



Université Claude Bernard



Lyon 1

Rapport de stage de M2 Bio-informatique :
Du 17/02 au 15/07

Application de méthodes de text-mining et
Natural Language Processing pour
l'extraction automatisée de données web
pour la création d'une base de données des
traits phénotypiques des adventices
tropicales

TRAVARD Jules

Encadrant : BLANCHARD Grégoire

2024-2025



Unité de recherche
Geco

Fonctionnement écologique
et gestion durable des
agrosystèmes bananiers et ananas

Résumé

Ce rapport présente le travail que j'ai effectué durant mon stage au CIRAD (Centre de coopération internationale en recherche agronomique pour le développement), plus précisément dans le département Persyst (Performances des systèmes de production et de transformation tropicaux), dans l'unité de recherche GECO (Fonctionnement écologique et gestion durable des agrosystèmes bananiers et ananas).

Ma mission a été de développer une base de données des traits phénotypiques et morphologiques des adventices tropicales en utilisant des techniques de web scraping et de text mining. Les adventices sont des plantes poussant spontanément dans les cultures, qui peuvent concurrencer les cultures mais aussi jouer un rôle écologique en protégeant les sols et en favorisant la biodiversité. Elles restent peu étudiées dans les régions tropicales et les informations disponibles quant à leur caractérisation sont assez éparses malgré leur importance pour optimiser la gestion de ces plantes et leur intégration dans des pratiques agroécologiques durables. L'objectif principal de ce stage était d'extraire, structurer et analyser des informations provenant de différentes sources en ligne afin de mieux comprendre les caractéristiques et la diversité des adventices tropicales présentes dans les cultures.

Pour ce faire, j'ai utilisé l'outil *FloraTraiter* pour identifier et extraire des traits phénotypiques et morphologiques à partir de textes issus de bases de données botaniques et scientifiques, notamment *Wiktrop*, *WorldFloraOnline*, *Wikipedia*, *Tropical.theferns*, *Feedipedia* et la base de données TRY. J'ai ensuite testé une approche alternative en utilisant une méthode inspirée du projet "AutomaticTraitExtraction" et en intégrant l'IA *Le Chat* pour affiner l'extraction et l'interprétation des données. Ces informations ont ensuite été organisées dans une base de données *SQLite*, permettant d'effectuer des analyses descriptives et d'évaluer l'homogénéité et la qualité des données récoltées.

Ce stage m'a permis de développer des compétences en programmation, en gestion de bases de données, en fouille de texte et en analyse de données. Il m'a également donné une vision plus approfondie des problématiques écologiques et agronomiques liées aux adventices tropicales et de l'intérêt d'une approche interdisciplinaire pour la gestion durable des agrosystèmes.

Remerciements

Je tiens à remercier toute l'équipe du CIRAD de la Station Neufchâteau en Guadeloupe, et plus particulièrement l'unité de recherche GECO. Leur accueil et leur disponibilité m'ont permis de travailler dans d'excellentes conditions et de m'investir pleinement dans ce stage.

Un remerciement particulier à Grégoire Blanchard, pour son encadrement et ses précieux conseils tout au long de mon stage. Merci également à Sandrine Auzoux, Aude Ripoche, Karim Barkaoui et Mathieu Roche pour leurs échanges enrichissants et leur appui scientifique.

Enfin, je remercie l'ensemble des chercheurs, techniciens, doctorants et stagiaires du CIRAD que j'ai pu côtoyer.

Tables des matières

Liste des abréviations.....	4
Logiciels utilisés.....	4
Table des figures.....	5
1. Structure d'accueil.....	6
2. Introduction.....	7
3. Matériel et Méthodes.....	9
3.1 Présentation des sources de données.....	9
3.2 Méthodes de collecte des données.....	10
3.3 Extraction et traitement des traits.....	11
3.4 Structuration et stockage des données.....	15
3.5 Analyse des données.....	16
4. Résultats.....	18
4.1 Collecte et extraction des données.....	18
4.2 Validation et qualité des données.....	20
4.3 Analyse des traits extraits.....	22
5. Discussion.....	28
6. Conclusion.....	29
7. Bibliographie.....	30
8. Annexes.....	31

