Birds Biodiversity - Technical Report

Baptiste Pras, Raphael Leonardi

1 Data Preparation and First Visualizations

1.1 Data Preparation

1.1.1 Cleaning the species dataset

The species file contained two empty technical columns generated during export. These were removed, and the remaining columns were renamed to **French Name**, **Latin Name**, and **Origin**. To ensure completeness of the reference list, one missing species *(Aigrette bleue)*, and the missing origin of another species *(Astrild à joues oranges -> Migrateur)* were manually added. Finally, we corrected the name of a misspelled species *(Pigeon à cou rouge that was written Pigeon à coup rouge)*. This dataset serves as the key link between each observation and its biological origin category (native, introduced, migratory, etc.).

1.1.2 Cleaning the site dataset

The sites dataset also contained two unused export columns, which were deleted. The first row duplicated the column headers and was removed. The remaining columns were renamed to **Transect**, **X**, **Y**, **Type**, **Site** and **Site+Point**. The column *Site+Point* encodes the sampling location using the format *S<number>P<number>* where S is the transect unique number, and P the location unique number inside the transect. We validated this format for each location and corrected the only inconsistent value *(SIPI instead of S1P1)*. This dataset serves as the key link between each observation and its habitat type.

1.1.3 Cleaning the observation dataset

The observation logs contained raw field entries: observer, date, transect, species, and multiple counting modes. We removed metadata column *code département* not needed for analysis, and also removed all the columns linked to **distances de contact** because they were mostly made of NaN and were not going to give useful insights for our analysis. The first 2 rows from the file contained residual header information and were dropped. Total birds counting columns were renamed to **Auditif**, **Visuel**, **A+V**, **A+V vol**. These columns contained some *negative or float values and occasional strings (empty or "v")*. We used a custom conversion function to coerce those values into integers. Visual and auditory counts were then aggregated into a unified **Amount** column, representing the total number of individual birds detected at this line. We added a column **Year** where we stored for each observation the year in a proper date format. Finally, environmental descriptors *(cloud, rain, wind, visibility)* had some inconsistent values *(negative values for wind, float and some str for cloud level)* and were projected onto a bounded integer scale (1–3).

1.1.4 Standardizing text fields for merging

Species names and transect names had inconsistencies (case sensitivity, spacing, accents). A string normalization function was applied to harmonize these fields: trimming whitespace, collapsing multiple spaces, converting to lowercase, and applying Unicode normalization for accents. We created a column **clean_name** in the table *species* and a column **clean_espece** in the table *observations* with the normalized species names. We also normalized the **transect names** in the tables *observations* and *sites*. This ensures that joins between datasets are accurate. Finally, one known naming conflict in transect names *("desmarinière" vs "desmarinières")* was resolved manually.

1.1.5 Handling ambiguous species identifications

During field observations, some individuals could not be identified at the exact species level, but only at a broader taxonomic group *(e.g., "Trochilidae", "Sterne sp.", "Moqueur sp.")*. These ambiguous records would artificially reduce

species richness and bias abundance estimates if kept as-is. To preserve them and avoid losing ecological information, we created a mapping from each ambiguous group to the list of credible candidate species based on the species reference table.

For each group, we computed the empirical probability of each candidate species using all unambiguous observations made in the dataset. If no observation existed for a group, a uniform probability was applied. Then, each ambiguous record was reassigned probabilistically to a specific species according to this distribution. For example, a record labeled *Trochilidae* was reassigned among the four possible hummingbirds with probabilities proportional to the number of individuals observed for each species across the dataset.

This probabilistic reassignment preserves the total number of individuals, avoids underestimating richness, and prevents artificially inflating the presence of rare species. After this step, every observation is associated with a uniquely identified species and can be linked to its biological origin category and to the habitat type, except for two observations noted *"RAS" (Rien à Signaler)*.

After these preprocessing steps, each observation can be linked to a habitat type and a species origin category, numeric fields are clean and usable, and identifiers are consistent across datasets. The prepared dataset is now suitable for computing biodiversity indicators and analyzing temporal trends. Table 1, Table 2 and Table 3 present an excerpt of all tables after preprocessing.

Table 1: Excerpt of the species dataset after preprocessing

ID	French Name	Latin Name	Origin	clean_name
0	Aigrette garzette	*Egretta garzetta*	Migrateur	aigrette garzette
1	Aigrette neigeuse	*Egretta thula*	Migrateur	aigrette neigeuse
...	...	...	...	...
85	Viréo à moustaches	*Vireo altiloquus barbadensis*	Autochtone	viréo à moustaches
86	Aigrette bleue	*Egretta caerulea*	Migrateur	aigrette bleue

Table 2: Excerpt of the sites dataset after preprocessing

ID	Transect	X	Y	Type	Site	Site+Point
1	aéroport	714593	1614233	Mangrove	S1	S1P1
2	aéroport	714416	1614194	Mangrove	S1	S1P2
...	...	...	...	...	...	...
649	bois pothau	723481	1627061	Forêt sèche	S65	S65P9
650	bois pothau	723598	1627125	Forêt sèche	S65	S65P10

Table 3: Excerpt of the observations dataset after preprocessing

ID	Nom observateur	Nom transect	date	Passage	nuages	pluie
2	BELFAN David	fond l'etang	2014-04-12	1.0	2.0	1.0
3	BELFAN David	fond l'etang	2014-04-12	1.0	2.0	1.0
...	...	...	...	...	...	...
114495	MAUGEE Lévy	post-colon	2025-05-01	1.0	2.0	1.0
114496	MAUGEE Lévy	post-colon	2025-05-01	1.0	2.0	1.0

vent	visibilité	N° point	heure début	ESPECE	Auditif
1.0	1.0	1.0	06:20:00	Sucrier à ventre jaune	1
1.0	1.0	1.0	06:20:00	Sporophile cici	0
...	...	...	...	...	...
1.0	1.0	10.0	08:45:00	Viréo à moustaches	3
1.0	1.0	10.0	08:45:00	Elénie siffleuse	4

Visuel	A+V	A+V Vol	Amount	year	clean_espece
0	1	1	2	2014	sucrier à ventre jaune
1	1	1	2	2014	sporophile cici
...	...	...	...	...	...
0	3	3	6	2025	viréo à moustaches
0	4	4	8	2025	elénie siffleuse

1.2 First Visualizations

We now give the first visualizations of the data we made to get familiar with it.

1.2.1 Bird Observations Overview

Figure 1 details the total number of birds observed each year. We notice that 2014 is much lower than other years, this is probably due to the fact that 2014 was the start year of the program and observations started only around mid-April. We therefore miss a third of the year of observations.

Figure 2 the average number of birds observed at each observation. A mean around 4.35 means that in average, every time we notice birds, there are between 4 and 5 individuals. This statistic stayed very stable over years.

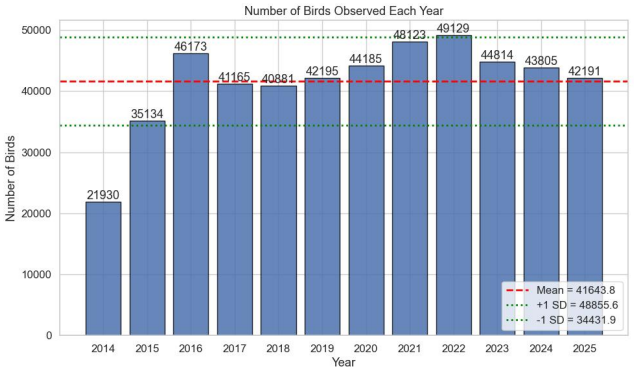

Figure 1: Total number of birds observed per year

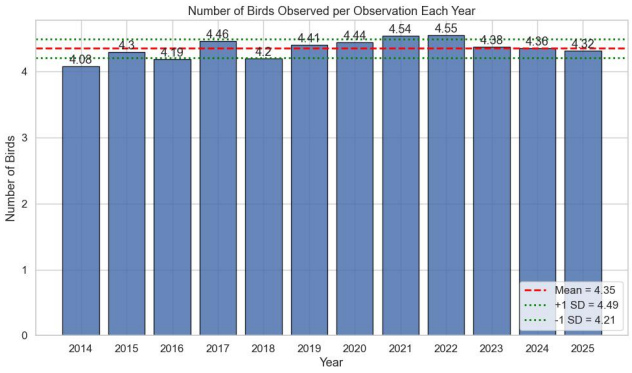

Figure 2: Number of birds observed per observation

1.2.2 Species Richness and Occurrence

Figure 3 shows the number of unique species observed each year. We notice a very stable trend over the years.

Figure 4 showcases the richest and poorest sites in number of unique species observed.

Figure 5 shows the species that were the most and least observed since the start of the program. We can note that the least observed ones are often migratory species, and that most observed ones are often native ones.

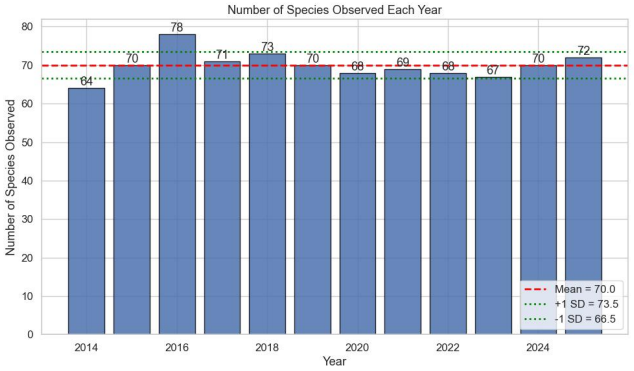

Figure 3: Number of species observed per year

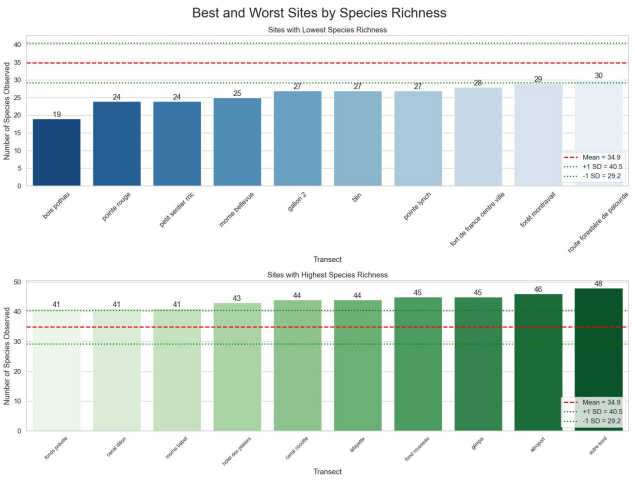

Figure 4: Number of species observed per site

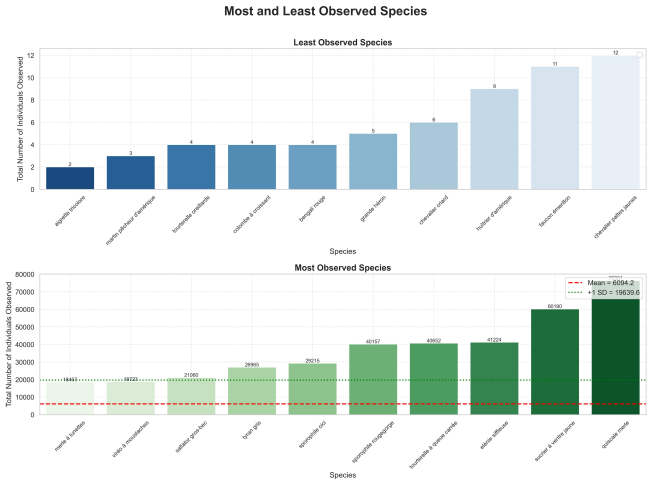

Figure 5: Most and least observed species

1.2.3 Spatial Habitat Overview

Figure 6 shows the distribution in coordinates X, Y of the different observation points, and their habitat.

