bioinformatics”. In: *Briefings in bioinformatics* 16.2 (2015), pp. 216–231. DOI: 10 . 5061 / dryad . nr353.
- [238] Gowtham Atluri et al. “Association analysis techniques for bioinformatics problems”. In: *International Conference on Bioinformatics and Computational Biology*. Springer. 2009, pp. 1–13. DOI: 10 . 1038 / npre . 2008 . 2184 . 1.
- [239] Wang Xin gang. “A Summary of Research on Frequent Itemsets Mining Technology”. In: *Procedia Computer Science* 131.C (2018), pp. 841–846. DOI: 10 . 1016 / j . procs . 2018 . 04 . 276.
- [240] Ashok Savasere, Edward Robert Omiecinski, and Shamkant B Navathe. *An efficient algorithm for mining association rules in large databases*. Tech. rep. Georgia Institute of Technology, 1995. DOI: 10 . 1007 / 978 - 3 - 540 - 28651 - 6 _ 52.
- [241] Hannu Toivonen et al. “Sampling large databases for association rules”. In: *Vldb*. Vol. 96. 1996, pp. 134–145. DOI: 10 . 1145 / 1514894 . 1514927.
- [242] Jong Soo Park, Ming-Syan Chen, and Philip S Yu. “An effective hash-based algorithm for mining association rules”. In: *Acm sigmod record* 24.2 (1995), pp. 175–186. DOI: 10 . 1145 / 568271 . 223813.
- [243] Zhi-Chao Li, Pi-Lian He, and Ming Lei. “A high efficient AprioriTid algorithm for mining association rule”. In: *2005 International Conference on Machine Learning and Cybernetics*. Vol. 3. IEEE. 2005, pp. 1812–1815. DOI: 10 . 1109 / ICMLC . 2005 . 1527239.
- [244] Yiwu Xie et al. “The Optimization and Improvement of the Apriori Algorithm”. In: *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*. Vol. 2. IEEE. 2008, pp. 663–665. DOI: 10 . 1109 / IITA . Workshops . 2008 . 170.

REFERENCES

- [245] Sergey Brin, Rajeev Motwani, and Craig Silverstein. “Beyond market baskets: Generalizing association rules to correlations”. In: *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. 1997, pp. 265–276. DOI: 10.1145/253262.253327.
- [246] Mohammed J Zaki. “Efficient enumeration of frequent sequences”. In: *Proceedings of the seventh international conference on Information and knowledge management*. 1998, pp. 68–75. DOI: 10.1145/288627.288643.
- [247] Ramesh C Agarwal, Charu C Aggarwal, and VVV Prasad. “A tree projection algorithm for generation of frequent item sets”. In: *Journal of parallel and Distributed Computing* 61.3 (2001), pp. 350–371. DOI: 10.1006/jpdc.2000.1693.
- [248] Junqiang Liu et al. “Mining frequent item sets by opportunistic projection”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 229–238. DOI: 10.1145/775047.775081.
- [249] Guimei Liu et al. “On computing, storing and querying frequent patterns”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 607–612. DOI: 10.1007/s10618-006-0059-1.
- [250] ZhiHong Deng, ZhongHui Wang, and JiaJian Jiang. “A new algorithm for fast mining frequent itemsets using N-lists”. In: *Science China Information Sciences* 55.9 (2012), pp. 2008–2030.
- [251] Zhihong Deng and Zhonghui Wang. “A new fast vertical method for mining frequent patterns”. In: *International Journal of Computational Intelligence Systems* 3.6 (2010), pp. 733–744.
- [252] Zhi-Hong Deng. “DiffNodesets: An efficient structure for fast mining frequent itemsets”. In: *Applied Soft Computing* 41 (2016), pp. 214–223. DOI: 10.1016/j.asoc.2016.01.010.
- [253] Wei Song, Bingru Yang, and Zhangyan Xu. “Index-BitTableFI: An improved algorithm for mining frequent itemsets”. In: *Knowledge-Based Systems* 21.6 (2008), pp. 507–513. DOI: 10.1016/j.knsys.2008.03.011.
- [254] Mohammad Karim Sohrabi and Ahmad Abdollahzadeh Barforoush. “Parallel frequent itemset mining using systolic arrays”. In: *Knowledge-Based Systems* 37 (2013), pp. 462–471. DOI: 10.1016/j.knsys.2012.09.005.
- [255] Tahrira Hashem et al. “An efficient dynamic superset bit-vector approach for mining frequent closed itemsets and their lattice structure”. In: *Expert Systems with Applications* 67 (2017), pp. 252–271. DOI: 10.1016/j.eswa.2016.09.023.
- [256] Mohammed J Zaki. “Hierarchical parallel algorithms for association mining”. In: *Advances in Distributed and Parallel Knowledge Discovery* (2000), pp. 339–76. DOI: 10.1007/978-3-540-73499-4_27.