Liste des abréviations

CIRAD : Centre de coopération internationale en recherche agronomique pour le développement

Persyst : Performances des systèmes de production et de transformation tropicaux

GECO : Fonctionnement écologique et gestion durable des agrosystèmes bananiers et ananas

IA : Intelligence Artificielle

LLM : Grand modèle de langage (Large Language Model)

NLP : Traitement automatique du langage naturel (Natural Language Processing)

SLA : Surface foliaire spécifique (Specific Leaf Area)

ACP : Analyse en Composantes Principales

NER : Named Entity Recognition

Logiciels utilisés

- R studio : version 2024.12.1+563
- R : version 4.4.3
- Visual Studio Code : version 1.98.0 et 1.98.1
- DBeaver : version 25.0.1
- GitBash : version 2.48.1
- Le Chat de Mistral : modèle "mistral-medium-latest"

Table des figures

Figure 1 : Schéma récapitulatif du pipeline de collecte, traitement et analyse des traits fonctionnels des adventices tropicales

Figure 2 : Les 30 traits présentant le plus grand nombre de correspondances exactes dans leurs valeurs entre celles extraites via LLM et via *FloraTraiter*, comparé au nombre total de valeurs de traits extraites par LLM (ex: pour le premier trait : leaf_length_mean, 1249 valeurs de traits ont été extraites via LLM, dont 616 correspondent exactement à celles extraites via *FloraTraiter*).

Figure 3 : Comparaison visuelle des vérifications manuelles avec les prédictions des deux méthodes (NLP et LLM). Cor: correct, inc: incorrect, mis: missing (valeur manquante), par: partiel, spu: spurious (invention, i.e. valeur générée par le model en l'absence de valeur réelle)

Figure 4 : Précision, Rappel et F1-Score des prédictions par trait pour LLM et *FloraTraiter*

Figure 5 : Répartition des valeurs pour les traits de dispersion, durée de vie, forme de vie et support selon les 2 méthodes d'extraction (LLM et *FloraTraiter*).

Figure 6 : Distribution logarithmique des valeurs de quatre traits fonctionnels extraites via la méthode basée sur LLM (L'utilisation de l'échelle logarithmique permet ici de mieux visualiser la large variabilité des valeurs)

Figure 7 : Nombre de traits différents extraits par espèce - LLM

Figure 8 : Nombre de traits extraits par mot-clé (à gauche) et nombre de traits différents (chaque trait comptés qu'une seule fois) par mot-clé (à droite) - méthode basée sur le LLM.

Figure 9 : 30 traits les plus fréquemment trouvés au moins une fois par espèce - LLM

Figure 10 : Analyse en composantes principales (ACP) des valeurs moyennes de différents traits fonctionnels quantitatifs des espèces (tailles des fruits, des plantes, des feuilles et des graines) selon les formes de vie végétale (a, b), la durée de vie (c, d) et le nombre de cotylédons (e, f), à gauche sans le SLA (Surface Foliaire Spécifique), à droite avec le SLA inclus (ce trait n'étant disponible que pour certaines espèces, réduisant le nombre d'espèces analysées). Les ellipses visibles sur les graphiques délimitent les regroupements d'espèces partageant une même modalité pour le trait catégoriel analysé. Elles permettent de visualiser la dispersion et le positionnement fonctionnel des groupes dans l'espace réduit des deux premières composantes principales. - Données issues de la méthode basée sur le LLM.

Figure 11 : Représentation en ACP des espèces d'intérêts et des espèces de la base de donnée TRY selon les traits fonctionnels - Données extraites par la méthode basée sur le LLM

1. Structure d'accueil

J'ai effectué mon stage au CIRAD (*Cirad, la recherche agronomique pour le développement durable des régions tropicales et méditerranéennes. | Cirad*), au sein de l'unité de recherche GECO, qui fait partie du département Persyst. Le CIRAD est un organisme français de recherche agronomique qui se consacre au développement durable des régions tropicales et méditerranéennes. Le CIRAD mène des projets collaboratifs dans de nombreux pays à travers le monde, notamment en Afrique, en Amérique latine, en Asie et dans les Caraïbes, avec une approche interdisciplinaire combinant agronomie, écologie, économie et sciences sociales.