Figure 7 is a map of the Martinique filled with an estimation of the different habitats on the island. This estimation was made using the habitat of the various observation points and using a KNN on these coordinates.

Figure 8 show the number of observation points for each habitat type. We notice that *Forêt sèche* is the habitat with the most observation points. *Agricole*, *Forêt humide* and *Périurbain* also have a lot of observations points, contrary to *Mangrove*, *Plage* and *Urbain* who have very few observation points. This distribution might be important to consider for later studies.

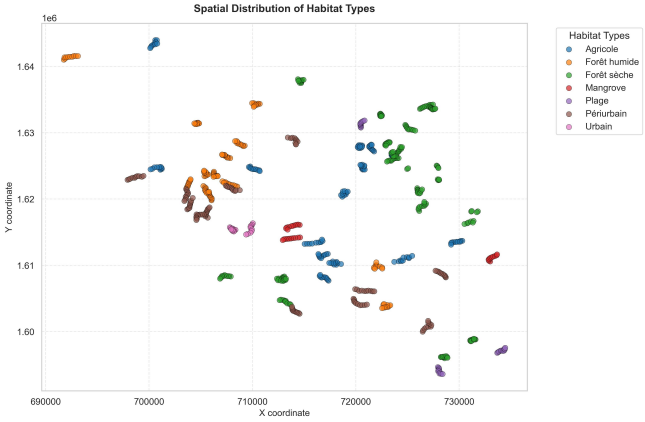

Figure 6: Spatial distribution of habitats

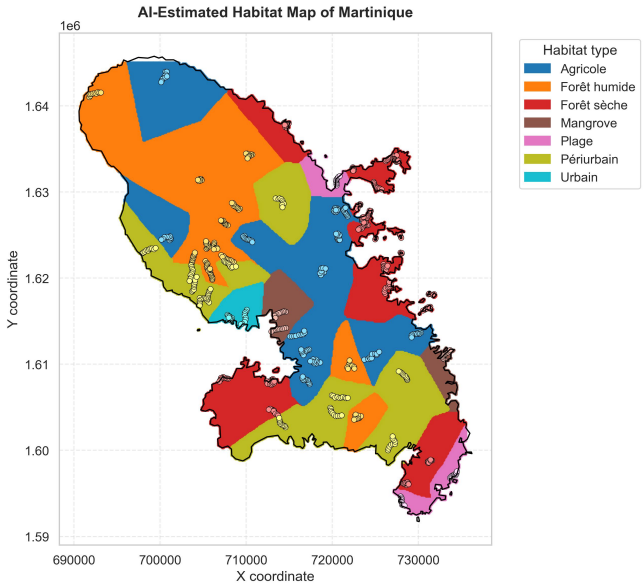

Figure 7: Map of the habitats of Martinique

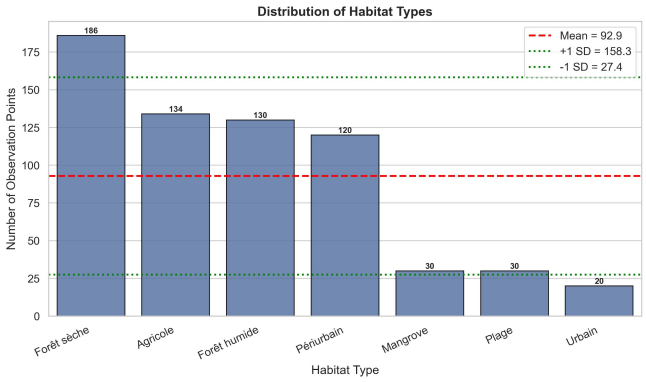

Figure 8: Distribution of habitat

1.2.4 Detection Modalities

Figure 9 shows the number of detections that were auditory versus the number of detections that were visual. Note that both auditory and visual detections are not counted here.

Figure 10 show the number of birds that were seen flying, and the number of birds that were seen not flying.

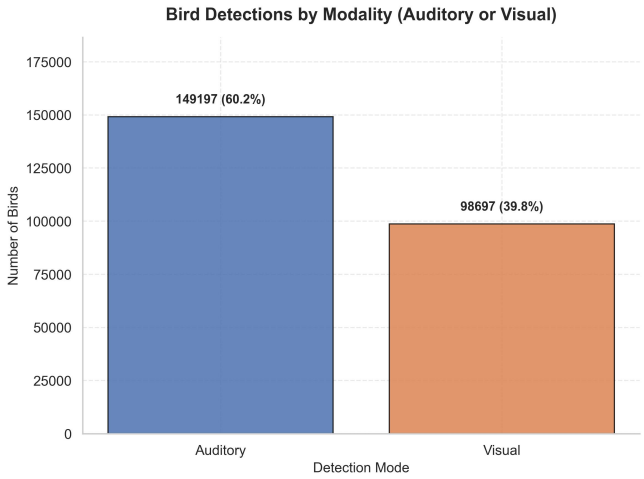

Figure 9: Auditory vs visual detections

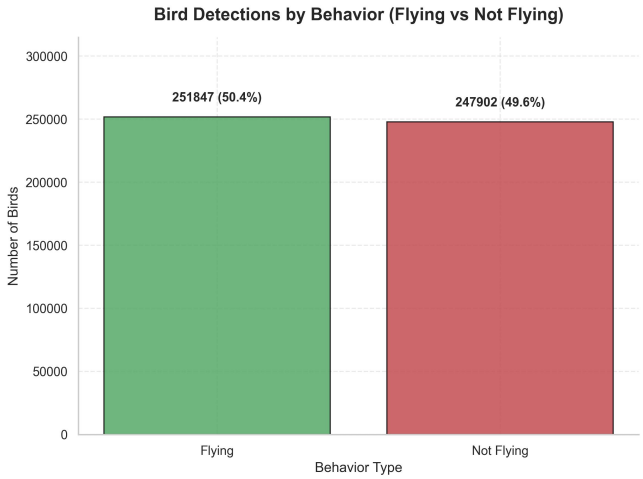

Figure 10: Fly vs no-fly detections

1.2.5 Environmental Conditions

Figure 11 shows the distribution of the weather conditions during the observations. The values can be either 1, 2 or 3. We do not have the information whether 1 is the lowest or highest level. We have the clouds level, visibility level, wind level and rain level.

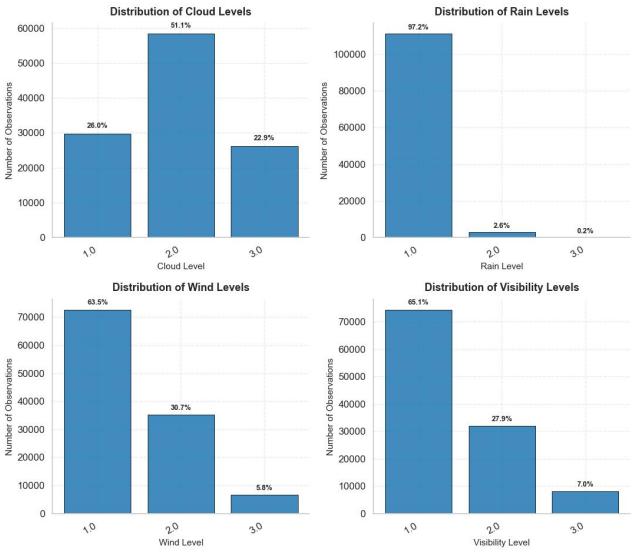

Figure 11: Weather conditions

2 Multi-Years Trends

In this section, we define three indicators: two biodiversity indicators and one sampling effort one. In the next three subsections, we first define and justify the indicators, we then compute and discuss the global values of the indicators over the whole observations, and then we display and discuss confidence intervals and trends over the years.

2.1 Species and Individual Richness per Habitat

The first indicator focuses on **species richness and total/normalized abundance** (normalized abundance is the number of individuals observed per unit of sampling effort) per habitat. This choice is motivated by the fact that **richness and abundance** are two of the **most fundamental ecological descriptors used to assess the state of an ecosystem**. Richness informs us about the variety of species present, while abundance reflects how intensively a habitat is used by wildlife. Monitoring how these values change across habitats and over years provides a direct way to **evaluate habitat quality and ecological stability**. A habitat that consistently hosts many species or large numbers of individuals is likely to offer suitable resources such as food, shelter, and breeding sites. Conversely, a decline in richness or abundance can signal environmental degradation, lack of resources, or anthropogenic pressure. This indicator therefore gives us a solid foundation for identifying at-risk habitats and guiding management actions to preserve biodiversity.

2.1.1 Global Overview

Figure 12 shows the total and normalized abundance of individuals per habitat, and the number of unique species observed in each habitat. The normalized abundance is computed by dividing the total abundance per habitat by the unique number of visits of the given habitat.

We notice a sharp difference between total and normalized; notably, *Agricole* is very high in total abundance while barely over the average in normalized abundance. Conversely, *Urbain* is very low in total abundance, while very high in normalized one. This can be explained by the fact that, as shown in Figure 8, we have many observation points in the habitat *Agricole*, but very few in *Urbain*, therefore biasing the total abundance.

Note that we can explain the high value of normalized abundance for *Urbain* by the nature of the observations, and that it does not necessarily mean that urban areas have a lot of birds and a healthy biodiversity. In urban areas, we can see a lot of birds such as pigeons because of the abundance of food, and have a much better visibility than in forests for instance, therefore making our time spent in such areas very efficient in terms of bird detection.

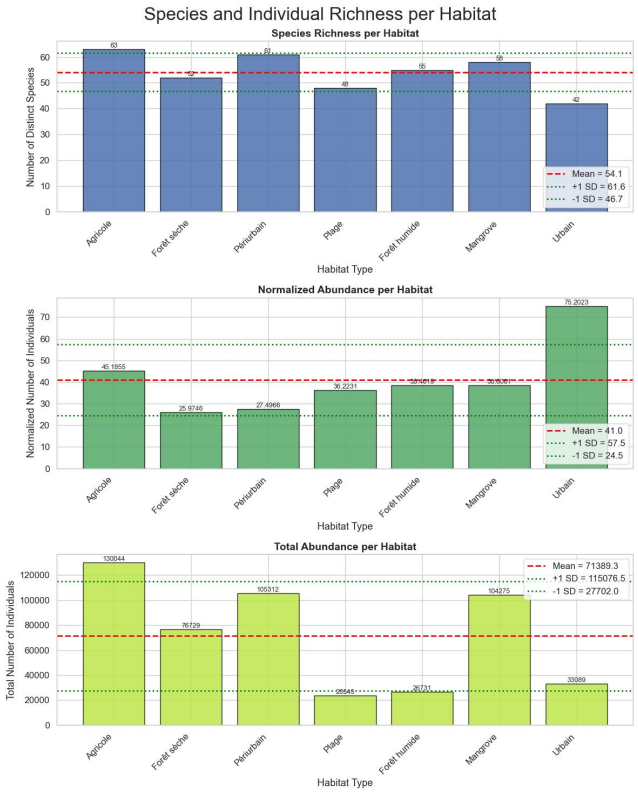

Figure 12: Species and individual richness per habitat

2.1.2 Yearly Averages and Confidence Intervals

To quantify the uncertainty of our indicators, we rely on a **non-parametric bootstrap resampling procedure**. Ecological monitoring datasets are often **unbalanced, noisy, and do not follow classical statistical distributions** (species and individuals counts are discrete, skewed, and may vary strongly between transects). Standard parametric confidence intervals assume normality and homogeneous variance, which are unrealistic in this context. The bootstrap avoids these assumptions by resampling the original transects with replacement, generating an empirical distribution of the indicator values. From this distribution, we derive confidence intervals without assuming any underlying distribution.

