Natural Resources, 2025, 16(13), 565-583
https://www.scirp.org/journal/nr
ISSN Online: 2158-7086
ISSN Print: 2158-706X

Extraction of Unique Plant Species Communities from the Sub-Humid Humid Bioclimate of Martinique

Jean-Emile Simphor, Jean-Philippe Claude, Philippe Joseph

UMR ESPACE DEV-BIORECA Laboratory, University of Antilles, Schœlcher, France
Email: jean-emile.symphor@univ-antilles.fr, claudejeanphilippe1@gmail.com

How to cite this paper: Simphor, J.-E., Claude, J.-P. and Joseph, P. (2025) Extraction of Unique Plant Species Communities from the Sub-Humid Humid Bioclimate of Martinique. *Natural Resources*, **16**, 565-583.
https://doi.org/10.4236/nr.2025.1613028

Received: January 15, 2025
Accepted: December 26, 2025
Published: December 29, 2025

Abstract

Biodiversity in forest ecosystems is crucial for regulating ecological processes and delivering essential ecosystem services. In this study, we investigate how specific plant species communities in secondary forest formations (FSS) reflect particular bioclimatic conditions and successional stages in a Sub-Humid Humid (SHH) environment. Our dataset comprises four survey stations (S51, S103, S119, S120) where minimal sampling areas were determined (ranging from 500 to 1000 m^2), yielding 29 to 45 recorded species per station. Environmental variables such as altitude, total biomass, and total basal area were also collected. Using the ECLAT algorithm to identify frequent species sets, followed by hierarchical clustering (CAH) and principal component analysis (PCA), we were able to highlight recurring species assemblages and singular, ecologically significant species. Further bivariate analyses of the distribution index (Id) against basal area (St_Totale) confirmed that the most frequent species often exhibit higher distribution and larger basal area. These findings underscore the potential of combining data mining techniques with conventional statistical methods to unravel complex patterns in forest ecosystem dynamics. Our results provide a valuable foundation for scaling up to more extensive datasets to better understand the ecological and environmental drivers shaping secondary forest communities under changing climate conditions.

Keywords

Biodiversity, Ecology, Item, Itemset, Frequent Itemset, Clustering

1. Introduction

The biological diversity of forest ecosystems plays a crucial role in regulating eco-

logical processes and providing essential ecosystem services. As forests evolve, their specific composition of plant species changes, reflecting adaptations to shifting bioclimatic conditions as well as complex biotic and abiotic interactions [1] [2]. Investigating unique plant species communities, whether very frequent or, conversely, very rare within specific bioclimates, and their association with different stages of succession can provide valuable insights into forest ecosystem dynamics in the face of environmental disturbances, particularly climate change.

In this article, our initial focus is on the bioclimatic characteristics and the stage of forest formation evolution. Specifically, we concentrate on secondary forest formations. Our objective is to demonstrate the viability of a methodological approach combining data mining techniques and statistical analyses to extract meaningful insights from a small dataset, based exclusively on station-level species surveys. By deliberately focusing on a limited number of stations (four), carefully selected to reflect significant ecological contrasts (altitude, biomass, management history), we propose a qualitative exploratory approach to evaluate the relevance of the methods employed before scaling them up to larger datasets.

This deliberate choice of a small number of stations is based on the need to test the robustness of the methodology in distinct ecological contexts, including *Swietenia macrophylla* plantations where vegetation regeneration is influenced by anthropic management, as well as "classic" stations with no known recent human intervention. This approach avoids biases linked to premature generalizations and lays the groundwork for future extrapolation. Thus, our central question is whether, using this limited dataset, it is possible to reveal unique plant species communities that reflect specific ecological conditions, while addressing the challenges posed by the complexity of forest ecosystems.

By validating this small-scale qualitative methodology, we hope to pave the way for its application on a larger scale, integrating richer and more diverse datasets to better understand ecological dynamics in tropical bioclimates.

2. Materials: Station Data from the Sub-Humid Humid Bioclimate for Secondary Sylvatic Stage Formations

We have four stations for plant species surveys in Sub-Humid Humid (SHH) bioclimate where physiognomic types correspond to Secondary Sylvatic Formations (SSF). These are stations S51, S103, S119, S120, referred to as such throughout. These four studied stations exhibit distinct ecological characteristics, reflecting significant variations in environmental conditions and management history. Stations S120 and S119 are classified as "classic," representing natural environments with no known recent or major human intervention. In contrast, stations S51 and S103 are Swietenia macrophylla (SWMAC) plantations, meaning their original vegetation was removed before being reconstituted under the canopy of these plantations.

In terms of altitude, S103 is the highest (130 m), followed by S120 (45 m), S119 (39 m), and finally S51, the lowest station (30 m). The station areas also vary: S103

is the largest (1000 m^2), while S119 is the smallest (500 m^2), with S120 and S51 having intermediate areas of 920 m^2 and 700 m^2, respectively.

For each station, plant species surveys were conducted on a surface corresponding to the minimal area. The minimal area refers to the smallest sampling surface beyond which adding additional surface area does not result in the appearance of new species (or very few). In other words, it is the smallest area that allows for an accurate representation of the floristic composition of an environment. The minimal area expressed in m^2 for stations S51, S103, S119, S120 is 700, 1000, 500, and 920 respectively.

Total biomass, an indicator of ecosystem density and productivity, is highest in station S119 (11.04) and lowest in S120 (3.31). The *Swietenia macrophylla* (SWMAC) plantations (S51 and S103) exhibit intermediate biomass levels (5.14 and 7.17, respectively). These ecological differences provide an opportunity to explore how plant community dynamics vary in response to the specific conditions of each station.

Stations S51, S103, S119, S120 contain 29, 45, 38, and 30 species respectively. We also have environmental information for each station. This includes altitude, total species biomass, total species density (m^2), station surface area, and forest type. Similarly, for each species in each station, we have the following variables: rf (relative frequency), Density, Di (Distribution index), BA_Total (Basal Area), DI (Dominance Index).

We emphasize that, within the framework of our qualitative methodological approach, only the species abundance matrix from the four stations serves as our initial dataset. All the aforementioned station characteristics were not known or used in the implementation of our approach.

Map of Martinique with Station Representation

3. Methods

In order to search for frequent and unique plant species communities in the SHH bioclimate at the SSF evolution stage, and based on the abundance matrix and environmental data table, we proceed in two main steps. We specify that our abundance matrix allows us to represent the presence and abundance of species for each of the species identified in the 4 stations of our study sample. Below we present an excerpt from our abundance matrix.

3.1. Descriptive Data Analysis

The descriptive analysis will allow us to summarize and characterize the data to provide an overall view. The objective is to describe and synthesize, using an observational approach, the main characteristics of our dataset.

3.2. Exploratory Data Analysis

With exploratory data analysis (EDA), we will search for relationships, structures, and trends in the data. The goal is to highlight patterns (trends, anomalies, groupings) that might not have been noticed with descriptive analysis alone. We aim to generate or validate hypotheses to better understand the underlying dynamics within the dataset.

3.2.1. Extraction of Frequent Plant Species Communities

1) Concept of Items, Itemsets, Frequent Itemsets, and Support

We rely on the concepts of itemsets and frequent itemsets, which are used in knowledge discovery in data or data mining. An itemset is a set of elements or items that appear together in a transaction or dataset. It represents a combination of features, attributes, or elements analyzed to uncover interesting patterns or associations within large databases. Itemsets are fundamental for analytical techniques such as association rules, where they help identify significant relationships and co-occurrences within the data.

Let $I = \{i1, i2, \ldots, im\}$ be a set of items. A transactional database D is a set of transactions, where each transaction T is a subset of items, *i.e.*, $T \subseteq I$.