L'unité GECO (*GECO - Fonctionnement écologique et gestion durable des agrosystèmes bananiers et ananas*) se spécialise dans l'étude du fonctionnement écologique des agrosystèmes tropicaux et la gestion durable des cultures, en particulier celles de la banane et de l'ananas. Son objectif est d'analyser les interactions entre les pratiques agricoles, la biodiversité et la fertilité des sols afin de proposer des solutions innovantes et durables. Les travaux de l'unité incluent la modélisation des dynamiques écologiques, la gestion intégrée des adventices, ainsi que l'étude des services écosystémiques rendus par les plantes et les micro-organismes du sol. Les axes de recherche de l'unité sont les suivants :

- Dynamique des bioagresseurs et des communautés en conditions d'intensification écologique : étude des interactions entre les bioagresseurs et les communautés biologiques dans des contextes d'intensification écologique des systèmes agricoles.
- Diversité végétale et services écosystémiques dans les systèmes cultivés : analyse de la diversité végétale et des services écosystémiques qu'elle procure au sein des agrosystèmes.
- Conception et évaluation de systèmes de cultures innovants et durables : développement et évaluation de systèmes de culture visant à améliorer la durabilité environnementale, économique et sociale des systèmes agricoles.

Mon stage s'inscrit directement dans le deuxième axe de recherche de l'unité (Diversité végétale et services écosystémiques dans les systèmes cultivés). Cet axe vise à comprendre comment la diversité des plantes, notamment des adventices, peut contribuer au bon fonctionnement des agrosystèmes tropicaux.

2. Introduction

Ce stage de fin d'étude a pour principal objectif de développer et de tester des méthodes d'extraction de données afin de constituer une base de données des traits phénotypiques et morphologiques des adventices tropicales (c.a.d. les espèces de plantes s'établissant de manière spontanée dans les cultures tropicales). Ces plantes, communément appelées mauvaises herbes, sont souvent perçues négativement bien qu'elles puissent jouer un rôle clé dans les écosystèmes agricoles. En milieu tropical, leur gestion constitue un enjeu particulièrement complexe car les conditions climatiques (absence de saison froide, chaleur constante et forte humidité) favorisent une croissance rapide et une production importante de biomasse. Cela implique des interventions fréquentes pour limiter leur développement (Oerke, 2006), qu'elles soient mécaniques ou chimiques, et soulève des problématiques environnementales et sanitaires majeures, notamment liées à l'usage intensif d'herbicides (Côte et al., 2025). Dans ce contexte, le réchauffement climatique risque d'accentuer encore ces dynamiques en renforçant les conditions favorables à la prolifération des adventices. (Anwar et al., 2021)

Depuis maintenant une trentaine d'années, l'étude des traits fonctionnels des plantes contribue à améliorer notre compréhension du fonctionnement et de la diversité des communautés végétales au sein des écosystèmes. Ces traits sont des caractéristiques morphologiques, physiologiques ou phénologiques mesurables à l'échelle de l'individu, qui influencent directement ou indirectement la performance des espèces en fonction de leur environnement (notamment la croissance, la reproduction et la survie (Violle et al., 2007)). Ils reflètent la manière dont les plantes accèdent aux ressources, interagissent avec leur environnement, et affectent le fonctionnement des écosystèmes. Nous distinguons généralement les traits de réponse, qui traduisent la capacité des plantes à faire face aux conditions environnementales, et les traits d'effet, qui décrivent leur influence sur les processus écologiques (ex: compétition avec les autres espèces) et sur l'environnement. En permettant de comparer les espèces au-delà de leur taxonomie, l'approche fonctionnelle offre une vision simplifiée et généralisable des rôles écologiques des plantes, indépendamment du contexte de biodiversité. Dans le cas des adventices, analyser leurs traits fonctionnels permet de mieux comprendre les processus écologiques qui déterminent leur composition et leur dynamique dans les agroécosystèmes. Cela permet également d'évaluer leur potentiel à devenir envahissantes ou non (Palma et al., 2021). En identifiant les traits associés aux réponses des espèces à l'environnement locale et aux pratiques agricoles, ou à leurs effets sur les plantes cultivées, cette approche permet d'évaluer le potentiel des espèces à fournir des services ou des disservices écosystémiques. Certaines adventices peuvent, par exemple, contribuer à la fertilité des sols, réduire leur érosion ou soutenir la biodiversité en fournissant des habitats pour certaines espèces (comme certains oiseaux ou insectes) (Gaba et al., 2017) ou encore en favorisant des ennemis naturels des ravageurs des cultures. D'un autre côté, les adventices peuvent aussi représenter une menace pour les cultures en entrant en compétition pour la lumière, l'eau et les nutriments du sol, ou en tant qu'hôtes de bioagresseurs. L'approche fonctionnelle ne vise donc pas uniquement à classer les adventices comme bénéfiques ou nuisibles, mais à mieux comprendre la gamme de leurs interactions avec les agroécosystèmes, selon le contexte spécifique.