For abundance, since the same birds can be **detected multiple times** or, conversely, some birds may be **undetected**, we use bootstrapping to produce a **symmetric confidence interval** representing how the measured abundance would fluctuate if the survey were repeated under similar conditions. The confidence intervals therefore represent the uncertainty around the true number of individuals observed.

For species richness, the interpretation is fundamentally different from abundance. The number of species observed S_{obs} is a **confirmed minimum**, not an estimate of the true richness (we assume that the observers did not say they saw a species if they are not sure). But some species present at a site may simply **remain undetected during surveys**. Using a symmetric bootstrap interval around the observed value would be inappropriate, since the **lower bound cannot fall below the set of species that were actually detected**. To account for undetected species, we use the **Chao2 estimator**, which infers the number of missed species by analyzing the frequency of rare species in the sample (specifically, species observed in only one transect (singletons, Q1) or two transects (doubletons, Q2)). The logic is intuitive: *if many species appear in only one or two transects, it suggests that additional rare species likely exist but were not encountered during sampling.*

However, the Chao2 estimator can become unstable when sampling coverage is low, potentially producing unrealistically high estimates. To address this, we apply two safeguards: an **absolute cap at 87 species** (the total number of species in the regional pool) and **Conservative adjustments** when the singleton/doubleton ratio indicates poor coverage

To quantify uncertainty, we **bootstrap the Chao2 estimator** by resampling transects with replacement and extract the **97.5th percentile** of the resulting distribution $chao2_{high}$. This yields an asymmetric confidence interval: $[S\_obs, min(chao2_{high}, 87)]$. We obtain a 97.5% bootstrap confidence interval, meaning that in 97.5% of the bootstrap resamples, the estimated richness does not exceed this upper bound. This approach acknowledges both data uncertainty and imperfect detection, providing a more realistic interval within which the true biodiversity level lies.

Figure 13 shows the total abundance per year for each habitat and the uncertainty around the observed values, as discussed previously.

Figure 14 shows the normalized abundance per year for each habitat and the uncertainty around the observed values.

For both figures, we notice a few outliers. On both figures, 2014 for *Mangrove* and 2014/2015 for *Urbain* don't have any confidence interval. This occurs when there is **only one unique transect visit**, making bootstrap resampling impossible. We also notice some confidence intervals being quite large, like *Mangrove* 2016/2022 *Plage* 2016/2017, or *Urbain* 2018. This indicates that the bootstrap could run, but that the dataset contains very **few independent visits with high variability between them**, which results in a high estimation uncertainty.

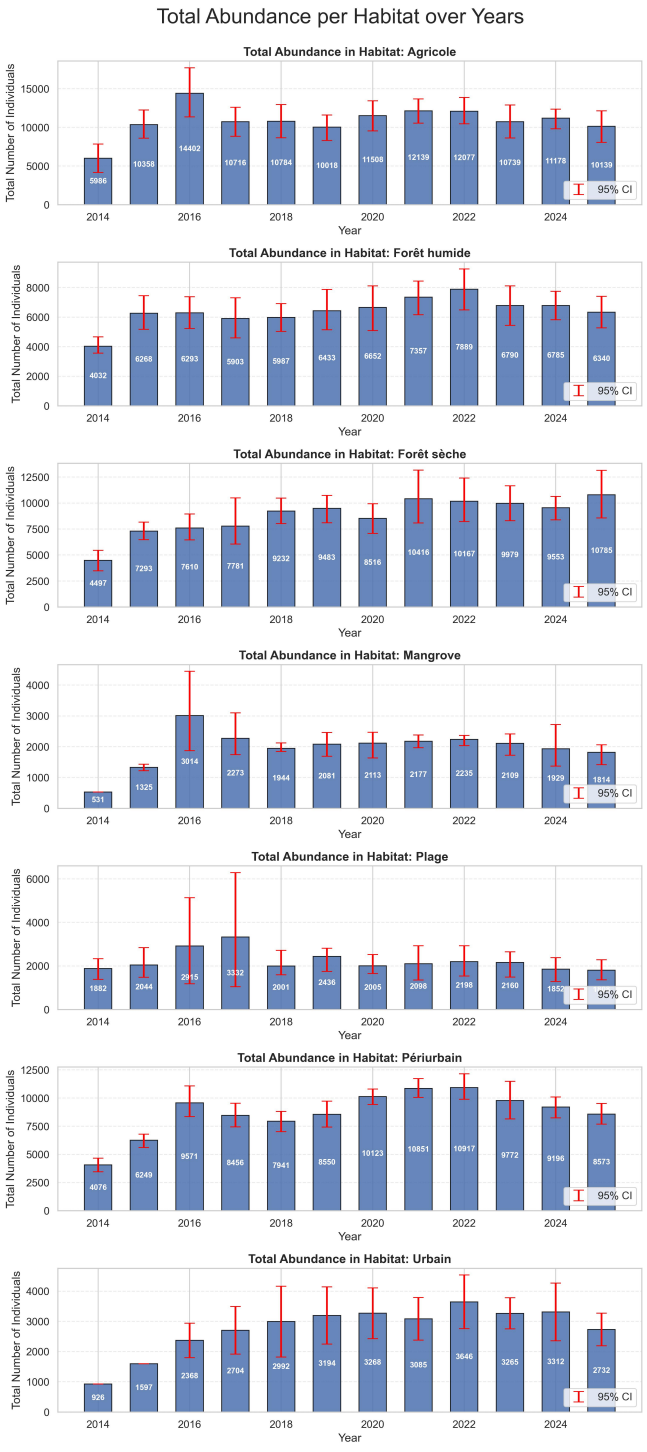

Figure 13: Total abundance per habitat per year

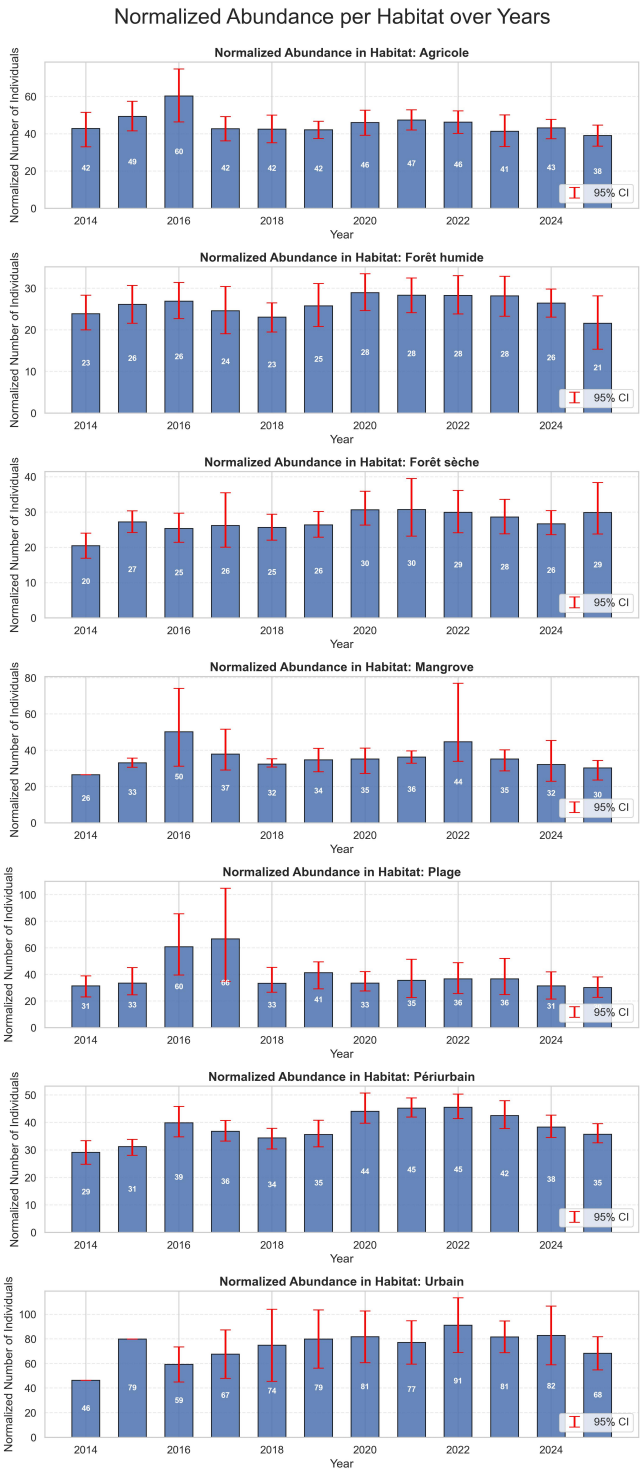

Figure 14: Normalized abundance per habitat per year

Figure 15 shows the observed species richness per habitat and per year, along with a confidence band derived from the Chao2 estimator[1], as described above.

Chao2 is an incidence-based richness estimator designed to infer the number of species that were potentially present but not detected. Wide intervals (as for *Forêt sèche*, *Agricole* or *Périurbain*) indicate that the estimator considers many species as "possibly undetected", which inflates the upper bound.

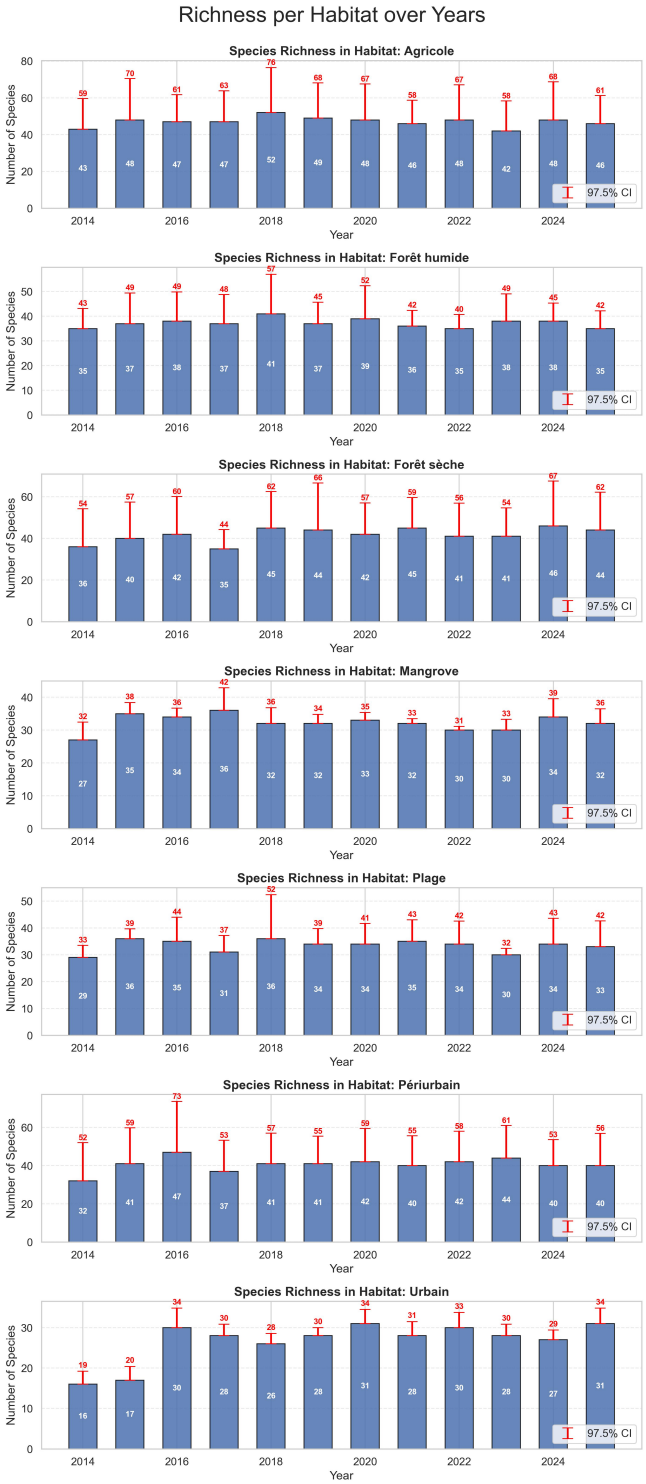

Figure 15: Species richness per habitat per year

[1]Note that this estimator is a standard incidence-based richness estimator in ecology. However, when the sampling effort is low or when many species are observed only once (singletons), the estimator becomes unstable and may inflate richness estimates.