An itemset X is any subset of I, such that $X \subseteq I$. The support of an itemset X in D is the proportion of transactions $T \in D$ for which $X \subseteq T$ [3] [4].

An itemset $X \subseteq I$ is considered frequent if it appears in a proportion of transactions exceeding a predefined threshold, known as the minimum support threshold, min_sup. Formally, if supp(X) denotes the support of the itemset X in the transactional database D, then X is frequent if $\text{supp}(X) \geq \text{min\_sup}$.

In our study, an item corresponds to a plant species, an itemset to a community of plant species, and a frequent itemset to a community of plant species that occurs together in a significant number of stations, exceeding the predefined support threshold. We hypothesize that frequent plant species communities can reveal groups of species that share preferences for certain ecological conditions or are similarly influenced by environmental disturbances.

2) Algorithms for Extracting Frequent Itemsets

In data mining, the first algorithm developed for extracting frequent itemsets is the Apriori algorithm developed by [3] [4]. This algorithm relies on the "anti-monotone" property, which states that if an itemset is not frequent, then all of its supersets cannot be frequent either. This significantly reduces the search space. Several other algorithms have been developed since, such as FP-growth [5], Fiasco [6]-[8].

Another notable algorithm is ECLAT, developed by [9] (Zaki *et al.*, 1997). ECLAT (Equivalence Class Clustering and bottom-up Lattice Traversal) is an efficient method used in data mining for finding frequent itemsets. Here are some key points about this algorithm:

- It employs a vertical data format, where each item is associated with a list of transaction IDs (TIDs) in which it appears, instead of a traditional horizontal layout of transactions.
- The algorithm performs a depth-first search (DFS) to explore the lattice of itemsets.
- By intersecting TID lists of items, ECLAT efficiently computes the support of itemsets without generating unnecessary candidates.
- It is well-suited for large and sparse datasets, providing better performance in such cases compared to Apriori.

In our ecological context, particularly with a large number of short-lived species communities and relatively high minimum support thresholds, the ECLAT algorithm is particularly well-suited. The input data in our case will consist of different sites with a list of the plant species identified for each site.

3.2.2. Hierarchical Clustering Using Agglomerative Classification

Hierarchical Ascendant Classification (HAC) [10] [11], is a clustering method used to classify objects (individuals, sites, variables) into homogeneous groups.

Initially, each object forms a separate cluster. At each step, the two most similar clusters are merged based on an agglomeration criterion (Ward, centroid, single linkage, etc.). The process continues until a single cluster containing all objects is formed. The results are visualized as a dendrogram, a tree-like diagram that helps determine the optimal number of groups. This technique is widely used in statistics, ecology, marketing, and any field requiring data segmentation.

1) Jaccard Distance Matrix

In this article, we have chosen to work with the Jaccard distance matrix, calculated from the Jaccard similarity index [11] [12] and [13]. The Jaccard similarity index is defined, for two stations A and B, as:

$$J(A,B)=\frac{\text{number of species present simultaneously in } A \text{ and } B}{\text{number of species present in at least } A \text{ or } B}$$

More formally, let:

- a: the number of species present in both stations;
- b: the number of species present in station A but absent in B;

- c: the number of species present in station B but absent in A.

The similarity index is then expressed as: $J(A,B)=\frac{a}{a+b+c}$.

The Jaccard distance is calculated as: $d_{\text{Jaccard}}(A,B)=1-J(A,B)$.

This approach highlights the species effectively shared between stations. The Jaccard index is an asymmetric index; the more species two stations share (*i.e.*, the larger a), the higher their similarity index and the lower their distance. Joint absences do not artificially inflate the similarity, as Jaccard focuses exclusively on actual presences. Thus, the similarity between two stations primarily reflects the number of shared species rather than the number of species absent in both.

4. Results

4.1. Species Data Abundance Matrix and Descriptive Analysis

We construct our abundance matrix, which includes 90 species columns for the four stations (S51, S103, S119, S120) representing the evolutionary stage of secondary sylvatic forest formations.

An excerpt of our initial abundance matrix is presented in Figure 1 below.

Description: df [4 × 90]

	ACRE <dbl>	AISP <dbl>	ANIN <dbl>	ANMU <dbl>	ARAL <dbl>	BAMU <dbl>	BOSU <dbl>	BRAL <dbl>	BUGL <dbl>
S103	1	3	2	0	0	2	4	2	0
S119	0	1	17	0	2	0	0	0	0
S120	0	0	44	2	0	0	0	9	0
S51	0	0	0	0	0	0	5	0	4

4 rows | 1–10 of 90 columns

Figure 1. Extract from the abundance matrix data.

For example, in the excerpt of this abundance matrix, we can observe the species ANIN (Andira inermis) in the third column, with 2 individuals in S103, 17 in S119, 44 in S120, and 0 in S51.

In terms of species richness per station, the results are represented in Figure 2:

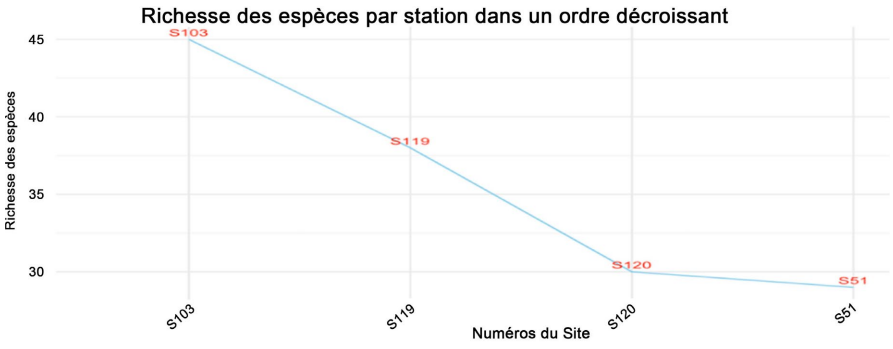
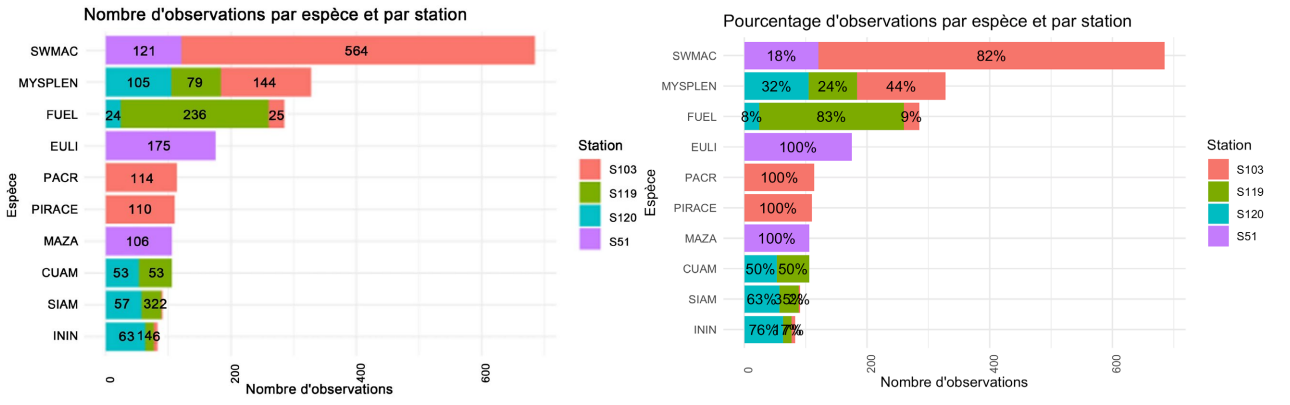


Figure 2. Richness per station.