- [257] R Patel, SS Rana, and KR Pardasani. “Model for load balancing on processors in parallel mining of frequent itemsets”. In: *Am. J. Applied Sci* 2 (2005), pp. 926–931. doi: 10.3844/ajassp.2005.926.931.
- [258] Chun-Hong Huang and Yungho Leu. “Multi-level dataset decomposition for parallel frequent itemset mining on a cluster of personal computers”. In: *Cluster Computing* 22.2 (2019), pp. 2851–2863. doi: 10.1007/s10586-017-1609-6.
- [259] Gülistan Özdemir Özdoğan and Osman Abul. “Task-parallel fp-growth on cluster computers”. In: *Computer and Information Sciences*. Springer, 2011, pp. 383–388. doi: 10.1007/3-540-36175-8_47.
- [260] Bart Goethals and Mohammed J Zaki. “Advances in frequent itemset mining implementations: report on FIMI’03”. In: *Acm Sigkdd Explorations Newsletter* 6.1 (2004), pp. 109–117.
- [261] Gösta Grahne and Jianfei Zhu. “Fast algorithms for frequent itemset mining using fp-trees”. In: *IEEE transactions on knowledge and data engineering* 17.10 (2005), pp. 1347–1362.
- [262] Takeaki Uno, Masashi Kiyomi, Hiroki Arimura, et al. “LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets”. In: *Fimi*. Vol. 126. 2004, pp. 1–11.
- [263] Douglas Burdick et al. “MAFIA: A Performance Study of Mining Maximal Frequent Itemsets.” In: *FIMI*. Citeseer. 2003, pp. 1490–1504.
- [264] Xing Sun and Andrew B Nobel. “On the size and recovery of submatrices of ones in a random binary matrix”. In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2431–2453.
- [265] Alessandro Colantonio et al. “ABBA: Adaptive bicluster-based approach to impute missing values in binary matrices”. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. <https://doi.org/10.1145/1774088.1774304>. 2010, pp. 1026–1033.
- [266] Selim Mimaroglu and Kuniaki Uehara. “Bit sequences and biclustering of text documents”. In: *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. <https://doi.org/10.1109/ICDMW.2007.38>. IEEE. 2007, pp. 51–56.
- [267] A. Prelić et al. “A systematic comparison and evaluation of biclustering methods for gene expression data”. In: *Bioinformatics* 22.9 (2006), pp. 1122–1129.
- [268] R. Agrawal and R. Srikant. “Fast algorithms for mining association rules”. In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. <http://doi.org/10.1007/BF02948845>. 1994, pp. 487–499.
- [269] Mohammed J Zaki and Karam Gouda. “Fast vertical mining using diffsets”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 326–335.
- [270] Mohammed Javeed Zaki. “Scalable algorithms for association mining”. In: *IEEE transactions on knowledge and data engineering* 12.3 (2000). <https://doi.org/10.1109/69.846291>, pp. 372–390.