2.1.3 Temporal Trends

We now discuss temporal trends for species richness and abundance. Figure 16 shows the evolution of observed richness and abundance over the years for each habitat.

To visualize long-term trends without assuming that biodiversity evolves linearly over time, we used a **non-parametric smoothing method** called **LOWESS** (Locally Weighted Scatterplot Smoothing). Unlike linear or polynomial regression, LOWESS does not impose a specific functional form between time and the indicator. Instead, it builds the trend curve by fitting multiple **weighted local regressions** on subsets of the data.

Mathematically, for each point x_i (a given year), LOWESS estimates the smoothed value $\hat{y}(x_i)$ as a weighted combination of neighboring observations:

$$\hat{y}(x_i) = \sum_{j=1}^{n} w_{ij}\, y_j$$

where the weights w_{ij} decrease as the distance between x_j and x_i increases. A commonly used weighting kernel is:

$$w_{ij} = \left(1 - \left(\frac{|x_j - x_i|}{d_i}\right)^3\right)^3$$

with d_i being the distance to the k-nearest neighbor (controlled by the parameter *frac*). In our case, $frac = 0.5$, meaning that for each year, LOWESS uses the closest 50% of the data to compute the local regression.

This smoothing technique is particularly suited to ecological datasets because:

- biodiversity indicators are **noisy** and may vary strongly from one year to the next
- counts are **non-linear and non-Gaussian**, violating classical assumptions of parametric models
- we do not want to impose a trend shape (linear, quadratic, etc.)

LOWESS thus highlights the **overall temporal tendency** of richness and abundance, while remaining flexible and data-driven.

For normalized abundance, *Urbain* has a sharp increase from 2014 to 2019, but then starts to flatten until 2022, when it starts decreasing. The other habitats stayed relatively flat (very small increase) until 2022, and then also started to slowly decrease too.

For total abundance, *Forêt humide*, *Agricole*, *Forêt sèche* and *Périurbain* had a sharper increase until 2022 than *Urbain*, *Plage* and *Mangrove*, but except for *Forêt sèche* that kept increasing, all other habitats started decreasing in 2023.

To summarize, both total and normalized abundance follow similar trends, with an increase from 2014 to 2022, and then a slow decrease. Note that since 2025 observations were interrupted in May in the dataset, the total abundance is biased (less sampling effort than for other years), and that the sharpest increase happen between 2014 and 2016, which fits 2014 and 2015 being the two years with the least observations made, as discussed previously.

We therefore rely more on normalized abundance that gives fairer trends over years. With normalized abundance, except for the urban areas that had a sharp increase from 2014 to 2020, all other habitats have almost no evolution in their abundance until 2022, and all habitats (except the *Forêt humide* that stays stable) seem to be declining after 2022.

For species richness, we notice an increase from 2014 to 2016 for all habitats, that can be attributed to lower sampling effort in 2014 and 2015. The overall trends are then very flat and it seems that there is no clear evolution in the number of species in any habitat, except for the urban areas that still increased from 2016 to 2018/2019.

In conclusion, despite some yearly fluctuations, the multi-year trends remain **stable**: species richness does not show any long-term increase or decline, and abundance only varies moderately across habitats. **Urban areas stand out with a marked rise in abundance and richness during the first half of the study period**, but this appears to have stabilized now. Across all habitats, the slight decrease in abundance after 2022 suggests that **bird populations may be starting to diminish**, though this pattern remains recent and should be interpreted with caution. Continued monitoring will be necessary to determine whether this marks the beginning of a true downward trend or simply natural variability.

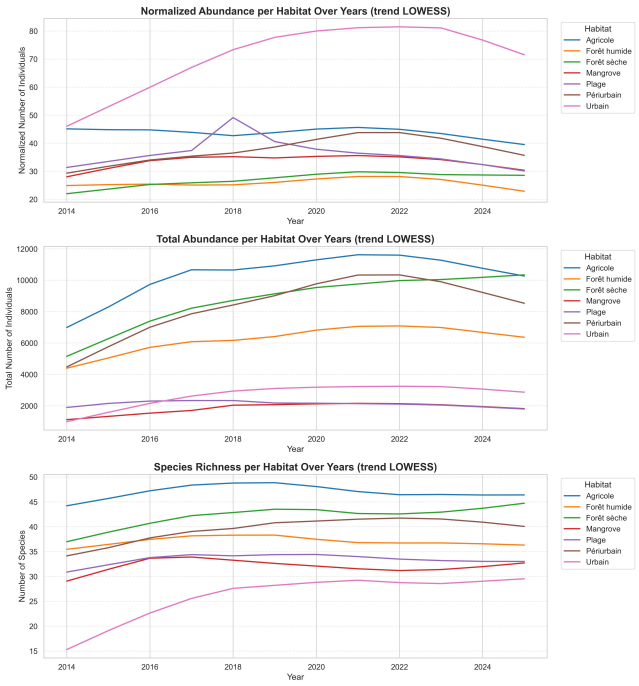

Figure 16: Species richness and abundance per habitat, trends over years

2.2 Shannon Diversity Index and Origin Proportions per Habitat

The second indicator combines two complementary metrics: the **Shannon Diversity Index** and the **biogeographical origin composition of species** observed in each habitat. Species richness alone does not capture how individuals are distributed among species, a habitat may host many species, but if a few dominant ones represent most observations, the ecosystem may still be unbalanced. The Shannon Index addresses this by measuring **both richness and evenness**: it increases when species are not only numerous but also equitably represented. A high Shannon value therefore reflects a habitat with limited ecological dominance and a diversified use of available resources.

However, ecological diversity does not automatically imply ecological quality. A habitat could appear diverse while being largely dominated by non-native or invasive species. For this reason, we couple Shannon with a second metric: **the proportion of individuals belonging to native, endemic, introduced, or migratory species**. While Shannon evaluates the structural balance of species within a habitat, **origin composition evaluates its ecological integrity**. A high proportion of native/endemic species indicates ecological stability, whereas a strong presence of introduced or invasive species may be an early warning signal of ecosystem disturbance, habitat degradation, or anthropogenic pressure.

Together, these two measures allow us to characterize habitats not only by how diverse their bird communities are, but also by how healthy and ecologically functional they remain over time.

2.2.1 Global Overview

Figure 17 shows the Shannon diversity index for each habitat, alongside the composition of species origin of each habitat. We first plot the precise origin, and then group origins into 3 categories: **native** *(endémique and autochtone)*, **introduced** *(exogène)*, **migratory/marine**.

The Shannon diversity index is defined as:

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

where:

- S is the total number of species

- p_i is the proportion of individuals belonging to species i, computed as

$$p_i = \frac{n_i}{N}$$

with n_i the number of individuals of species i, and N the total number of individuals observed in the habitat.

Values close to 0 indicate low diversity (one dominant species). Higher values indicate habitats where species are both numerous and evenly represented. We plot the maximum value in red (computed as $ln(S)$) and indicate the difference between the observed value and the theoretical maximum. The maximum is reached if all species have the same number of individuals.

Most habitats have a similar H' value, but some have a lower difference from their theoretical max. In particular, *Mangrove* has a difference of 1.02, and *Forêt humide* of 1.11, when most others are between 1.23 and 1.24. This would suggest that these two habitats have a more diverse population. Note that urban areas have the largest difference with 1.28, while also the smallest theoretical max, making a much bigger relative difference than others. This is no surprise when we know that some invasive and dominant species like pigeons tend to live in cities, where they can thrive thanks to the abundance of food. With a **ratio of the difference from the max over the max** that is **over 2/3** for all habitats, we can still conclude that the habitats **do not seem to be impacted by dominant species**.

For grouped species proportions, most habitats have about 85% or more native individuals, highlighting a **natural and well-balanced biodiversity**. *Agricole* and *Urbain* stand out with respectively 72.3% and 53.5% of native species. This could suggest that **introduced species might thrive** in these habitats and cause a threat to the native species, especially in urban areas.

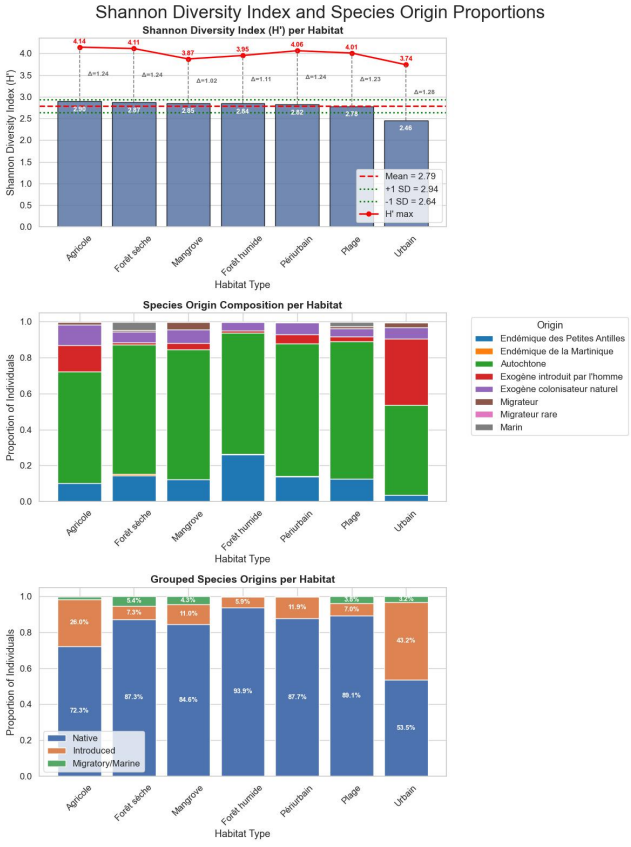

Figure 17: Shannon diversity index and origin proportions per habitat

2.2.2 Yearly Averages and Confidence Intervals

Unlike abundance or richness, the Shannon diversity index and species origin proportions are **derived metrics** (functions of relative abundance): they do not count individuals, but rather quantify how individuals are distributed among categories (species for Shannon, origins for native/introduced/migratory composition).

Because these metrics are based on **relative abundances** (ratios and logarithms), classical parametric confidence intervals are inappropriate. They rely on assumptions such as normality and homogeneous variance, which do not

hold for ecological community data.

To quantify the uncertainty, we use a **non-parametric bootstrap at the transect level**. For each habitat and each year:

1. transects are resampled with replacement
2. the bird communities of the resampled transects are reconstructed
3. Shannon index and origin proportions are recomputed

Repeating this procedure many times produces an empirical distribution of Shannon values and origin proportions. We then extract the **2.5th and 97.5th percentiles** as the confidence interval.

These confidence intervals do not express uncertainty over the total number of birds, but rather over the **structure of the community**. In other words, they quantify how much the Shannon index or origin proportions would fluctuate if the sampling were repeated under similar conditions.

This method is particularly suited for community diversity analysis because it:

- does not assume any specific distribution of the data
- naturally incorporates variability in sampling effort
- is sensitive to rare species or rare origins, which strongly affect Shannon and composition ratios

Figure 18 shows the observed Shannon diversity index per habitat every year, along with the confidence interval.

Figure 19 shows the origin proportions per year for all habitats, with confidence intervals showing the uncertainty around this estimation of the proportion.