We observe that station S103 has the highest number of species, with a total of 45, followed by S119 with 38. Stations S120 and S51 are less diverse, with 30 and

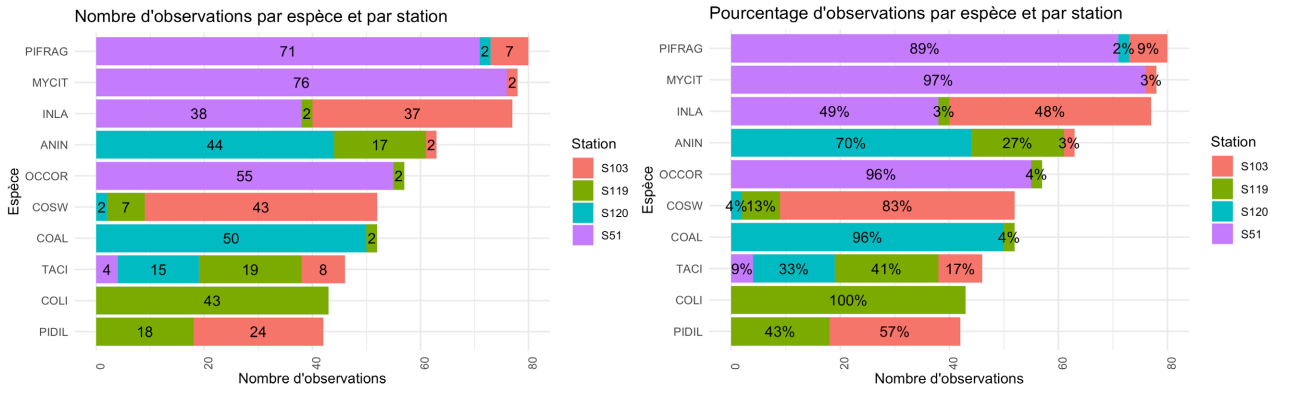
29 species, respectively. The stations S103, S119, S120, and S51 contain 1216, 666, 530, and 801 floristic units, respectively.

To enrich the information provided in the above figure with the species names and their counts per station, we use stacked bar charts to represent the most abundant species in the four stations. Specifically, in **Figure 3(a)** and **Figure 3(b)** below, we show the top 10 most abundant species by number of observations and by percentage (Top 10). In **Figure 4(a)** and **Figure 4(b)** below, we display the species ranked 11th to 20th by number of observations and percentage (Top 11-20). Charts for other less abundant species can be provided if needed.



(a) (b)

Figure 3. Top 10: Species abundance by station (a): Number, (b): Percentage.



(a) (b)

Figure 4. Top 11 to 20: Species abundance by station (a): Number, (b): Percentage.

We observe from the above charts that the species SWMAC (*Swietenia macrophylla*, *Meliaceae*, Tree) is the most abundant species in terms of observations, with a total count of 685, including 564 units in S103 and 121 units in S51. Notably, SWMAC is present in only two out of the four stations.

We also note that the next most abundant species is MYSPLEN (*Myrcia splendens*, *Myrtaceae*, Shrub), with a total count of 328, distributed as 144 units in S103, 79 units in S119, and 105 units in S120. MYSPLEN is found in three out of the

four stations. Similarly, the species FUEL (*Funtumia elastica*, *Apocynaceae*, Tree) has a total observation count of 285, with 25 units in S103, 236 in S119, and 24 in S120. This species is also present in three out of the four stations, the same three as MYSPLEN.

Additionally, the species TACI is present in all four stations, with counts of 4, 15, 19, and 8 units in stations S51, S120, S119, and S103, respectively.

Furthermore, we provide in Figure 5 below a boxplot representation of the species abundance distribution. For readability purposes, we have only included species found in at least two of the four stations. Species specific to a single station are therefore not represented.

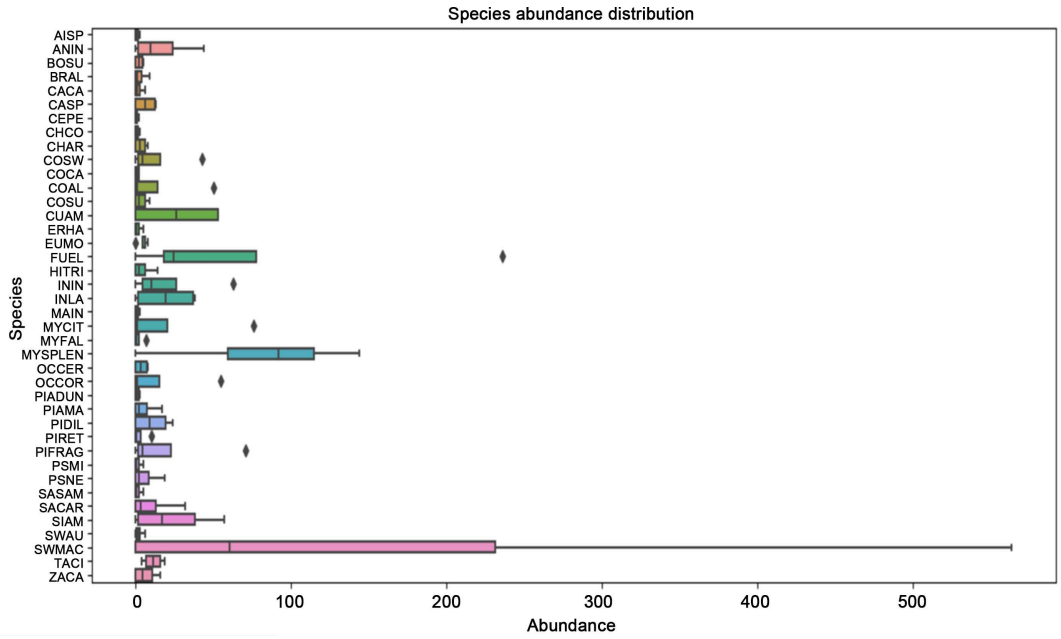


Figure 5. Boxplot: Abundance by species.

Just like in Figure 3, Figure 5 also highlights the dominance in terms of abundance of the species SWMAC, MYSPLEN, and FUEL. An additional piece of information provided by Figure 5 concerns the species FUEL, which shows an "outlier" in station S119 with 236 units. This indicates that for FUEL, which is highly abundant in the stations where it is present (S103, S119, S120), there is an overabundance in station S119 compared to stations S103 and S120.

Similarly, Figure 5 reveals the presence of other "outliers" concerning the species COSW (Coccoloba swartzii, Polygonaceae, Tree), COAL (Cordia alliodora, Boraginaceae, Tree), ININ (Inga ingoides, Mimosaceae, Tree), MYCIT (Myrcia citrifolia, Myrtaceae, Shrub), MYFAL (Myrcia fallax, Myrtaceae, Tree), OCCOR (Ocotea coriacea, Lauraceae, Tree), PIRET (Piper reticulatum, Piperaceae, Shrub), and PIFRAG (Pisonia fragans, Nyctaginaceae, Tree).

To provide an exhaustive view, Figure 6 below presents the total abundance by species for all species across the four stations. This table is sorted in descending

order of species abundance.