Cependant, l'étude de l'écologie des adventices tropicales via une approche fonctionnelle est limitée par l'absence de base de connaissances solides, standardisées et centralisées, qui permettrait à la fois aux chercheurs de bénéficier d'un accès rapide et généralisé à des données fiables, et d'effectuer facilement des analyses ou des requêtes en fonction d'objectifs scientifiques ou techniques particuliers. D'une part, les connaissances sur les traits fonctionnels des adventices tropicales restent encore lacunaires. D'autre part, de nombreuses données potentiellement utiles concernant ces connaissances existent, mais

elles sont largement dispersées dans différentes sources, souvent sous forme de texte non structuré et organisées de manière hétérogène, ce qui rend leur exploitation difficile. C'est à cette problématique que ce stage cherche à répondre.

Ce projet s'inscrit dans une démarche d'innovation scientifique visant à mieux caractériser les adventices tropicales à partir d'informations disponibles en ligne, en exploitant des méthodes modernes d'extraction automatique de données. Ces dernières années, les progrès dans le traitement automatique du langage naturel (NLP) et l'émergence des grands modèles de langage (LLM) ont ouvert de nouvelles perspectives pour analyser des textes non structurés, en particulier dans le domaine de la science du vivant. Le web scraping permet d'extraire automatiquement des contenus depuis des pages web, qu'il s'agisse de textes structurés (fiches espèces ou tableaux) ou non structurés (descriptions narratives). Une fois ces textes récupérés, les méthodes de text mining et de NLP (notamment la bibliothèque SpaCy qui est un outil open-source adapté pour l'extraction d'informations depuis de grands corpus de texte) peuvent être mobilisées pour détecter et structurer les informations pertinentes. Cela inclut des techniques classiques comme l'analyse syntaxique, la reconnaissance d'entités nommées (NER, Named Entity Recognition) (Farrell et al., 2022) ou la détection de patrons linguistiques. D'autre part, il existe aussi des approches plus récentes fondées sur des modèles intégrant des algorithmes relatifs à l'Intelligence Artificielle (IA), pré-entraînés sur de très grandes bases de données et capables de comprendre le contexte des phrases (LLM, Large Language Models). Bien que généralistes, ces LLM permettent de développer des stratégies d'extraction d'information spécifiques et contextuelles, basées sur le prompting qui consiste à formuler des instructions en langage naturel pour guider l'IA dans l'interprétation du texte (Marcos et al., 2025). Dans le domaine de la biologie, ces techniques plus souples sont aujourd'hui utilisées pour extraire des informations à partir d'articles scientifiques, de bases de données naturalistes ou de documents taxonomiques. Plusieurs études ont montré leur efficacité dans le domaine biomédical (Farrell et al., 2022), de la génomique (Oikonomou et al., 2024) ou de la biodiversité (Thessen et al., 2012). Ce type d'approche est notamment exploité dans des projets comme *Automatic Trait Extraction* (Marcos et al., 2025), où des LLM sont utilisés pour extraire des traits fonctionnels spécifiques à partir de descriptions textuelles complexes. Dans le cas particulier des plantes, l'extraction automatique de traits fonctionnels à partir de descriptions botaniques constitue un défi spécifique, car ces traits sont souvent exprimés de manière hétérogène et contextuelle dans le texte.