For both figures, we notice the same outliers. 2014 for *Mangrove* and 2014/2015 for *Urbain* don't have any confidence interval. As stated during the analysis of the confidence intervals for the indicator 1, this occurs when there is **only one unique transect visit**, making bootstrap resampling impossible. Otherwise, only the origin proportions in urban areas seem odd, with very large confidence intervals, probably because of **few independent visits with a high variability between them**.

2.2.3 Temporal Trends

We now discuss temporal trends for Shannon diversity index and origin proportions. Figure 20 shows the evolution of the Shannon index over the years for each habitat, using the smoothing method **LOWESS** (this was explained in section 2.1.3). Figure 21 shows the evolution of the origin proportions.

For the Shannon diversity index, we notice various trends depending on the habitat. *Forêt sèche*, *Périurbain* and *Urbain* continuously increased their H' value, while *Mangrove* increased until 2017 and then stabilized, and while *Plage* increased until 2020 but then started decreasing until 2025. Finally *Agricole* remained somewhat stable but with a slight increase while *Forêt humide* also remained somewhat stable but with a slight decrease. These trends suggest that in general, habitats do not seem to suffer from dominant species, and that the diversity and evenness of species has increased for some habitats, and is even still increasing for *Urbain*.

Origin proportions remained very stable for most habitats, except for *Urbain*, which shows large year-to-year fluctuations with an overall decrease in the proportion of native species between 2014 and 2025, and for *Agricole* and *Périurbain* that experienced a continuous decrease in the proportion of native species, and an increase in the proportion of introduced ones. The loss in the proportion of natives for *Agricole* is even very concerning, as it lost about 15 to 20 percentage points in the span of 10 years.

Interestingly, we can notice that for *Mangrove*, the proportion of migratory and marine species decreased over time, while it increased for *Forêt sèche*, suggesting that such species might have moved from one habitat to another over the years.

In conclusion, trends suggest that most habitats seem to **remain ecologically stable and do not appear to be impacted by dominant or invasive species**. But we must note that *Agricole* and *Périurbain* are **losing their native population** and may require targeted management actions. **Urban habitats require particular attention**. The **increase in Shannon diversity** indicates a **more even distribution of species**, yet the parallel **decrease in the proportion of native species** suggests that this evenness is **driven by the establishment of introduced species** rather than by the recovery of native biodiversity. Thus, the observed improvement in diversity may **mask an underlying process of biotic homogenization**. Finally, we note the **shift in the presence of migratory and marine populations** from *Mangrove* to *Forêt sèche*.

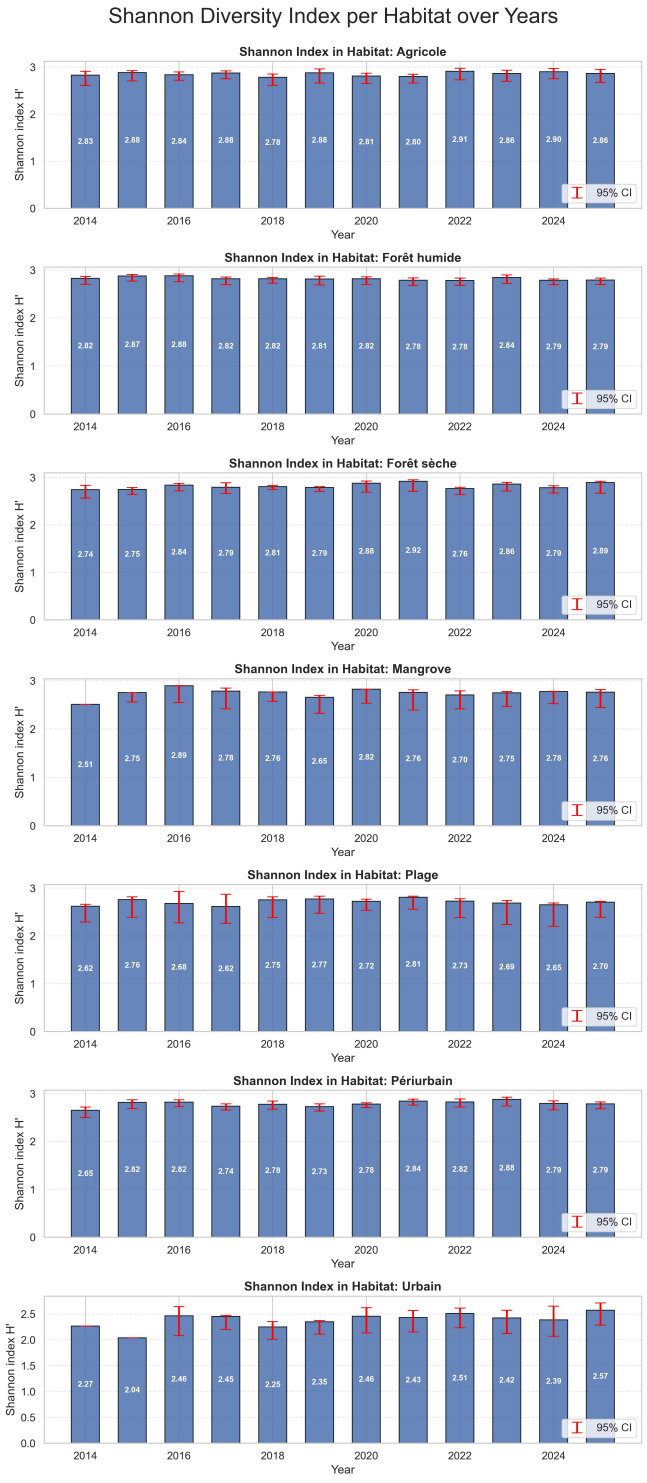

Figure 18: Shannon index per habitat per year

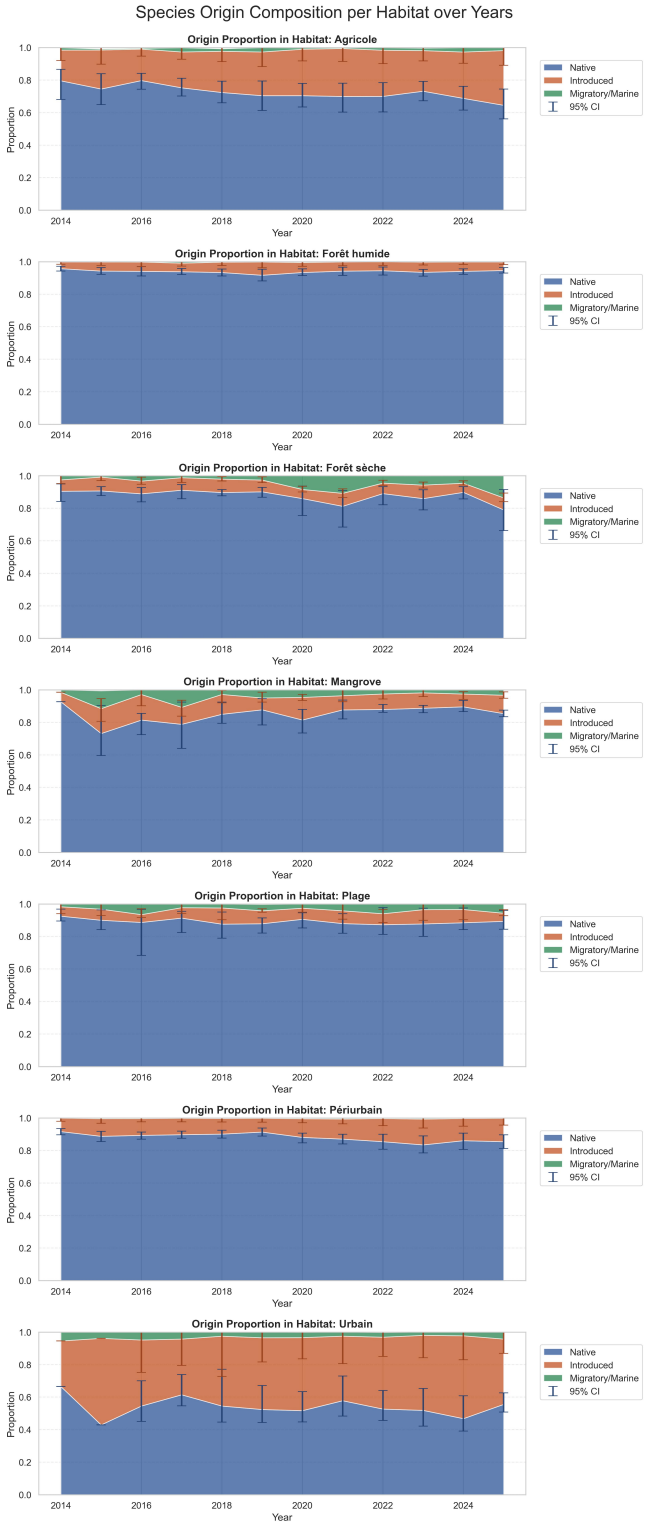

Figure 19: Origin proportions per habitat per year

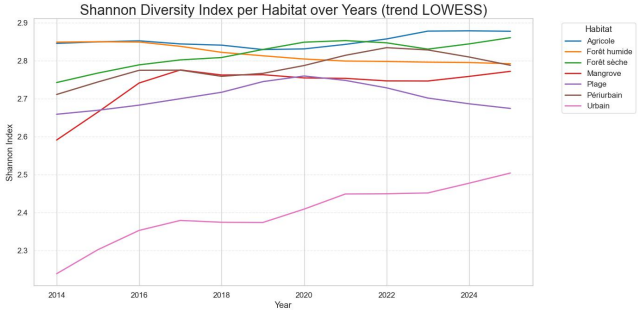

Figure 20: Shannon index per habitat, trends over years

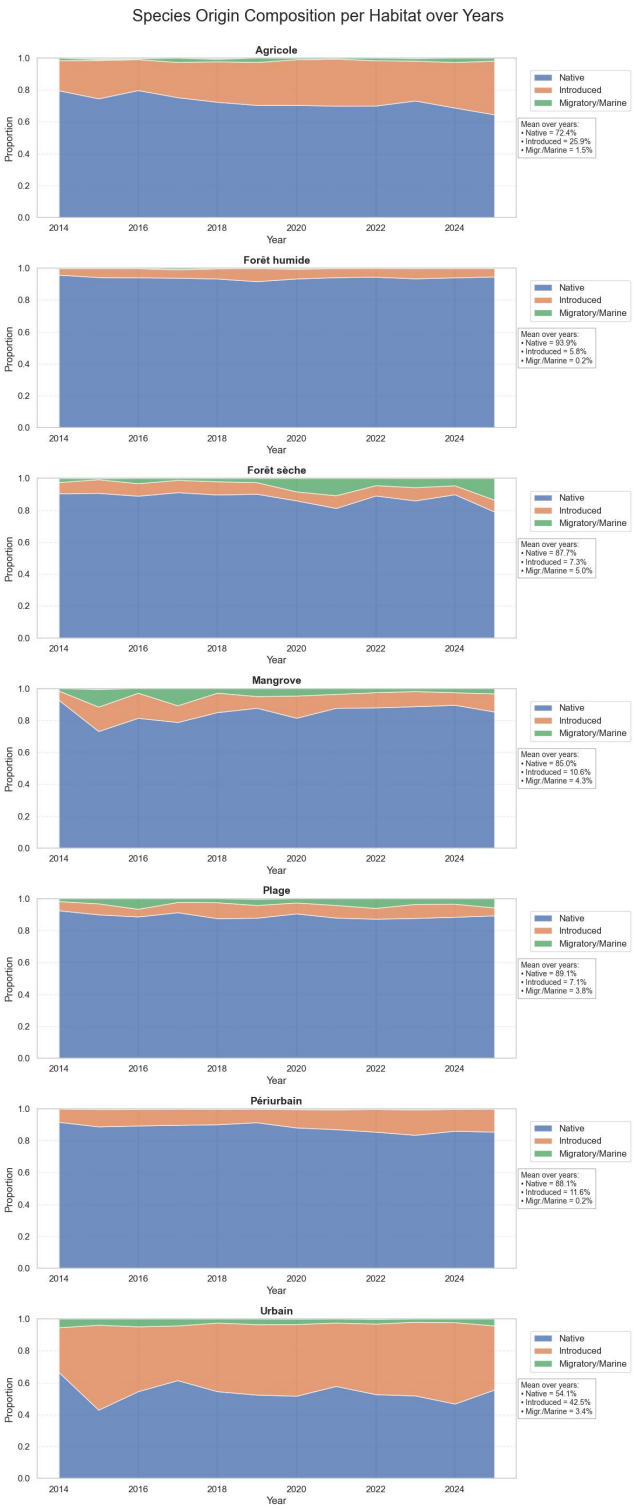

Figure 21: Origin proportions per habitat, trends over years

2.3 Sampling Effort (Unique Transects Visited and Visit Intensity)

The third indicator evaluates the **sampling effort** through two complementary metrics: the **number of unique transects surveyed each year** and the **intensity of visits per transect**. This indicator is essential because biodiversity estimates (richness, abundance, or community composition) are **meaningful only if the underlying sampling effort is consistent**. Variations in the number of sites visited or in survey frequency can produce artificial trends that mimic ecological changes, while in reality they may simply reflect differences in observer effort. By explicitly tracking sampling effort, we **ensure that increases or decreases in species observations are interpreted correctly and not confounded with uneven fieldwork**. This also allows us to verify the spatial representativeness of the monitoring program and to detect potential gaps. Ultimately, this indicator strengthens