SWMAC	MYSPLEN	FUEL	EULI	PACR	PIRACE	CUAM	MAZA	SIAM	ININ	PIFRAG	MYCIT	INLA	ANIN	OCCOR	COSW	COAL
685	328	285	175	114	110	106	106	91	83	80	78	77	63	57	52	52
TACI	COLI	PIDIL	TRTR	SACAR	PSMA	CAIN	CASP	ZACA	PSNE	PIAMA	EUMO	HITRI	HYCOU	OCCER	CHAR	COSU
46	43	42	41	39	26	25	25	25	24	21	20	18	16	15	14	14
BRAL	PIRET	SIFO	BOSU	CADE	ODNY	SWAU	CACA	MYFAL	ERHA	EUPS	PSMI	SASAM	COCO	PIADUN	AISP	BUGL
11	11	11	9	9	9	9	8	8	6	6	6	6	5	5	4	4
CELA	CESP	CHAL	CHCO	CLHI	COBA	COCA	MAIN	MEB	OCPAT	OUGUIL	CEPE	COPU	LOHE	ANMU	ARAL	BAMU
4	4	4	4	4	4	4	4	4	4	4	3	3	3	2	2	2
CACY	CISP	HACA	HORAC	LOPU	MABI	PIPE	RIHUM	TAHET	ACRE	BUSI	CESC	COMO	EUAL	FICI	IXFER	MAAME
2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
MALAE	RAACU	SIOB	SWMAH	TEC												
1	1	1	1	1												

Figure 6. Total abundance by species.

We also present **Figure 7** below, which provides information on the absolute frequency of all species by station, *i.e.*, the number of stations where each species was observed. This table is sorted in descending order of the number of stations where the species were observed.

TACI	ANIN	COSW	EUMO	FUEL	ININ	INLA	MYSPLEN	PIFRAG	SIAM	SWAU	AISP	BOSU	BRAL	CACA	CASP	CEPE
4	3	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2
CHCO	CHAR	COCA	COAL	COSU	CUAM	ERHA	HITRI	MAIN	MYCIT	MYFAL	OCCER	OCCOR	PIADUN	PIAMA	PIDIL	PIRET
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
PSMI	PSNE	SASAM	SACAR	SWMAC	ZACA	ACRE	ANMU	ARAL	BAMU	BUGL	BUSI	CACY	CAIN	CADE	CESC	CELA
2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1
CESP	CHAL	CISP	CLHI	COPU	COBA	COLI	COMO	COCO	EUAL	EULI	EUPS	FICI	HACA	HORAC	HYCOU	IXFER
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
LOHE	LOPU	MAAME	MABI	MAZA	MALAE	MEB	OCPAT	ODNY	OUGUIL	PACR	PIPE	PIRACE	PSMA	RAACU	RIHUM	SIFO
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SIOB	SWMAH	TAHET	TEC	TRTR												
1	1	1	1	1												

Figure 7. Species frequency.

From **Figure 4** and **Figure 6**, we can note that there is only one species, TACI (Tabernaemontana citrifolia, Apocynaceae, Tree), which is present in all four stations. Following this, the species ANIN (Andira inermis, Fabaceae, Tree), COSW (Coccoloba swartzii, Polygonaceae, Tree), FUEL (Funtumia elastica, Apocynaceae, Tree), ININ (Inga ingoides, Mimosaceae, Tree), MYSPLEN (Myrcia splendens, Myrtaceae, Shrub), SIAM (Simarouba amara, Simaroubaceae, Tree), INLA (Inga laurina, Mimosaceae, Tree), and PIFRAG (Pisonia fragans, Nyctaginaceae, Tree) are present in three out of the four stations.

From **Figure 1**, **Figure 5** and **Figure 6**, we can also highlight species that are found in only one of the four stations. For example, ACRE (Acacia retusa) is found exclusively in station S103, and BUGL (Bunchosia glandulosa) is present only in station S51. These species will henceforth be referred to as station-specific species.

From this descriptive analysis of the data, two major characteristics emerge:

- Abundance or overabundance of certain species compared to others, as observed, for example, with the species SWMAC, MYSPLEN, and FUEL.
- Frequency of species occurrence across stations. For instance, the species TACI, though far less abundant than SWMAC, is present in all four stations, while SWMAC is found in only two. Moreover, several species are present in only one station.

4.2. Exploratory Analysis

With the descriptive analysis in the previous paragraph, we clearly highlighted the

characteristics of abundance, and even overabundance, of certain species across stations. In this paragraph, we aim to advance our analysis by identifying groupings or communities of distinctive species that are significant indicators of particular ecological situations and conditions.

We adopt a community-based approach, focusing on the specific composition of the species present before considering their abundance. This analysis builds upon the "frequent" species communities explained in paragraph 3 above.

Frequent and Specific Plant Species Communities

All the results in this paragraph presented below were obtained using the ECLAT algorithm.

1) Frequent plant species communities across the four stations

With a minimum support value of min_supp = 100%, only one species, TACI, is present in all four stations: S51 ∩ S103 ∩ S119 ∩ S120 = {TACI}.

The species TACI (*Tabernaemontana citrifolia*, *Apocynaceae*, Tree) from the Apocynaceae family is the only species present across all four stations of the secondary sylvatic forest formations. This is the sole species exhibiting this characteristic in our dataset.

In terms of its abundance, it was observed in stations S103, S119, S120, and S51 with counts of 8, 19, 15, and 4, respectively. This species is consistently found across the four secondary sylvatic formations.

2) Frequent plant species communities across three stations

With a minimum support value of min_supp = 75%, the following species communities are present in three out of the four stations:

- S103∩S119∩S120 = {ANIN (*Andira inermis*, Fabaceae, Tree), COSW (*Coccoloba swartzii*, Polygonaceae, Tree), EUMO (*Eugenia monticola*, Myrtaceae, Shrub), FUEL (*Funtumia elastica*, Apocynaceae, Tree), ININ (*Inga ingoides*, Mimosaceae, Tree), MYSPLEN (*Myrcia splendens*, Myrtaceae, Shrub), SIAM (*Simarouba amara*, Simaroubaceae, Tree), SWAU (*Swietenia aubrevilleana*, Meliaceae, Tree), TACI (*Tabernaemontana citrifolia*, Apocynaceae, Tree)}.

This community is present in stations S103, S119, and S120. It is referred to as maximal because it cannot be expanded further without reducing its support (3/4), which is the ratio of the number of stations containing the community (3) to the total number of stations (4).

- S51∩S103∩S119 = {TACI (*Tabernaemontana citrifolia*, Apocynaceae, Tree), INLA (*Inga laurina*, Mimosaceae, Tree)},
- S51∩S103∩S120 = {TACI *Tabernaemontana citrifolia*, Apocynaceae, Tree, PIFRAG (Pisonia *fragans*, Nyctaginaceae, Tree).

These two communities, namely {TACI, INLA} and {TACI, PIFRAG}, are also maximal itemsets.

3) Frequent plant species communities across two stations

Similarly, with a minimum support value of min_supp = 50%, we obtain:

- S51 ∩ S119 = {INLA, OCCOR, PIAMA, TACI},
- S51 ∩ S120 = {PIFRAG, TACI},

- S51 ∩ S103 = {BOSU, CHCO, CHAR, COCA, ERHA, INLA, MYCIT, PIFRAG, PSMI, PSNE, SWMAC, TACI},
- S103 ∩ S119 = {AISP, ANIN, COSW, EUMO, FUEL, ININ, INLA, MAIN, MYSPLEN, PIDIL, SIAM, SWAU, TACI},
- S103 ∩ S120 = {ANIN, BRAL, CACA, COSW, EUMO, FUEL, ININ, MYFAL, MYSPLEN, PIFRAG, SIAM, SWAU, TACI},
- S119 ∩ S120 = {ANIN, CASP, CEPE, COSW, COAL, COSU, CUAM, EUMO, FUEL, HITRI, ININ, MYSPLEN, OCCER, PIADUN, PIRET, SASAM, SACAR, SIAM, SWAU, TACI, ZACA}.