REFERENCES

- [271] Y. K. Woon, W. K. Ng, and E. P. Lim. “A support-ordered trie for fast frequent itemset discovery”. In: *IEEE Transactions on Knowledge and Data Engineering* 16.7 (2004). <https://doi.org/10.1109/TKDE.2004.1318569>, pp. 875–879.
- [272] Nicolas Pasquier et al. “Discovering frequent closed itemsets for association rules”. In: *International Conference on Database Theory*. https://doi.org/10.1007/3-540-49257-7_25. Springer. 1999, pp. 398–416.
- [273] Mohammed J Zaki, Ching-Jui Hsiao, et al. *CHARM: An efficient algorithm for closed association rule mining*. Tech. rep. <https://doi.org/10.1.1.17.2956>. Technical Report 99, 1999.
- [274] Gösta Grahne and Jianfei Zhu. “Efficiently using prefix-trees in mining frequent itemsets.” In: *FIMI*. Vol. 90. <https://doi.org/10.1.1.3.6241>. 2003, p. 65.
- [275] John L Schnase et al. “Information technology challenges of biodiversity and ecosystems informatics”. In: *Information systems* 28.4 (2003). <https://doi.org/10.1016/S0306-4379>, pp. 339–345.
- [276] Ganesh Chandra and R Sagar. “Fisheries in Sundarbans: problems and prospects”. In: *Available at SSRN 2084014* (2003). <http://doi.org/10.2139/ssrn.2084014>.
- [277] Ivan Valiela, Jennifer L Bowen, and Joanna K York. “Mangrove Forests: One of the World’s Threatened Major Tropical Environments: At least 35% of the area of mangrove forests has been lost in the past two decades, losses that exceed those for tropical rain forests and coral reefs, two other well-known threatened environments”. In: *Bioscience* 51.10 (2001), pp. 807–815.
- [278] Mark Huxham et al. “Mangroves and people: local ecosystem services in a changing climate”. In: *Mangrove Ecosystems: A Global Biogeographic Perspective*. Springer, 2017, pp. 245–274.
- [279] Md Monirul Islam, Helena Borgqvist, and Lalit Kumar. “Monitoring Mangrove forest landcover changes in the coastline of Bangladesh from 1976 to 2015”. In: *Geocarto International* 34.13 (2019), pp. 1458–1476.
- [280] Stephanie S Romañach et al. “Conservation and restoration of mangroves: Global status, perspectives, and prognosis”. In: *Ocean & Coastal Management* 154 (2018), pp. 72–82.
- [281] Manoj Kumer Ghosh, Lalit Kumar, and Chandan Roy. “Mapping long-term changes in mangrove species composition and distribution in the Sundarbans”. In: *Forests* 7.12 (2016), p. 305.
- [282] B Ganter. *R. Wille Formal Concept Analysis: Logical Foundations*. 1999.
- [283] Eva Kriegova et al. “A theoretical model of health management using data-driven decision-making: the future of precision medicine and health”. In: *Journal of translational medicine* 19.1 (2021), pp. 1–12.

- [284] Prem Kumar Singh, Cherukuri Aswani Kumar, and Abdullah Gani. “A comprehensive survey on formal concept analysis, its research trends and applications”. In: *International Journal of Applied Mathematics and Computer Science* 26.2 (2016), pp. 495–516.
- [285] Nyoman Juniarta, Miguel Couceiro, and Amedeo Napoli. “Order-preserving Biclustering Based on FCA and Pattern Structures”. In: *Complex Pattern Mining*. Springer, 2020, pp. 47–62.
- [286] John A Hartigan. “Direct clustering of a data matrix”. In: *Journal of the american statistical association* 67.337 (1972), pp. 123–129.
- [287] Yizong Cheng and George M Church. “Biclustering of expression data.” In: *Ismb*. Vol. 8. 2000, pp. 93–103.
- [288] Amos Tanay, Roded Sharan, and Ron Shamir. “Discovering statistically significant biclusters in gene expression data”. In: *Bioinformatics* 18.suppl_1 (2002), S136–S144.
- [289] Dmitry I Ignatov, Sergei O Kuznetsov, and Jonas Poelmans. “Concept-based biclustering for internet advertisement”. In: *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE. 2012, pp. 123–130.
- [290] DV Gnatyshak et al. “Analysing online social network data with biclustering and triclustering”. In: *Proceedings of the «Concept Discovery in Unstructured Data» conference*. Vol. 871. Citeseer. 2012, pp. 30–39.
- [291] Mehdi Kaytoue et al. “Biclustering meets triadic concept analysis”. In: *Annals of Mathematics and Artificial Intelligence* 70.1-2 (2014), pp. 55–79.
- [292] Nyoman Juniarta, Miguel Couceiro, and Amedeo Napoli. “Formal Concept Analysis for Identifying Biclusters with Coherent Sign Changes”. In: 2019.
- [293] Swapan Kumar Sarker et al. “1980s–2010s: The world’s largest mangrove ecosystem is becoming homogenous”. In: *Biological Conservation* 236 (2019), pp. 79–91.
- [294] Jacob B Socolar et al. “How should beta-diversity inform biodiversity conservation?” In: *Trends in ecology & evolution* 31.1 (2016), pp. 67–80.
- [295] NA Siddiqi. *Mangrove forestry in Bangladesh*. Institute of Forestry & Environmental Sciences, University of Chittagong, 2001.
- [296] KR Clarke and RM Warwick. “The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels”. In: *Marine Ecology Progress Series* 184 (1999), pp. 21–29.
- [297] K Robert Clarke and Richard M Warwick. “A taxonomic distinctness index and its statistical properties”. In: *Journal of applied ecology* 35.4 (1998), pp. 523–531.
- [298] Christopher J Humphries, Paul H Williams, and Richard I Vane-Wright. “Measuring biodiversity value for conservation”. In: *Annual review of ecology and systematics* 26.1 (1995), pp. 93–111.