the reliability of the conclusions drawn from indicators 1 and 2, and provides valuable feedback to improve future monitoring strategies.

2.3.1 Global Overview and Temporal Trends

Figure 22 shows the number of unique transects visited each year and the heatmap of the visit intensity per transect. The visit intensity is displayed as a heat map, where high values (in blue) for a transect means that this transect was visited a lot of times during a year, and where low values (in white) means that this transect was not visited a lot of times during a year. We also display on the right the mean value per year for each transect. We do not count years where the transect was never visited in the mean to avoid biasing the mean per transect.

Figure 23 shows the **LOWESS** trend of the number of unique transects visited each year.

We notice that the number of unique transects visited each year has increased by 12 from 2014 to 2015, then by 6 in 2016, by 1 in 2017 and by 4 in 2018 to reach 64. From 2020 to 2022, the number of unique transects visited decreased to 63, before going back to 64 in 2023 and 2024, and finally increasing to 65 in 2025. This shows that the program is continuously trying to increase the number of transects visited. The decrease to 63 between 2020 and 2022 is probably due to the inability to access a specific transect (for instance, temporary restrictions such as COVID-19 sanitary measures, or physical causes such as landslides or terrain instability that made the site inaccessible).

The heatmap shows in details the observations made previously. All white cells are transects never visited during a year, we clearly notice some outliers, such as the one that is white until 2025 (the one never visited before 2025, the 65th transect), or the one that stopped being visited between 2020 and 2022. The mean per transect suggests some disparities in the visit intensity among transects, some being visited over 200 times per year in average, while some are visited between 100 and 150 times. However, the colors in the heatmap seem more homogeneous in 2024 and 2025 than they were previously, suggesting that the program is trying to have a more even visit intensity among all transects. Note that some transects might be more or less visited also because of their accessibility or beauty, and that most observers probably are volunteers.

Note that these indicator values are not estimates derived from a statistical model or from a sample used to infer a population parameter. They are directly observed operational data, determined by the survey protocol (how many transects were scheduled and how many were actually visited). There is no sampling variability: visiting 42 transects is a factual event, not a stochastic outcome. Therefore, computing confidence intervals do not make sense, confidence intervals quantify uncertainty around an estimate, not around fixed logistics decisions or recorded effort.

In conclusion, the program seems to be **developing well and trying to make up for the disparities in visit intensity**, which is a good sign and might help **improve the robustness and comparability of biodiversity estimates**.

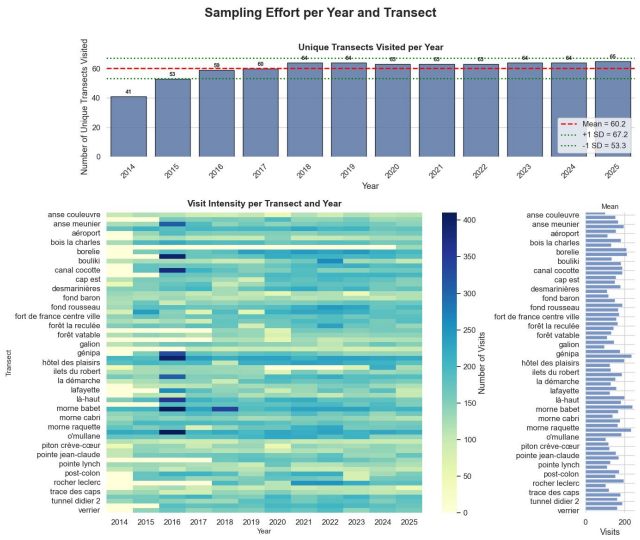

Figure 22: Sampling effort per year and per transect

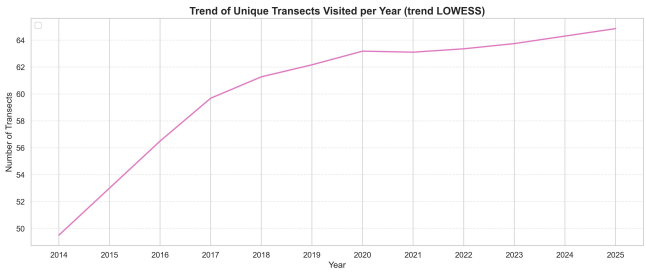

Figure 23: Sampling effort per year and per transect, trends over years

3 Species-Level Trends

In this section, we now focus our study on the evolution of species populations. To keep **reliable conclusions and avoid noise**, we only considered species for which **at least 220 specimens were observed**. Since the project began in April 2014, the **total number of observations for that year is way lower than for other years**, we therefore decided to exclude it from our analysis to **avoid bias** in the results.

We study the evolution of species based on two criteria:

- First we compare the species that shows the **largest decrease of its population** with the species that shows the **largest increase of its population**.
- Then we compare the species that show the **largest diversity of habitat** with the species that show the **lowest diversity of habitat**.

3.1 Largest Increase and Decrease in Population

The purpose of this work is to assess the biological health of the island of Martinique. One effective way to do this is by **identifying species with a declining population**, as identifying one could then **help us assess the reasons** to give insights on biodiversity preservation. Likewise, **studying species whose population is increasing** can **help us understand why some species thrive** while others do not. Exploring these trends allows us to better understand how to **preserve and support biodiversity**.

We also aim to predict the number of specimens per species in the coming years. Such predictions can help us **anticipate potential future declines and prepare conservation strategies**. In addition, they can reveal which species might become overly abundant and therefore help **avoid an excessive population growth** that can disrupt natural ecosystems and create strong competition for other species. Moreover, an overabundance of certain species could negatively impact agriculture on the island. To summarize, predicting those populations can help us understand the future and prepare a preservation strategy.

The size of the database will naturally increase as the monitoring program continues. Although the dataset will contain more rows over time, the prediction task itself remains low–dimensional: we want to predict a single value (the population of a species), based only on the year. In this setting, a **linear model** of the form

$$\hat{Y} = aX + b$$

could be appropriate, since it is easy to train and interpret. The idea would be to fit the model on **historical observations** and use it to forecast future population values.

However, we must be cautious. For many species, the **number of available historical observations is very limited**, which could make the model **unstable or unreliable**. To properly assess whether such predictions are meaningful, it is necessary to quantify their uncertainty, analyze confidence intervals, and interpret the trends accordingly.

3.1.1 Linear Model Confidence Intervals

Let x and y denote two variables. We consider a linear model that determines the coefficients a and b of the equation $\hat{y} = ax + b$ that best fits the observed data. We want to compute the confidence interval for this model.

Compute the residues

$$\mathbf{r} = y - \hat{y}$$

Compute :

$$\mathbf{SSE} = \sum_{i=1}^{n} r_i^2$$

It is called Sum of Squared Errors, measuring how far the predicted values vary from the real values.

The Mean Squared Error (MSE) is defined as

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

where SSE is the sum of squared residuals and n is the number of observations.

We also define :

$$S_{xx} = \sum (x_i - \bar{x})^2$$

which measures the variability of the predictor variable x.

The critical value of the t-distribution for a confidence level $1 - \alpha$ and $n - 2$ degrees of freedom is given by

$$t_{\text{crit}} = t_{1-\alpha/2,\, n-2}$$

We compute two standards errors : one to calculate the Confidence Interval for the mean of our data, given by :

$$SE_{\text{mean}} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)}.$$

And another for the Confidence Interval for the prediction of values given by

$$SE_{\text{pred}} = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)}.$$

The confidence intervals are then :

$$CI_{mean} = [\bar{x} - t_{crit} * SE_{mean}, \bar{x} + t_{crit} * SE_{mean}]$$

and

$$CI_{pred} = [\hat{y} - t_{crit} * SE_{pred}, \hat{y} + t_{crit} * SE_{pred}]$$

We chose to compute 95% confidence intervals, meaning that $\alpha = 0.05$. A 95% **confidence interval for the mean** indicates that if we were to resample our data many times under the same conditions, then in 95% of those resamples, the interval CI_{mean} would contain the true mean of the population. For the **prediction interval**, a 95% interval means that if we were to collect a new observation under the same conditions, then in 95% of cases, the true value of that future observation would fall within the prediction interval CI_{pred}.

3.1.2 Evolution of Populations

We compute for each species the amount of individuals observed each year. For each species, we select the lines in the table **df_observations** concerning this species. We perform a **groupBy** by the key **year**, and calculate the sum of the column **Amount**.

To compute the rate of change, we compute the increase (or decrease) in percentage between the first last year, and store it in a **pandas.Series** named **changeSpecies**. We then choose the two species to study by computing which ones have the largest and smallest value. We find that the species that shows the **largest decrease** in population is the species named **colibri falle-vert**. The species that shows the **largest increase** is the species named **tourterelle turque**. We can visualize this evolution, the linear model and the Confidence interval on **matplotlib**. After choosing the species and storing their evolutions, we use the Python library **sklearn** to fit a linear model for our data. We compute the confidence intervals to study how confident we are on this linear model.

On Figure 24, we can notice on the upper graph the decrease of the population of the *colibri falle-vert*, going from 236 specimens in 2015 to 108 in 2025. The dotted line in green shows the linear trend estimated by the model. The green surface represents the confidence interval for the mean of the values. The black surface represents the confidence interval for the prediction model, that ensures that all the data is inside it with a 95% confidence. Finally the yellow bars are the prediction of the observed population for the next 2 years. We can see that the confidence interval is quite large, which suggests that the model is quite unsure about its predictions, suggesting that a linear model might not be the best way to fit our data.

The lower graph displays the increase of the population of the *tourterelle turque*, going from 329 specimens in 2015 to 885 in 2025. The confidence interval for the prediction model encompasses well the data. The confidence interval for the predictions is still quite big, which suggest that our linear model is still not a good fit in this scenario.