4) Plant species specific to each site

Finally, with min_supp = 25%, we obtain the species that are specific to the different stations. These are species found in only one of the four stations. For example, the species BUGL below, which is part of the set S51_SpecificSpecies, is a species found exclusively in station S51 and not in the other three stations, namely S103, S119, and S120.

- S51_SpecificSpecies = {BUGL, CACY, CAIN, COPU, COBA, COCO, EULI, MAZA, MEB, OUGUIL, RIHUM, SIFO, SIOB, TAHET, TRTR},
- S103_SpecificSpecies = {ACRE, BAMU, CADE, CHAL, EUAL, EUPS, FICI, HACA, IXFER, LOHE, LOPU, MABI, MALAE, ODNY, PACR, PIPE, PIRACE, PSMA, RAACU},
- S119_SpecificSpecies = {ARAL, CESC, CELA, CESP, CISP, COLI, COMO, MAAME, OCPAT, SWMAH, TEC},
- S120_SpecificSpecies = {ANMU, BUSI, CLHI, HORAC, HYCOU}.

4.3. Species Clustering

Using the different frequent species communities obtained in the previous paragraph with the ECLAT algorithm, we implement another approach based on clustering using Hierarchical Ascendant Classification (HAC) and Principal Component Analysis (PCA) to verify if the obtained communities are indicative of specific ecological conditions. We perform a Hierarchical Ascendant Classification (HAC) on our dataset. For readability purposes, species specific to only one station are not displayed in the dendrograms below. In **Figure 8(a)**, we present the dendrogram obtained using the Ward method, and in **Figure 8(b)**, we show the dendrogram obtained using the "complete linkage" method.

Below, we present the cross-tables that allow for a comparison of the clusters obtained using the Ward, Single, and Complete Linkage methods.

From the cross-tabulation tables in **Figure 9**, the three methods Ward, Single, and Complete produce identical results when the dendrograms are divided into 9 clusters. The cross-tabulation tables are diagonal, meaning the species clusters obtained are exactly the same across all three methods.

Thus, we opt for the selection of 9 clusters. The clusters obtained are represented on the factorial plane in **Figure 10**.

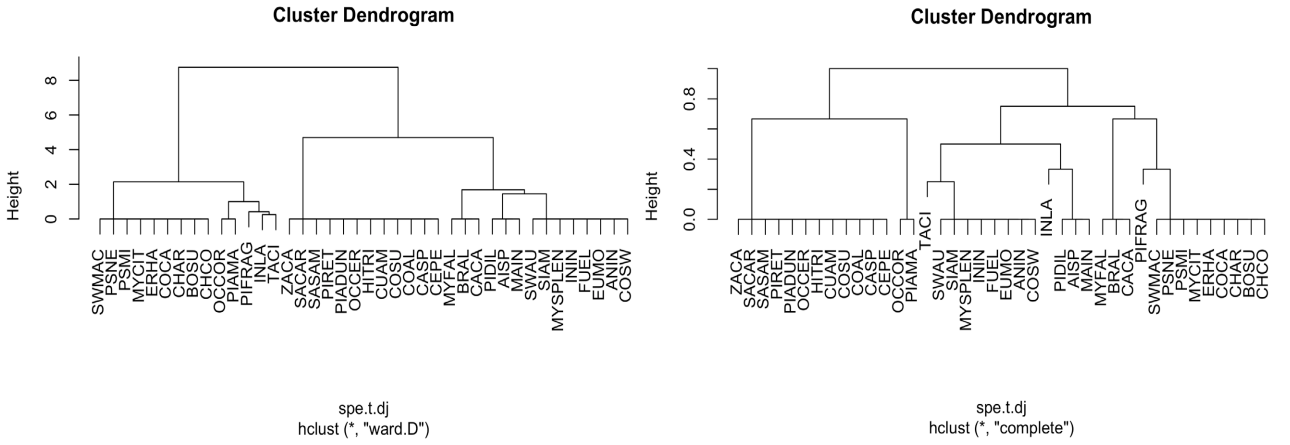


Figure 8. CAH_Jaccard_distance, (a): ward, (b): «complete» linkage.

spe.t.dj.ward

spe.t.dj.single	1	2	3	4	5	6	7	8	9
1	3	0	0	0	0	0	0	0	0
2	0	8	0	0	0	0	0	0	0
3	0	0	9	0	0	0	0	0	0
4	0	0	0	3	0	0	0	0	0
5	0	0	0	0	12	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	2	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	1

(a)

spe.t.dj.complete

spe.t.dj.single	1	2	3	4	5	6	7	8	9
1	3	0	0	0	0	0	0	0	0
2	0	8	0	0	0	0	0	0	0
3	0	0	9	0	0	0	0	0	0
4	0	0	0	3	0	0	0	0	0
5	0	0	0	0	12	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	2	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	1

(b)

Figure 9. cross-tables (a): ward & single linkage, (b): complete & single linkage.

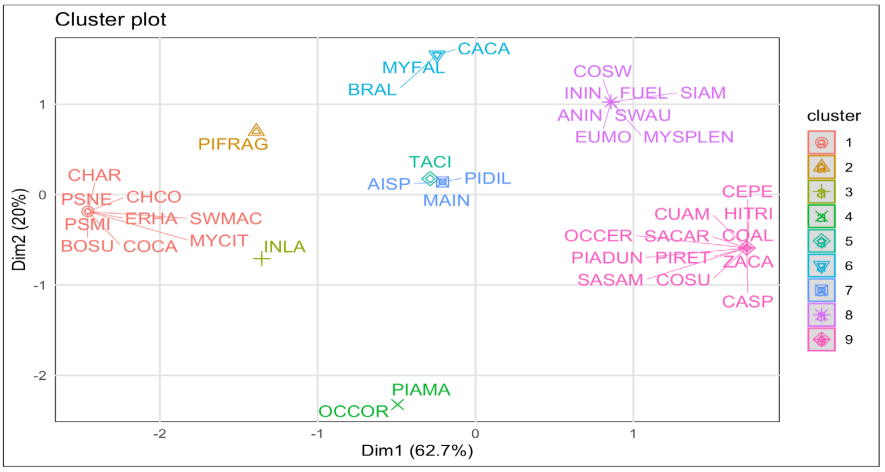


Figure 10. Clusters in the factorial plane.

The first factorial plane explains 62.7% for dimension 1% and 20% for dimension 2, accounting for a total of 82.7% of the variance in the abundance matrix.

This factorial plane is therefore highly significant.

4.4. Bivariate Analysis of Basal Area as a Function of the Species Distribution Index for Frequent Species Communities

Among the variables characterizing the species in the stations fr (relative frequency), Density, Id (Distribution Index), St_Totale (Basal Area), and ID (Dominance Index), we know that ID = Id * St_Totale and Id = fr * Density. Therefore, we select the two uncorrelated variables, Id and St_Totale, to perform a bivariate analysis. The objective is to analyze the positioning of the previously identified frequent species with respect to these two variables.

The relevance of analyzing the distribution index in relation to basal area was demonstrated in [14]. **Figure 11** below illustrates the case of station S51.

We use three colors for coding the species.

- Green: for Species found exclusively in station S51 and absent in all others. These species, marked in green on the chart, are specific to station S51 (Speci_S51).
- Blue: for Species present in two stations including S51. These are shown in blue Freq_2_St).
- Red: for Species found in three or four stations including S51. These are shown in Freq_3_4_St).

This color coding is consistent across **Figures 11-13**.

In **Figure 11(a)**, we observe that one species, SWMAC, has a significantly higher basal area and distribution index compared to other species. Statistically, this species is an outlier, as shown by the corresponding boxplots for basal area and distribution index, located at the top and right of the graph, respectively. In such a case, we alert the ecologist, indicating that it is necessary to provide an ecological explanation.