Dans ce travail, deux approches ont été explorées dans ce projet pour répondre à ce défi. La première repose sur *Flora Traitier* (Folk et al., 2024), un pipeline basé sur des règles linguistiques et des listes de traits pré-établies, conçu pour extraire automatiquement les traits phénotypiques depuis des descriptions botaniques. La seconde approche mobilise un LLM, en l'occurrence *Le Chat* de *Mistral AI* (*Le Chat - Mistral AI*), capable d'analyser des textes complexes et d'en extraire des traits ciblés, y compris des valeurs numériques, des unités, ou des modalités descriptives, à partir d'un prompt guidé. Grâce à ces techniques, l'objectif de ce travail était de collecter, structurer et analyser automatiquement des traits morphologiques et fonctionnels des adventices tropicales à partir de descriptions textuelles issues de différentes sources en ligne, dans le but de les intégrer dans une base de données exploitable par les chercheurs et facilitant leur analyse. Dans ce rapport, nous détaillerons la méthodologie mise en place, les outils et techniques employés, ainsi que les résultats obtenus. Enfin, nous discuterons des défis rencontrés, des perspectives d'évolution de ce type d'approche, ainsi que de son intérêt pour améliorer la compréhension de l'écologie des adventices tropicales et faciliter leur gestion dans les agroécosystèmes.

3. Matériel et Méthodes

3.1 Présentation des sources de données

Ce travail s'est appuyé sur différentes sources de données en ligne pour l'extraction de données sur les traits fonctionnels des plantes. Premièrement, nous avons défini une liste d'espèces adventices tropicales (voir **Annexe 10**) à partir de l'intégralité des espèces référencées sur la plateforme *Wiktrop* (1049 espèces). Ensuite, nous avons sélectionné, en plus de *Wiktrop*, un certain nombre de sources en ligne regroupant des données descriptives des espèces de plantes en lien avec l'objectif d'en extraire de l'information sur les traits fonctionnels des adventices. Les sources mobilisées étaient les suivantes:

- *Wiktrop (WIKTROP - Weed Identification and Knowledge in the Tropical and Mediterranean Areas)* est une plateforme dédiée à l'identification et à la connaissance des adventices des régions tropicales et méditerranéennes. Elle fournit des descriptions botaniques et écologiques détaillées, utilisées ici comme source principale.
- *World Flora Online (WFO - The World Flora Online)* est une base collaborative internationale regroupant des descriptions botaniques issues de multiples institutions, utilisée pour compléter les informations morphologiques.
- *Feedipedia (List of feeds | Feedipedia)* propose des fiches sur les plantes utilisées en alimentation animale, incluant souvent des descriptions morphologiques utiles à ce projet.
- *Tropical (Useful Tropical Plants)* fournit des données sur les plantes tropicales, avec un accent sur leurs usages. Ses descriptions détaillées ont permis d'extraire divers traits phénotypiques.
- *Wikipedia* (« Wikipédia », 2024) a été utilisée en tant que source complémentaire. Bien que variable en fiabilité, elle a permis de recouper ou valider des informations.
- *RTaxRef (An R Client for TaxRef the French Taxonomical Database)* a permis de récupérer la taxonomie complète des espèces présentes dans la base de données, notamment les familles, ordres et classes.
- *TRY (TRY Proposal Submission)* est une base de référence pour les traits fonctionnels des plantes à l'échelle mondiale. Elle a été utilisée pour compléter ou comparer les données issues de l'extraction textuelle.

3.2 Méthodes de collecte des données

Pour extraire les descriptions des espèces végétales à partir des différentes sources, nous avons développé un script spécifique pour chacune d'elles. Bien que la structure générale des scripts soit similaire, certaines adaptations ont été nécessaires en raison des spécificités de chaque source. Nous allons prendre en exemple les scripts pour *Wiktrop* et pour *World Flora Online*.