REFERENCES

- [299] Neil Campbell et al. “Species richness, taxonomic diversity, and taxonomic distinctness of the deep-water demersal fish community on the Northeast Atlantic continental slope (ICES Subdivision VIa)”. In: *ICES Journal of Marine Science* 68.2 (2011), pp. 365–376.
- [300] Kritish De et al. “Relationship between species richness, taxonomic distinctness, functional diversity, and local contribution to β diversity and effects of habitat disturbance in the riparian spider community of the Ganga River, India”. In: *Ecological Processes* 12.1 (2023), pp. 1–13.
- [301] Stanislao Bevilacqua et al. “The use of taxonomic relationships among species in applied ecological research: Baseline, steps forward and future challenges”. In: *Austral Ecology* 46.6 (2021), pp. 950–964.
- [302] Stephanie N. Spohn and Vincent B. Young. “Chapter 32 - Gastrointestinal Microbial Ecology With Perspectives on Health and Disease”. In: *Physiology of the Gastrointestinal Tract (Sixth Edition)*. Ed. by Hamid M. Said. Sixth Edition. Academic Press, 2018, pp. 737–753. ISBN: 978-0-12-809954-4. DOI: <https://doi.org/10.1016/B978-0-12-809954-4.00032-3>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128099544000323>.
- [303] Liza Goldberg et al. “Global declines in human-driven mangrove loss”. In: *Global change biology* 26.10 (2020), pp. 5844–5855.
- [304] Abdul Malik, Rasmus Fensholt, and Ole Mertz. “Mangrove exploitation effects on biodiversity and ecosystem services”. In: *Biodiversity and Conservation* 24 (2015), pp. 3543–3557.
- [305] Fareha Hilaluddin et al. “Disturbance of mangrove forests causes alterations in estuarine phytoplankton community structure in Malaysian Matang mangrove forests”. In: *Marine environmental research* 158 (2020), p. 104935.
- [306] Ruhuddien Pandu Yudha et al. “Impact of logging on the biodiversity and composition of flora and fauna in the mangrove forests of Bintuni Bay, West Papua, Indonesia”. In: *Forest Ecology and Management* 488 (2021), p. 119038.
- [307] Nazlin Asari et al. “Mangrove as a Natural Barrier to Environmental Risks and Coastal Protection”. In: *Mangroves: Ecology, Biodiversity and Management* (2021), pp. 305–322.
- [308] Abhiroop Chowdhury et al. “Alternative green livelihood initiatives: an effective way to achieve the sustainable development goals at disaster vulnerable, Indian Sundarbans”. In: *IOP Conference Series: Earth and Environmental Science*. IOP Publishing. 1077 - 2022, p. 012007.
- [309] Thomas Worthington et al. “Mangrove restoration potential: a global map highlighting a critical opportunity”. In: *University of Cambridge* (2019).