Regarding the species PIFRAG (Pisonia fragrans), INLA (Inga laurina), and TACI (Tabernaemontana citrifolia), marked in red on the figure and part of the community of frequent species present in three or four stations, the corresponding boxplots for both basal area and distribution index show that they have basal areas and distributions comparable to other species, except for the outliers. These species provide indications about the ecological development level of the station.

We note that all species are represented on the figure, but not all are labeled for readability purposes. This will also be the case for the three figures below. Similarly, for species referenced as Freq_2_St (excluding SWMAC), their distribution is equivalent to that of species referenced as Freq_3_4_St.

Concerning the species specific to station S51, in terms of the distribution index, two species behave as outliers: MAZA (Manilkara zapota) and TRTR (Triphasia trifolia). In terms of basal area, three species exhibit outlier behavior: EULI (Eugenia ligustrina), MAZA (Manilkara zapota), and TRTR (Triphasia trifolia).

To improve the readability of the graph, **Figure 11(b)** is presented as a modified version of **Figure 11(a)**, excluding the representation of species SWMAC and

EULI. It is observed that, in both distribution and basal area, species referenced as Freq_3_4_St, Freq_2_St, and Speci_S51 are distributed equivalently.

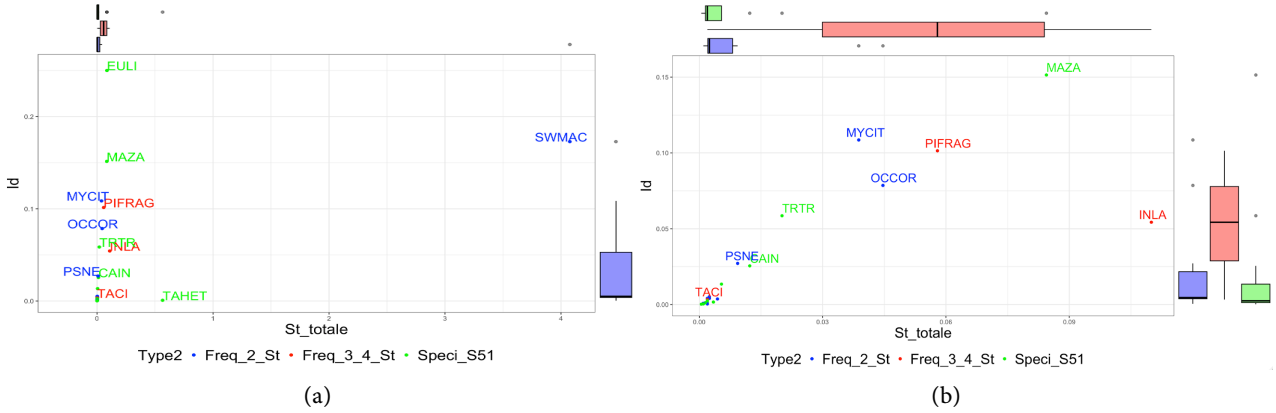


Figure 11. S51 Id = f(St), (a) with all species; (b) without swmac, euli, tahet.

In **Figure 12** below, we present the bivariate graph for station S103.

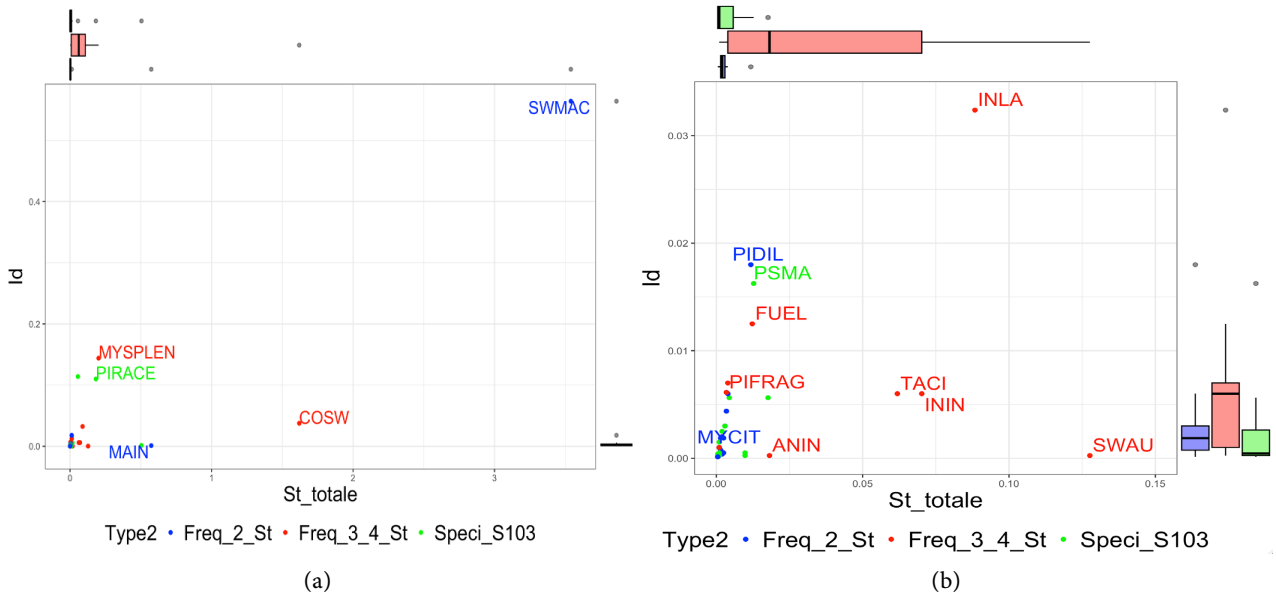


Figure 12. S103 Id = f(St), (a) with all species; (b) without swmac, cosw, mysplen, pirace, main.

Regarding station S103 and as shown in **Figure 12(a)**, several species, including SWMAC (Swietenia macrophylla), COSW (Coccoloba swartzii), MAIN (Mangifera indica), MYSPLEN (Myrcia splendens), and PIRACE (Pimenta racemosa), are positioned as "outliers." The other frequent species referenced as Freq_3_4_St in **Figure 12(b)**, namely SWAU (Swietenia aubrevilleana), INLA (Inga laurina), ININ (Inga ingoides), TACI (Tabernaemontana citrifolia), ANIN (Andira inermis), FUEL (Funtumia elastica), and PIFRAG (Pisonia fragrans), exhibit distributions and basal areas that are highly significant for this station.

In Figure 13 below, we present the bivariate graph for station S119.