3.2.1 Collecte des données depuis Wiktrop

Le script dédié à *Wiktrop* commence par parcourir une liste d'espèces d'intérêt extraite depuis le site directement via un script en java. Pour chaque espèce, une requête HTTP est envoyée pour récupérer la page HTML correspondante. Ensuite, nous avons utilisé la bibliothèque *BeautifulSoup* pour analyser le code source et extraire les descriptions en ciblant certains mots qui nous ont servi de balise, ce qui nous a permis de sélectionner les informations d'intérêt. Toutefois, cette approche a présenté plusieurs difficultés. Les noms des balises n'étaient pas toujours uniformes d'une page à l'autre. Certains champs comportaient des fautes de frappe, des variations dans leur nom ou encore des tabulations et espaces supplémentaires, ce qui compliquait leur identification. Pour pallier cela, nous avons dû affiner le traitement des balises en intégrant des techniques de nettoyage et de gestion des erreurs.

3.2.2 Collecte des données depuis Word Flora Online

L'extraction des données depuis *World Flora Online* suit une méthode similaire à celle employée pour *Wiktrop*, mais avec certaines particularités. Étant donné que le site regroupe plusieurs sources pour une même espèce, nous avons dû extraire et stocker chaque description séparément dans des fichiers distincts. Une difficulté majeure a été l'identification des noms des sources, qui présentaient des variations et des incohérences, rendant leur extraction plus complexe. De plus, les descriptions étaient disponibles en plusieurs langues, ce qui nécessitait une harmonisation des données. Pour y remédier, nous avons développé un second script, capable de détecter automatiquement la langue de chaque texte et, si nécessaire, de le traduire en anglais à l'aide du package *deep_translator* qui contient *GoogleTranslator*. Cette étape a permis d'assurer une cohérence linguistique à l'ensemble des données collectées.

3.3 Extraction et traitement des traits

3.3.1 Extraction des valeurs de traits : approche par NLP

Une fois les textes extraits des différentes sources, nous avons utilisé l'outil *FloraTraiter*, pour identifier et structurer les traits fonctionnels des adventices. *FloraTraiter* est un pipeline d'extraction de traits conçu pour analyser des descriptions botaniques et en extraire automatiquement des informations pertinentes sous forme de fichiers JSON. Son fonctionnement repose sur le trait parsing, une approche qui segmente les textes en unités linguistiques et applique des règles pour identifier les éléments clés associés aux traits des plantes. Plus précisément, *FloraTraiter* utilise une méthode d'analyse hybride, combinant des règles et des méthodes statistiques. Cette approche s'appuie fortement sur la bibliothèque de traitement du langage naturel *SpaCy*, qui fournit à la fois des modèles statistiques de catégorisation grammaticale (POS) et un cadre flexible pour l'élaboration de règles personnalisées adaptées aux textes scientifiques.

L'utilisation de *FloraTraiter* est relativement simple, une seule commande permet de traiter un ensemble de descriptions textuelles situées dans un répertoire spécifique et de générer un fichier HTML et des fichiers JSON contenant les résultats :

```
parse-treatments --treatment-dir ./input --json-dir ./output --html-file ./output/traits.html  
--encoding utf-8
```

L'outil détecte les traits en s'appuyant sur une liste de termes stockés dans différents fichiers CSV. Cependant, ces fichiers ne contenaient pas tous les termes nécessaires pour identifier correctement certains traits phénotypiques spécifiques aux adventices tropicales. Il a donc été nécessaire d'enrichir ces listes en ajoutant des termes supplémentaires afin d'améliorer la détection des traits recherchés. Un autre point important est que *FloraTraiter* associe les termes d'une phrase à leur contexte, ce qui peut poser problème dans certains cas. Par exemple, si une valeur est mentionnée dans une phrase mais que le nom du trait correspondant n'y est pas explicitement indiqué, le programme va chercher cette information dans les phrases adjacentes (au-dessus ou en dessous). Cela peut parfois conduire à attribuer une valeur à un mauvais trait, créant ainsi des erreurs d'interprétation.