- [310] Jorge Soberón and Townsend Peterson. “Biodiversity informatics: managing and applying primary biodiversity data”. In: *Philosophical Transactions of the Royal Society of London. Series B Biological Sciences* 359.1444 (2004), pp. 689–698.
- [311] Roy R Lewis. “Mangrove restoration-Costs and benefits of successful ecological restoration”. In: *Proceedings of the Mangrove Valuation Workshop, Universiti Sains Malaysia, Penang*. Vol. 4. 2001.
- [312] Catherine E Lovelock and Benjamin M Brown. “Land tenure considerations are key to successful mangrove restoration”. In: *Nature ecology & evolution* 3.8 (2019), pp. 1135–1135.
- [313] Babak Kamali and Roslan Hashim. “Mangrove restoration without planting”. In: *Ecological Engineering* 37.2 (2011), pp. 387–391.
- [314] Roslan Hashim et al. “An integrated approach to coastal rehabilitation: mangrove restoration in Sungai Haji Dorani, Malaysia”. In: *Estuarine, Coastal and Shelf Science* 86.1 (2010), pp. 118–124.
- [315] Claudia Teutli-Hernández et al. “Nurse species could facilitate the recruitment of mangrove seedlings after hydrological rehabilitation”. In: *Ecological Engineering* 130 (2019), pp. 263–270.
- [316] PER Dale, JM Knight, and PG Dwyer. “Mangrove rehabilitation: a review focusing on ecological and institutional issues”. In: *Wetlands ecology and management* 22 (2014), pp. 587–604.
- [317] Mehebab Sahana, Gopala Areendran, and Haroon Sajjad. “Assessment of suitable habitat of mangrove species for prioritizing restoration in coastal ecosystem of Sundarban Biosphere Reserve, India”. In: *Scientific Reports* 12.1 (2022), p. 20997.
- [318] Tien Dat Pham et al. “Remote sensing approaches for monitoring mangrove species, structure, and biomass: Opportunities and challenges”. In: *Remote Sensing* 11.3 (2019), p. 230.
- [319] Ben Brown. “Ecological rehabilitation in mangrove systems: the evolution of the practice and the need for strategic reform of policy and planning”. In: *Routledge Handbook of Ecological and Environmental Restoration*. Routledge, 2017, pp. 295–311.
- [320] Maria Elisa B Gerona-Daga and Severino G Salmo III. “A systematic review of mangrove restoration studies in Southeast Asia: Challenges and opportunities for the United Nation’s Decade on Ecosystem Restoration”. In: *Frontiers in Marine Science* (2022), p. 1865.
- [321] Jurgene H Primavera and Janalezza Morvenna A Esteban. “A review of mangrove rehabilitation in the Philippines: successes, failures and future prospects”. In: *Wetlands Ecology and Management* 16 (2008), pp. 345–358.
- [322] Fuguang Bao et al. “An Improved Evaluation Methodology for Mining Association Rules”. In: *Axioms* 11.1 (2022), p. 17.

REFERENCES

- [323] VB Arifanti. “Mangrove management and climate change: a review in Indonesia”. In: *IOP conference series: earth and environmental science*. Vol. 487. IOP Publishing. 2020, pp. 012–022.
- [324] Kodikara Arachchilage Sunanda Kodikara et al. “Have mangrove restoration projects worked? An in-depth study in Sri Lanka”. In: *Restoration Ecology* 25.5 (2017), pp. 705–716.
- [325] Abner Barnuevo et al. “Drawbacks of mangrove rehabilitation schemes: Lessons learned from the large-scale mangrove plantations”. In: *Estuarine, Coastal and Shelf Science* 198 (2017), pp. 432–437.
- [326] Betty May R Villamayor et al. “Impact of Haiyan on Philippine mangroves: Implications to the fate of the widespread monospecific *Rhizophora* plantations against strong typhoons”. In: *Ocean & Coastal Management* 132 (2016), pp. 1–14.
- [327] Severino G Salmo, Ian Tibbetts, and Norman C Duke. “Colonization and shift of mollusc assemblages as a restoration indicator in planted mangroves in the Philippines”. In: *Biodiversity and Conservation* 26 (2017), pp. 865–881.
- [328] Severino G Salmo, Ian R Tibbetts, and Norman C Duke. “Nekton communities as indicators of habitat functionality in Philippine mangrove plantations”. In: *Marine and Freshwater Research* 69.3 (2017), pp. 477–485.
- [329] Adriana Gracia et al. “Use of ecosystems in coastal erosion management”. In: *Ocean & coastal management* 156 (2018), pp. 277–289.
- [330] HP Rauch et al. “Ecological challenges* for the use of soil and water bioengineering techniques in river and coastal engineering projects”. In: *Ecological Engineering* 176 (2022), p. 106539.
- [331] Aji Ali Akbar et al. “The role of breakwaters on the rehabilitation of coastal and mangrove forests in West Kalimantan, Indonesia”. In: *Ocean & Coastal Management* 138 (2017), pp. 50–59.
- [332] Denny S Suripin and H Muhammad. “Mangrove Restoration with Environment Friendly Permeable Breakwater. Asian Jr. of Microbiol”. In: *Biotech. Env. Sc* 19.1 (2017), pp. 102–107.
- [333] Rosela Pérez-Ceballos et al. “Determining hydrological flow paths to enhance restoration in impaired mangrove wetlands”. In: *PloS one* 15.1 (2020), e0227665.
- [334] Jorge López-Portillo et al. “Mangrove forest restoration and rehabilitation”. In: *Mangrove Ecosystems: A Global Biogeographic Perspective: Structure, Function, and Services* (2017), pp. 301–345.
- [335] Tu Le Xuan et al. “Wave energy dissipation through a hollow triangle breakwater on the coastal Mekong Delta”. In: *Ocean Engineering* 245 (2022), p. 110419.
- [336] Aida Sartimbul et al. “Monitoring of Sedimentation on Geosynthetic Bags Installation Area in Banyuurip Mangrove Center, Ujung Pangkah, Gresik, Indonesia.” In: *Indonesian Journal of Marine Sciences/Ilmu Kelautan* 26.3 (2021).