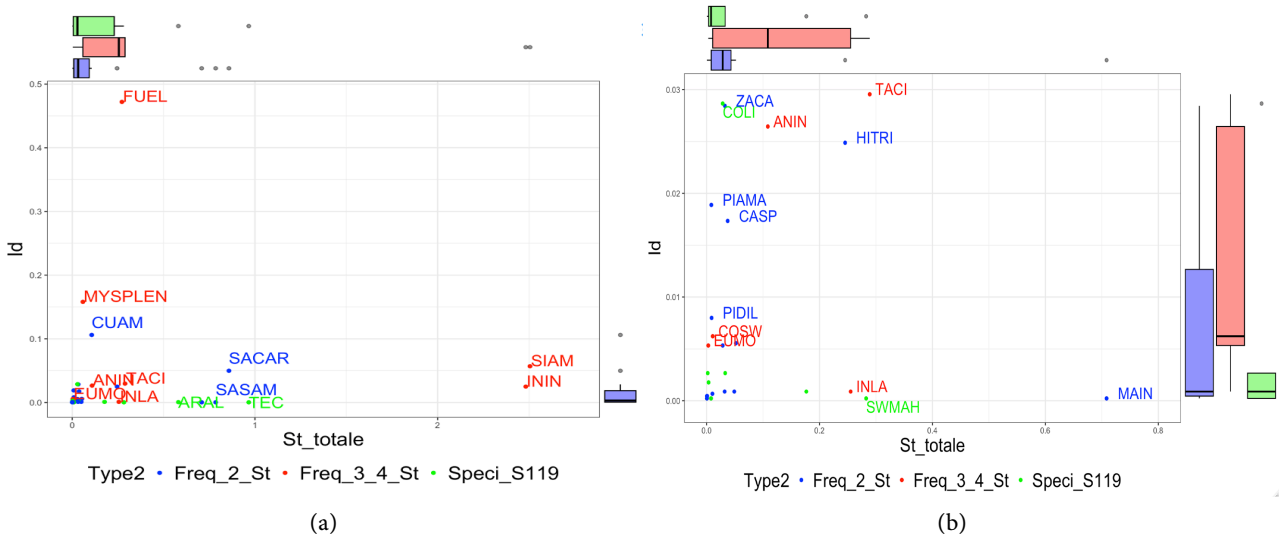


Figure 13. S119 Id = f(St), (a) with all species; (b) without fuel, siam, inin, sacar, sasam, tec, apal, mysplen, cuam.

Regarding station S119, as shown in Figure 13(a), for species referenced as Freq_3_4_St, the distribution index highlights FUEL (Funtumia elastica) and MYSPLEN (Myrcia splendens) as "outliers." These species are thus significantly distributed in this station. Similarly, for Freq_3_4_St species, but considering basal area, SIAM (Simarouba amara) and ININ (Inga ingoides) exhibit significantly higher basal areas than the others.

A similar observation applies to species referenced as Freq_2_St, with CUAM (Cupania americana) and SACAR (Samanea saman) as outliers for distribution, and SACAR (Sapium caribaeum), SASAM (Samanea saman), MAIN (Mangifera indica), and HITRI (Hirtella triandra) as outliers for basal area.

Regarding species specific to station S119 Speci_S119, TEC (Terminalia catappa) and ARAL (Artocarpus altilis) are the most significant in terms of basal area, while COLI (Coffea liberica) is the most significant in terms of the distribution index.

In Figure 14 below, we present the bivariate graph for station S120.

Station S120 exhibits a profile similar to that of station S119, in the sense that the species assemblages Freq_3_4_St and Freq_2_St are the most widely distributed and have the largest basal areas.

More specifically, regarding distribution for the Freq_3_4_St assemblage, species such as MYSPLEN (Myrcia splendens), ININ (Inga ingoides), FUEL (Funtumia elastica), TACI (Tabernaemontana citrifolia), EUMO (Eugenia monticola), SWAU (Swietenia aubrevilleana), and COSW (Coccoloba swartzii) are highly distributed in this station.

For the Freq_2_St assemblage, the most widely distributed species are CUAM (Cupania americana), COAL (Cordia alliodora), COSU (Cordia sulcata), and

SACCAR (Sapium caribaeum). Notably, COAL also exhibits a significantly higher basal area compared to other species.

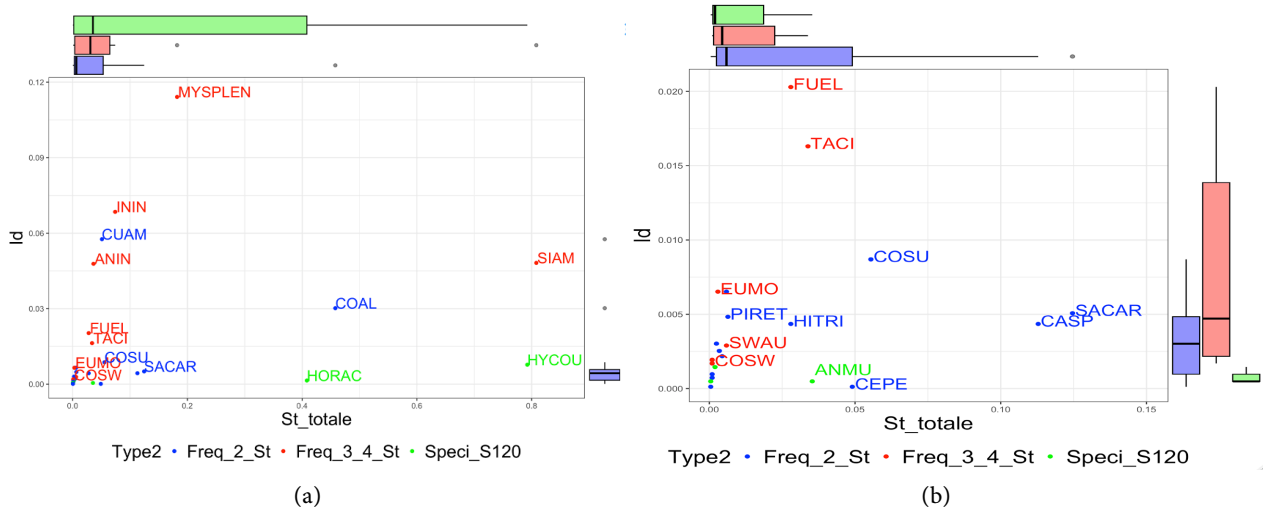


Figure 14. S120 Id = f(St), (a) with all species; (b) without mysplen, inin, cuam, anin, siam, coal, hycou, horac.

Regarding species specific to station S120 (Speci_S120), notable cases include HYCOU (Hymenaea courbaril), HORAC (Homalium racemosum), and ANMU (Annona muricata), which primarily show significant development in terms of basal area.

5. Discussion

Through this work, we aimed to demonstrate that by applying methods for knowledge discovery in data, data mining, and statistical analyses, it is possible to highlight unique species communities that are indicative of specific ecological situations and conditions. We selected a small dataset to test our knowledge extraction methodology, given the complexity of ecological ecosystems.

The most notable result concerns the identification of *Tabernaemontana citrifolia* (TACI), the only species common to all four stations, regardless of altitude, biomass, or management history. The presence of this species across all stations suggests its ability to adapt to diverse ecological conditions, ranging from the "classic" stations (S119 and S120) to the anthropized conditions of the *Swietenia macrophylla* plantations (S51 and S103). Based on the bivariate analysis from the previous section, this species is well distributed, with a significant basal area, particularly in the "classic" stations S119 and S120. This finding highlights *T. citrifolia* as a potentially key species in the dynamics of the studied secondary forest formations, likely playing a central role in ecological resilience.

The analysis using a support threshold of 75% identified groups of species present in three out of the four stations, indicating shared resilience despite notable ecological differences. For example, the community {ANIN, COSW, EUMO, FUEL, ININ, MYSPLEN, SIAM, SWAU, TACI} includes species widely distrib-

uted across stations S103, S119, and S120. Although these stations differ in altitude, biomass, and management history, they appear to share relatively homogeneous mesophilic conditions. This species assemblage includes trees (*Andira inermis*, *Simarouba amara*, and *Swietenia aubrevilleana*) as well as shrubs (*Eugenia monticola* and *Myrcia splendens*). This may reflect a balanced community dynamic between pioneer and mature species.

The species common to S51 and two other stations (S103 and either S119 or S120) also reveal interesting ecological patterns. For instance, the presence of *Inga laurina* (INLA) in S51, S103, and S119, as well as *Pisonia fragrans* (PIFRAG) in S51, S103, and S120, indicates that these species are capable of tolerating the specific ecological conditions of the *Swietenia macrophylla* plantations. This suggests that, despite the management history of these stations, certain species are able to persist under conditions of high anthropic influence.