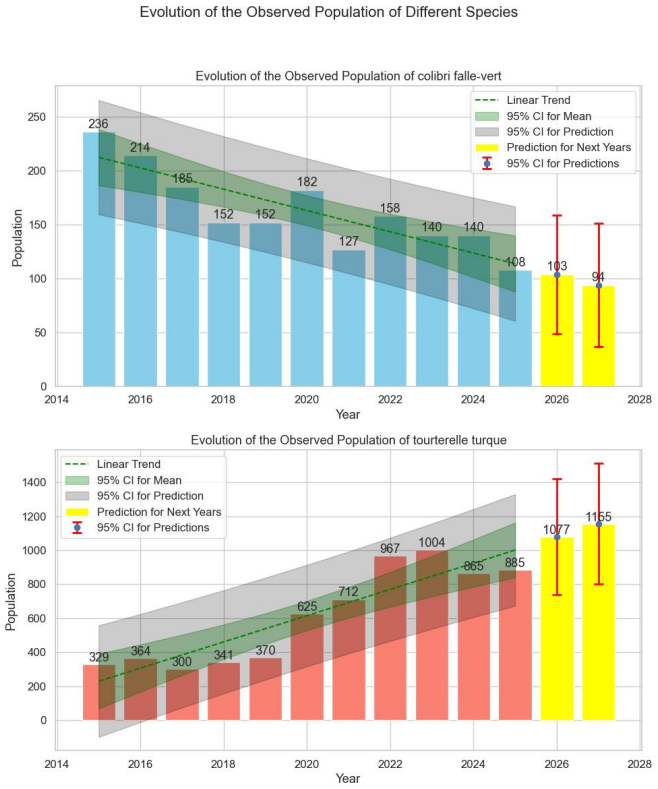

Figure 24: Evolution of the population of colibri falle-vert and tourterelle turque

We now study the distribution of these species' habitats. For this, we created a function that takes as a parameter the name of a species, and calculate the distribution of their habitats. We do this by iterating through **df_observations**, for each species and linking the column **N° point** to the correct site in the table **df_sites**, to get what type of habitat this is. For each type of habitat, the function maps how many specimens of this species have been observed in the habitat. Then by dividing each value by the sum of all the values, we get a normalized distribution of the habitats.

Figure 25 shows the distribution of the habitats of *tourterelle turque* and *colibri falle-vert*. We observe that the colibri falle-vert has a lot of habitat diversity, 27% live in *Périurbain*, 28% in *Agricole*, 17% in *Forêt humide* and 17% in *Forêt sèche*.

However for the tourterelle turque, almost half of them live in urban areas, 22% in *Agricole*, and 19% in *Périurbain*.

Observing the **diversity of habitats** and how species **evolve in different environmental contexts** could help us understand how environmental complexity influences species adaptation and population dynamics and reveals how flexible species are when facing **environmental changes**.

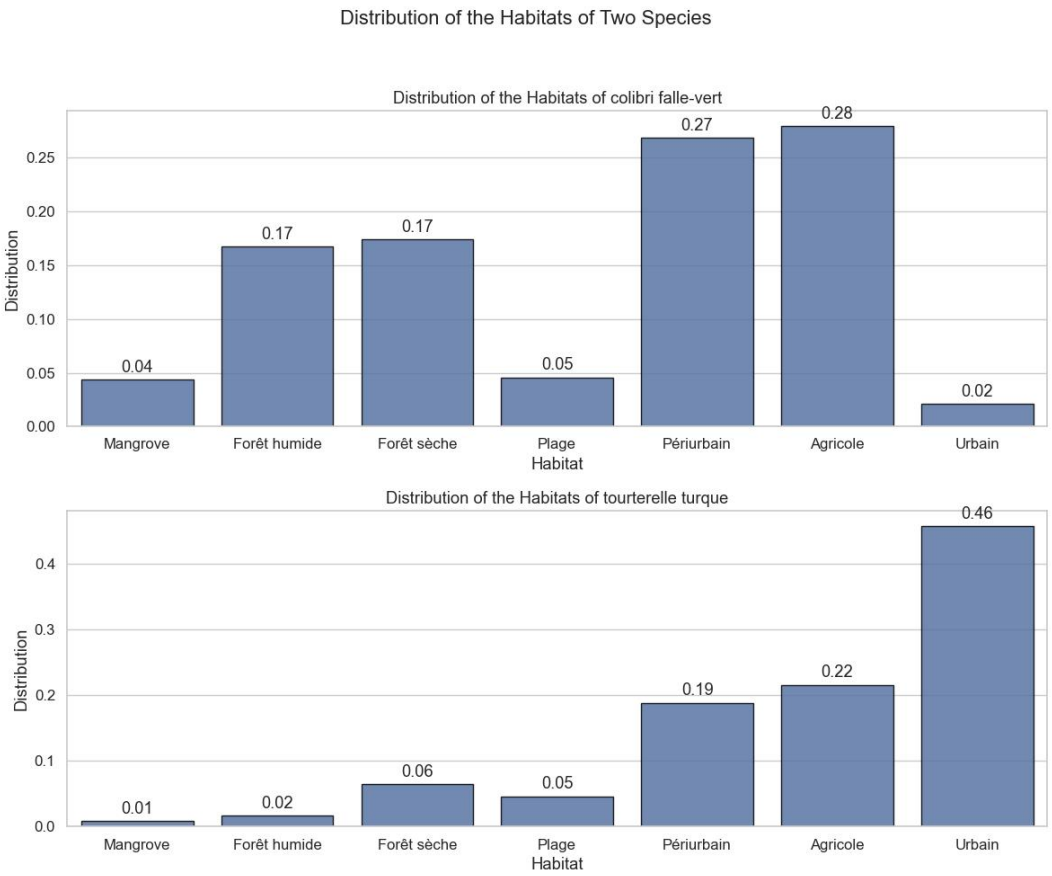

Figure 25: Distribution of the habitats of colibri falle-vert and tourterelle turque

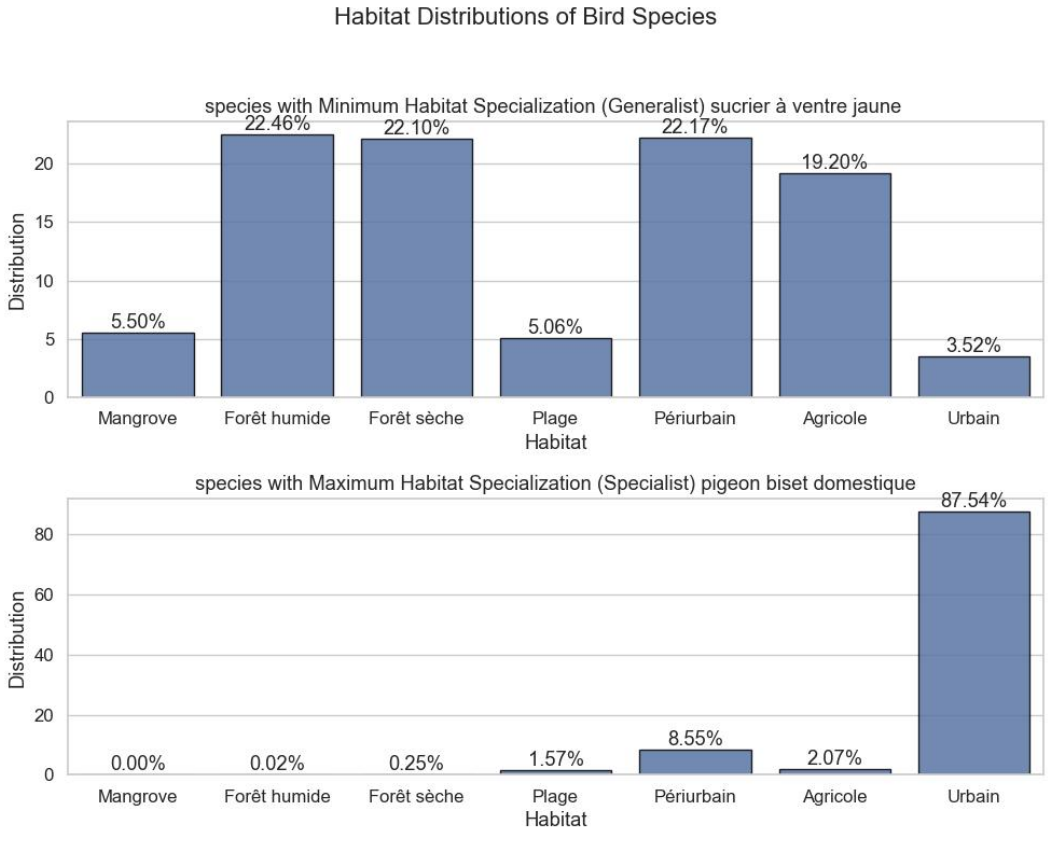

Figure 26: Distribution of habitats of sucrier à ventre jaune and pigeon biset domestique

3.2 Species with Least and Most Evenly Distributed Habitats

In this part, we study two species, one having the **lowest habitat diversity** and the other one having the **largest habitat diversity**. To do this, we compute the **distribution of the habitats** of every species and their **standard deviation (SD)**. We select the species with the **lowest diversity** as the species **whose SD is the highest**. For the species with the **largest habitat diversity**, we select the species with the **lowest SD**. These two species are respectively the **pigeon biset domestique** *(lowest diversity)* and the **sucrier à ventre jaune** *(largest diversity)*.

Figure 26 displays on the upper graph the distribution of habitats of the species that shows the most diverse habitats *(sucrier à ventre jaune)*, and on the lower graph the species that shows the least diverse habitats *(pigeon biset domestique)*.

The upper graph has a very centered habitats distribution (minimizing the SD). We notice that there are 3 habitats with a proportion of 22%, and another being close at 19%. Meanwhile, the lower graph has a distribution not centered at all (maximizing the SD). We see that the pigeon biset domestique lives at a 87% rate in urban areas, and in 8% of cases in *Périurbain*. This means that 95% of this species live in areas in or near cities.

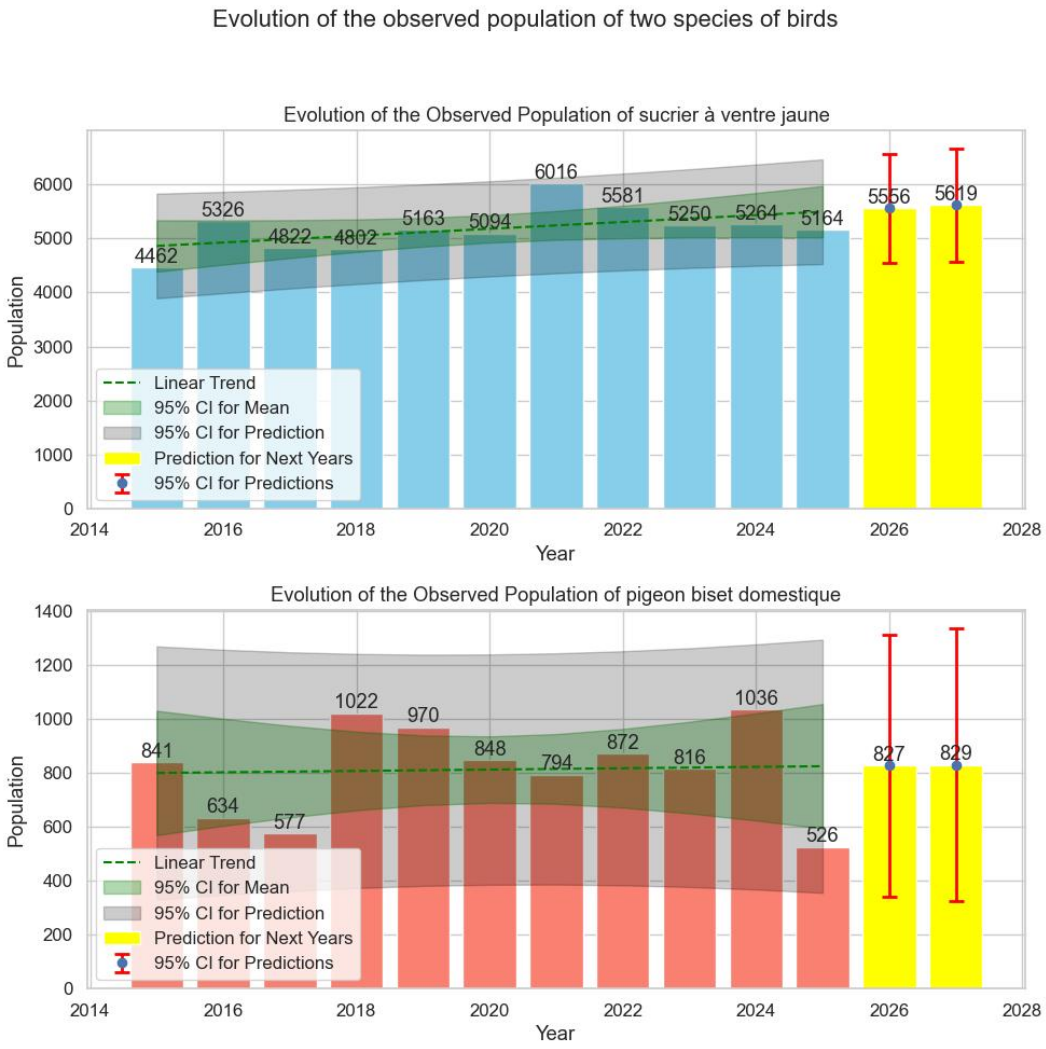

Figure 27: Evolution of the population of sucrier à ventre jaune and pigeon biset domestique

Figure 27 shows the evolution of the population of the species which have the most and least diverse habitat.