- [337] Stephane Lallich, Olivier Teytaud, and Elie Prudhomme. “Association rule interestingness: Measure and statistical validation”. In: *Quality measures in data mining* (2007), pp. 251–275.
- [338] Jones Ranstam. “Multiple P-values and Bonferroni correction”. In: *Osteoarthritis and cartilage* 24.5 (2016), pp. 763–764.
- [339] Tal Galili. “dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering”. In: *Bioinformatics* 31.22 (2015), pp. 3718–3720.
- [340] P Ragavan et al. “A review of the mangrove floristics of India.” In: *Taiwania* 61.3 (2016).
- [341] Jyotiskona Barik et al. “Mangrove species distribution and water salinity: an indicator species approach to Sundarban”. In: *Journal of Coastal Conservation* 22.2 (2018), pp. 361–368.
- [342] Hema Joshi and M Ghose. “Forest structure and species distribution along soil salinity and pH gradient in mangrove swamps of the Sundarbans”. In: *Tropical Ecology* 44.2 (2003), pp. 195–204.
- [343] Kumudranjan Naskar, DN Guha Bakshi, et al. *Mangrove swamps of the Sundarbans*. Naya Prokash, 1987.
- [344] Neha Tusharbhai Patel and Amar Nath Pandey. “Salinity tolerance of *Aegiceras corniculatum* (L.) Blanco from Gujarat coasts of India”. In: *Anales de Biología*. 2009.
- [345] Asit Kumar Bhattacharjee et al. “Impact of salinity on above ground biomass and stored carbon in a common mangrove *Excoecaria agallocha* of Indian Sundarbans”. In: *The American journal of bio-pharmacology biochemistry and Life Sciences* 2 (2013), pp. 1–11.
- [346] B Clough. “Continuing the journey amongst mangroves ISME Mangrove Educational Book Series No 1 International Society for Mangrove Ecosystems (ISME) Okinawa Japan and International Tropical Timber Organization (ITTO) Yokohama Japan”. In: *Yokohama, Japan* (2013), pp. 370–373.
- [347] Shamim Ahmed et al. “Salinity reduces site quality and mangrove forest functions. From monitoring to understanding”. In: *Science of the Total Environment* 853 (2022), p. 158662.
- [348] Ta-Jen Chu et al. “Incorporating Species-Conditional Co-Occurrence When Selecting Indicator Species to Monitor Restoration after Mangrove Removal from the Siangshan Wetland, Taiwan”. In: *Journal of Marine Science and Engineering* 9.10 (2021), p. 1044.
- [349] Samyuktha Ashokkumar and Zareena Begum Irfan. *Current status of mangroves in India: benefits, rising threats policy and suggestions for the way forward*. Tech. rep. Madras School of Economics, 2018.

REFERENCES

- [350] Kanksha Mahadevia Ghimire and Mayank Vikas. “Climate change–impact on the Sundarbans, a case study”. In: *International Scientific Journal: Environmental Science* 2.1 (2012), pp. 7–15.
- [351] Abhijit Mitra, Sufia Zaman, and Prosenjit Pramanick. “Blue Economy and Livelihoods in Indian Sundarbans”. In: *Blue Economy in Indian Sundarbans: Exploring Livelihood Opportunities*. Springer, 2022, pp. 175–259.
- [352] PIB Delhi. *Conservation and protection of mangroves*. Tech. rep. Ministry of Environment, Forest & Climate Change, 2022.

Moumita Ghosh