Après l'exécution de *FloraTraiter*, les données extraites étaient disponibles sous format JSON. Afin de les exploiter plus facilement, nous avons transformé ces fichiers en tableaux CSV à l'aide de *R*, permettant ainsi une analyse plus approfondie des traits. Cependant, en raison des limites mentionnées précédemment, nous avons dû envisager une seconde méthode plus souple et précise pour extraire les traits. Pour cela, nous avons choisi d'utiliser l'IA *Le Chat* de Mistral.

3.3.2 Extraction des valeurs de traits : approche par LLM

Nous avons constaté que l'outil *FloraTraiter* ne parvenait pas toujours à extraire correctement les traits recherchés. Certains éléments étaient mal identifiés, notamment dans les cas de descriptions longues ou de formulations complexes, et l'outil montrait des lacunes pour certains types de traits, en particulier les traits numériques.

Pour améliorer la qualité et l'efficacité de l'extraction, nous nous sommes inspirés d'une autre méthode : celle proposée dans le projet *AutomaticTraitExtraction* (Marcos et al., 2025). Cependant, cette approche imposait de fournir en amont une liste exhaustive des traits à détecter, avec toutes les valeurs possibles pour chacun. La méthode renvoyait alors un tableau où chaque colonne représentait un trait, et chaque valeur était binaire (0 ou 1) selon la présence ou non du trait dans le texte. Ce fonctionnement était trop rigide pour notre cas d'étude, car il ne permettait pas d'extraire des valeurs numériques ou des expressions complexes (ex. "hauteur entre 30 et 60 cm", "fleurs de 5 mm de diamètre", etc.) qui ne peuvent pas être dans une liste fixe.

Pour contourner ces limitations, nous avons donc mis en place une approche plus souple en exploitant l'IA *Le Chat*, un grand modèle de langage (LLM) développé par l'entreprise française *Mistral AI* utilisable selon une interaction conversationnelle ("prompt"). Cette IA est basée sur un ensemble de modèles intégrant de nombreux paramètres, la rendant capable d'analyser et d'extraire dynamiquement des informations depuis des textes complexes. Nous avons utilisé cette approche via un script *Python* (voir **Annexe 1**). Ce script lit les fichiers texte contenant les descriptions d'espèces, les soumet un par un à l'IA via une requête API, et reçoit en retour un tableau structuré au format CSV, dans lequel chaque ligne correspond à un trait identifié (nom du trait, valeur, unité éventuelle). Nous avons construit un prompt explicite, qui guide l'IA pour qu'elle transforme la description brute de chaque espèce extraite de chaque source en une table de traits composée de trois colonnes : le nom du trait, sa valeur et son unité.

Pour obtenir un prompt efficace, plusieurs essais ont été nécessaires. Nous avons constaté que l'ordre des phrases dans le prompt influençait fortement le résultat : les instructions placées en fin de prompt étaient généralement mieux prises en compte par l'IA. D'autre part, bien que nous ayons fourni une liste précise de traits à extraire, l'IA avait tout de même tendance à détecter et structurer d'autres traits non demandés, ce qui est à la fois un avantage (enrichissement potentiel des données) et un inconvénient (nécessité de filtrer

ou de contrôler les données). Cette méthode présente ainsi plusieurs avantages : elle est flexible, car elle ne nécessite pas de liste prédéfinie de traits ou de valeurs ; elle est adaptée aux traits numériques, qui sont extraits avec leurs unités et plages de variation ; et elle permet un traitement automatisé à grande échelle. Cette approche permet d'obtenir un format de tableau structuré, facilitant leur intégration dans une base de données pour les analyses ultérieures.

Dans le cadre de l'extraction des traits fonctionnels, les dimensions des feuilles représentent une information particulièrement utile, car elles sont directement liées à des fonctions écologiques majeures (surface photosynthétique, compétition pour la lumière) (Díaz et al., 2016). Afin d'exploiter au mieux les données de longueur et de largeur extraites par l'IA, nous avons souhaité estimer l'aire des feuilles et des folioles de manière standardisée. Pour cela, nous nous sommes appuyés sur les formules proposées dans l'article "*Leaf size estimation based on leaf length, width and shape*" (Schrader et al., 2021), qui propose des ratios multiplicateurs (voir la formule dessous du facteur de corrélation CF, calculé à partir du rapport entre l'aire mesurée des feuilles et le produit longueur × largeur, à partir de la banque de données d'images de feuilles) dépendant de la forme de la feuille (par exemple : elliptique, obovale, lancéolée...). L'aire a été estimée en multipliant ce ratio par la longueur et la largeur des feuilles.