The analysis using a 50% support threshold revealed species associations specific to station pairs. These results enhance our understanding of ecological gradients. For example, the assemblage shared by S103 and S120 {ANIN, BRAL, CACA, COSW, EUMO, FUEL, ININ, MYFAL, MYSPLEN, PIFRAG, SIAM, SWAU, TACI} includes a mix of pioneer and mature species, which aligns with the intermediate biomass levels of these stations and their relatively higher altitude. In contrast, the species associations specific to S51 and S119, such as {INLA, OCCOR, PIAMA, TACI}, likely reflect ecological responses specific to the anthropized conditions of S51 and the advanced successional characteristics of S119.

The species specific to each station provides additional insights into local ecological characteristics. For example, Bougainvillea glabra (BUGL) and Cecropia schreberiana (CESC), specific to S51 and S119, respectively, reflect the pronounced differences between a plantation station and a "classic" station at a more advanced successional stage. These indicator species could be useful in future studies to refine the ecological diagnostics of these stations.

Aside from Tabernaemontana citrifolia (TACI), the species assemblages identified in the Swietenia macrophylla plantations (S51 and S103) exhibit distinct compositions compared to the "classic" stations (S119 and S120). Specifically, S51 ∩ S103 = {BOSU, CHCO, CHAR, COCA, ERHA, INLA, MYCIT, PIFRAG, PSMI, PSNE, SWMAC, TACI}, while S119 ∩ S120 = {ANIN, CASP, CEPE, COSW, COAL, COSU, CUAM, EUMO, FUEL, HITRI, ININ, MYSPLEN, OCCER, PIADUN, PIRET, SASAM, SACAR, SIAM, SWAU, TACI, ZACA}. These distinctions likely reflect the influence of management history on plant community dynamics.

Additionally, station S103, which has the highest altitude (130 m) and the largest surface area (1000 m^2), harbors species assemblages (S103_EspecesSpecifiques = {ACRE, BAMU, CADE, CHAL, EUAL, EUPS, FICI, HACA, IXFER, LOHE, LOPU, MABI, MALAE, ODNY, PACR, PIPE, PIRACE, PSMA, RAACU}) that include species commonly associated with cooler, mesophilic, and better-preserved conditions. These results highlight the role of SWMAC plantations in shap-

ing community structures while also revealing interactions between post-plantation regeneration and abiotic factors such as altitude.

The bivariate analyses conducted on the Distribution Index (Id) and total basal area (St_Totale) revealed that the most widespread species exhibit some of the highest Id values and, in certain cases, also a high basal area. This may indicate a dominant position within the community. More broadly, these results suggest that a species combining high frequency with substantial coverage plays a significant role in ecological processes such as competition, resource availability, and canopy architecture, potentially influencing successional dynamics.

6. Conclusions

This study highlights the relevance of an approach for identifying frequent plant species communities in the sub-humid humid bioclimate and mesophilic forests. The ECLAT algorithm, originally designed for analyzing transactional data in supermarket sales, successfully identified frequent and specific plant species assemblages while avoiding biases related to overabundance caused by extensive anthropization. The results were obtained without incorporating prior knowledge of the environmental or historical characteristics of the stations.

Robust co-occurrence patterns were identified, accurately reflecting the ecological characteristics of the studied stations. The identification of frequent plant species assemblages, as well as station-specific species, helped to delineate characteristic groups, supporting the hypothesis that certain floristic associations are more recurrent and likely better adapted to specific bioclimatic and edaphic conditions.

The validation of these results using well-established statistical methods (CAH and PCA) and the dominant positioning of frequent communities in relation to the variables Id and St_Totale further reinforces the robustness of our methodology.

Additionally, the results demonstrate the relevance of a qualitative approach in an exploratory context, where the primary objective is to assess the feasibility and reliability of the methods before applying them to larger datasets and broader scales. By relying on a small dataset with significant ecological contrasts, our qualitative approach proves effective in addressing the challenges posed by the complexity of forest ecosystems while opening avenues for more extensive analyses. It provides a strong methodological foundation for further investigating vegetation succession processes and contributing to the study of ecosystem resilience under environmental pressures.

Future work should extend this study to a larger set of stations and bioclimatic conditions to determine whether the observed trends persist across broader ecological contexts.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] François, R. (2012) Éléments d'écologie: Écologie appliquée. 7th Edition, Dunod.

[2] Joseph, P. (2009) La végétation forestière des Petites Antilles: Synthèse biogéographique et écologique, bilan et perspectives. Karthala.

[3] Agrawal, R., Imieliński, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the* 1993 *ACM SIGMOD International Conference on Management of Data*, Washington D.C., 26-28 May 1993, 207-216. https://doi.org/10.1145/170035.170072

[4] Agrawal, R.R.S. (1994) Fast Algorithms for Mining Association Rules. *Proceedings of the* 20*th VLDB Conference Santiago*, Chile, 12-15 September 1994, 487-499.

[5] Han, J., Pei, J. and Yin, Y. (2000) Mining Frequent Patterns without Candidate Generation. *Proceedings of the* 2000 *ACM SIGMOD International Conference on Management of Data*, Dallas, 16-18 May 16-18, 2000. https://doi.org/10.1145/342009.335372

[6] Symphor, J.E., Mancheron, A., Vinceslas, L. and Poncelet, P. (2008) Le FIA: Un nouvel au-tomate permettant l'extraction efficace d'itemsets fréquents dans les flots de données. *Extraction et gestion des connaissances* (*EGC*'2008), *Actes des* 8*èmes journées Extraction et Gestion des Connaissances*, Sophia, 29 janvier au 1er février 2008, 157-168.

[7] Vinceslas, L., Symphor, J.E., Mancheron, A. and Poncelet, P. (2008) FIASCO: Un nouvel algorithme d'extraction d'itemsets fréquents dans les flots de données. *Extraction et gestion des connaissances* (*EGC*'2008), *Actes des* 8*èmes journées Extraction et Gestion des Connaissances*, Sophia, 29 janvier au 1er février 2008, 235-236. http://editions-rnti.fr/?inprocid=1000603

[8] Laur, P., Nock, R., Symphor, J. and Poncelet, P. (2007) Mining Evolving Data Streams for Frequent Patterns. *Pattern Recognition*, **40**, 492-503. https://doi.org/10.1016/j.patcog.2006.03.006

[9] Zaki, M.J., Parthasarathy, S., Ogihara, M. and Li, W. (1997) New Algorithms for Fast Discovery of Association Rules.

[10] Zhang, Y., Taylor, W.W. and Liu, T. (2022) Hierarchical Clustering Reveals Distinct Fish Community Structures in Response to Environmental Variation in the Yangtze River. *Aquatic Conservation*: *Marine and Freshwater Ecosystems*, **32**, 1567-1579.

[11] Legendre, P. and Legendre, L. (2012) Numerical Ecology. 3rd Edition, Elsevier.

[12] Kissling, W.D. and Field, R. (2022) Using Species Co-Occurrence Networks to Explore Biodiversity Patterns and Processes with the Jaccard Index. *Journal of Biogeography*, **49**, 973-987.

[13] Borcard, D. and Gillet, F. (2018) Numerical Ecology with R. Springer.

[14] Joseph, P., Symphor, J.É., Baillard, K., Elymarius, S., Claude, J.P., Abati, Y. and Jean-françois, Y. (2017) The Effects of Topography on Martinique's Mesological and Floristic Differentiations: The Case of Morne Carrière (Commune of VAUCLIN). *IOSR Journal of Environmental Science Toxicology and Food Technology*, **11**, 74-96.