The linear trend suggests a slight increase in the population of sucrier à ventre jaune, but year-to-year fluctuations indicate that the trend remains uncertain, while the population of the pigeon biset domestique has a very noisy and inconsistent trend, making it hard to get a general trend. Our linear model predicts a flat trend, but with very large confidence intervals, therefore meaning it is really uncertain of the trend and that a linear model is not optimal here. However, the confidence interval for the sucrier à ventre jaune is quite normal and suggests that the linear model fits reasonably well in this scenario.

3.3 Non-linear Modeling Attempt: GAM

The linear model provided very **large confidence intervals**, indicating that it does not capture the temporal variability of ecological data. Bird populations fluctuate due to migration, reproduction cycles and local environmental disturbances, which makes the assumption of a linear trend unrealistic.

To address this, we fitted a **Generalized Additive Model (GAM)**, a non-linear regression method commonly used in ecology because it allows the response variable to **vary smoothly over time** without assuming a fixed trend shape. GAMs are **particularly suited for 1 time series** where fluctuations are driven by complex and unknown processes (seasonality, migration, climatic events). This method estimates:

$$Y = \beta_0 + f(\text{year})$$

where $f(\cdot)$ is a smooth and flexible function learned from the data and β_0 is the model intercept: it represents the expected value of the response variable when the smooth function equals zero. In practice, f is represented as a weighted sum of spline basis functions:

$$f(\text{year}) = \sum_{k=1}^{K} \theta_k \, B_k(\text{year})$$

where B_k are spline basis functions and the θ_k are coefficients estimated by penalized least squares. Unlike LOWESS (locally weighted regression), which fits *local* regressions independently around each point, GAM learns a *global smooth function* over the entire time series, ensuring consistency and reducing overfitting.

Unlike LOWESS, which only smooths the data locally, GAM is a true statistical model. It estimates a **parametric smooth function** that can be **extrapolated and extended** to include additional predictors (e.g., habitat, origin category, climatic variables), and unlike linear models, it does not force a straight-line and therefore **better captures fluctuations**. This makes GAM a more rigorous framework for prediction and hypothesis testing in ecological monitoring.

We applied the model to the four species previously studied. Figures 28 and 29 display the GAM trend and its 95% confidence band, along with the predictions for the next two years.

Overall, the GAM produces a smoother and more ecologically realistic trend than a linear regression, as it **does not force a straight-line relationship over time**. However, the uncertainty associated with future predictions remains very high. The confidence intervals of the GAM are still wide and not substantially smaller than those obtained with the linear model.

This is not a limitation of the model itself, but a consequence of the data: the time series contains only 12 yearly observations, and no additional explanatory variables (environmental conditions, habitat, sampling effort, etc.) are included in the model. With so few points and no covariates, the model cannot reliably estimate how populations will evolve, regardless of whether the trend is linear or non-linear.

Thus, while the GAM provides a more flexible and biologically plausible trend, it does **not improve predictive certainty**. This highlights that uncertainty is driven by **limited temporal data**, not by the choice of model.

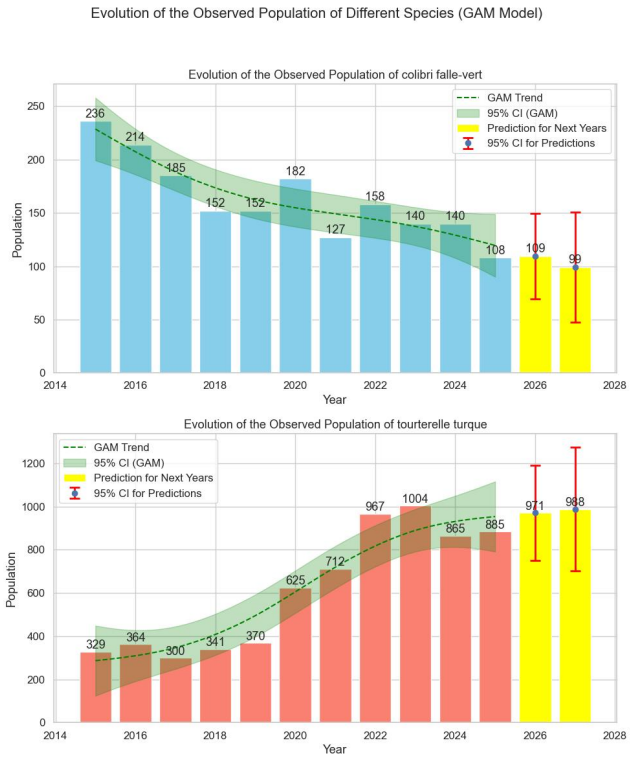

Figure 28: Evolution of the population of colibri falle-vert and tourterelle turque (GAM)

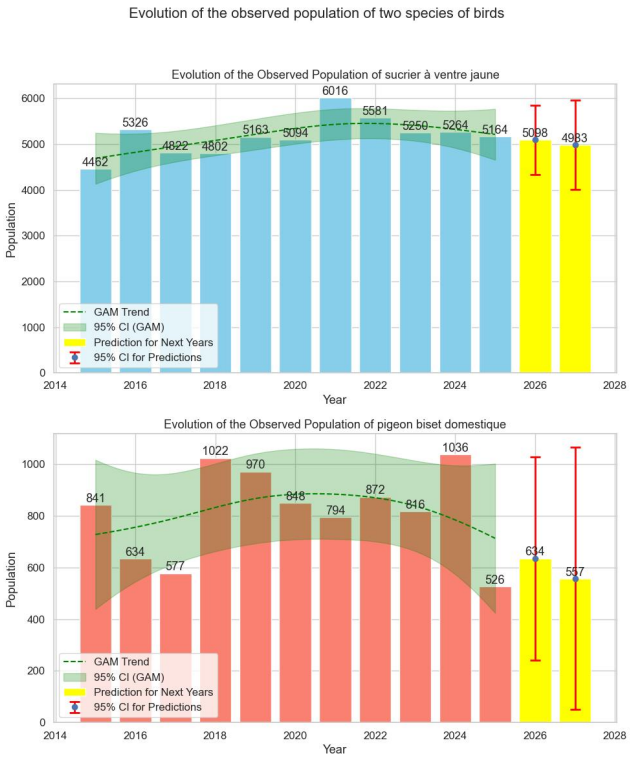

Figure 29: Evolution of the population of sucrier à ventre jaune and pigeon biset domestique (GAM)

4 Synthesis and Recommendations

4.1 Global Synthesis

The results from Indicators 1 and 2 (abundance, species richness, Shannon index, and origin proportions) show that the overall biodiversity of birds in Martinique appears to be in a **globally healthy state**. Most habitats show **stable or increasing diversity**, with **little evidence of dominance** by a small number of species. However, our analyses highlight **several points of vigilance**:

1. **Loss of native species in agricultural and peri-urban habitats.** Both habitats exhibit a continuous decrease in the proportion of native individuals and a corresponding increase of introduced species. This suggests ecological degradation and possible competitive pressure from exotic species.

2. **Shift in habitat use by migratory/marine species.** In Indicator 2, we observe that the proportion of migratory and marine species decreases in *Mangrove* while increasing in *Forêt sèche*. This shift may indicate a **change in habitat attractiveness or suitability**, potentially linked to environmental disturbances (e.g., coastal pressure, human activity, or habitat modification).

3. **Urban areas: increased diversity but loss of native integrity.** Normalized abundance and Shannon index both increase strongly in *Urbain*, indicating more birds and more evenly distributed species. However, the proportion of native individuals decreases steadily, meaning that diversity is increasingly driven by **introduced species**. In the species-level analysis, the species with the strongest population increase is the *tourterelle turque*, a predominantly **urban** and **introduced** species. This confirms that the improvement of diversity in urban areas reflects a process of **biotic homogenization**, not a recovery of native biodiversity.

4. **Recent decline in normalized abundance across most habitats.** While species richness and Shannon diversity remain stable, the normalized abundance shows a **decreasing trend in nearly all habitats over the last two years**. This means that the total number of individuals observed is declining, even though diversity metrics do not yet reflect this drop. This pattern can be an early warning signal of ecosystem degradation, where **populations decrease before diversity collapses**.

4.2 Monitoring Strategy

Indicator 3 shows that the **monitoring program is progressing well**: more transects are visited over the years, and visit intensity is becoming more evenly distributed across sites. This is a strong point of the project and should be pursued. We recommend:

- continuing to **increase the number of surveyed transects** when possible
- **improving balance in visit intensity** to avoid under-sampled areas
- **maintaining at least one transect in each habitat** category every year to ensure comparability

A consistent and evenly distributed sampling effort is essential to guarantee reliable biodiversity trends.

4.3 Modeling and Prediction Perspective

The linear model applied in Section 3 highlights an important methodological lesson: **linear regression rarely fits ecological time series**. Bird populations fluctuate due to migration, reproduction cycles, and inter-annual environmental variability. In addition, the dataset contains only a few annual observations (10–12 years), which makes any trend estimation statistically fragile. This leads to large confidence intervals and unstable predictions.

Even when using a **Generalized Additive Model (GAM)**, the uncertainty remained high. This is not a limitation of the GAM itself, but rather the result of two constraints: **very few time points**, and **no additional explanatory variables** were provided (e.g., habitat, sampling effort, climatic data).

More suitable approaches to improve predictive capacity include:

- **enriching GAM models with additional ecological covariates** (habitat type, origin category, sampling effort), to explain population fluctuations rather than fitting time only
- **Bayesian time-series models**, which naturally express uncertainty and are well suited for sparse ecological data.

These approaches do not force a linear trend and can better capture ecological dynamics when the dataset grows or when more relevant predictors are included.

4.4 Final Recommendations

- **Prioritize actions in agricultural and peri-urban habitats** where native species are declining.
- **Monitor urban areas closely** to prevent uncontrolled establishment of introduced species.
- Investigate why migratory/marine species show reduced presence in mangrove habitats.
- Maintain and **reinforce the sampling effort and spatial representativeness** of transects.

Overall, Martinique's bird biodiversity is in **good global condition**, but **several habitats exhibit signals** that justify increased ecological attention and proactive management.

GitHub: https://github.com/baptistepras/birds_biodiversity