$$CF = \frac{\text{leaf size}}{\text{leaf length} * \text{leaf width}}.$$

Pour déterminer la forme des feuilles, nous avons réutilisé le même script *Python* que celui utilisé précédemment, en modifiant le prompt pour qu'il renvoie uniquement la forme de la feuille ou de la folioles (parmi 7 formes provenant de (Schrader et al., 2021)). Nous avons ensuite lié chaque forme au ratio correspondant, et lorsqu'une espèce était associée à plusieurs formes, nous avons pris la moyenne des ratios. Cependant, une difficulté a été rencontrée à cette étape : le LLM renvoyait très souvent les formes associées aux folioles, même lorsqu'il s'agissait de feuilles simples. Pour résoudre ce problème, nous avons généré une liste d'espèces ayant des feuilles composées, grâce au trait "compounds" qui a comme valeur "yes" ou "no", et nous avons utilisé cette liste pour corriger manuellement le dataframe des formes : si une espèce n'était pas dans cette liste, sa forme était alors interprétée comme celle de la feuille, et non des folioles. Cette vérification a permis d'améliorer considérablement la précision des calculs d'aires.

Enfin, cette même méthode a été utilisée pour extraire les formes de vie des plantes. Un prompt spécifique a été conçu pour extraire de manière indépendante les valeurs "yes"/"no" pour chacun des types de formes de vie parmi une liste de 15 types (comme parasitic, aquatic, climbing, tree, shrub, etc...), et une valeur spécifique pour "cotyledon_number" ("monocotyledonous" / "dicotyledonous"). Par ailleurs, pour compléter les données manquantes sur les formes de vie (notamment pour les types tree, shrub, graminoid et forb sur lesquels nous nous sommes focalisés pour les analyses de la base de donnée), une approche itérative a été menée sur un ensemble de 108 espèces pour lesquelles l'information n'avait pas initialement été récupérée. Nous avons utilisé *Le Chat* avec un prompt ciblé sur cette liste d'espèces, afin d'extraire uniquement cette information et de la réintégrer dans la base de données. Cette étape a permis d'améliorer à la fois la proportion des espèces pour lesquelles les traits ont été extraits et de limiter les erreurs dans ces extractions.

3.3.3 Traitement

Après l'extraction des traits phénotypiques à partir des différentes sources et leur structuration sous forme de tableaux, une phase de tri et de nettoyage a été nécessaire afin

de ne conserver que les informations pertinentes pour notre étude. En effet, certains tableaux contenaient un très grand nombre de noms de traits, dont certains pouvaient être non pertinents pour relayer des informations sur l'écologie des espèces, ou bien pouvaient présenter une incertitude quant à leur définition. Nous avons donc choisi d'effectuer une sélection de traits avec une terminologie claire et fiable, et considérés comme écologiquement pertinents pour le reste de notre étude.

De même, dans les tableaux issus des autres sources, nous avons sélectionné un ensemble de traits correspondant aux caractéristiques phénotypiques les plus intéressantes pour l'étude des adventices tropicales (voir **Annexe 2**). Les traits retenus incluent :

- Les dimensions (longueur, épaisseur, largeur, diamètre, hauteur) des fruits (akène, capsule, samares, arille, baie, caryopse, drupe), tiges, feuilles, folioles, limbes, pétioles, graines, troncs, arbres, arbustes et racines.
- Les caractéristiques générales de la plante : type de support, mode de reproduction, mode de dispersion, durée de vie des feuilles, stolon, racine, corne, tubercule, rhizome, durée de vie de la plante, ligneux.

D'autre part, la base TRY recensait 725 traits différents pour la liste d'espèces adventices considérées dans notre étude, parmi lesquels nous avons retenu 63 traits directement correspondant ou présentant un lien fort avec la liste de trait jugés pertinents définie précédemment.

En plus de cette sélection, nous avons effectué un nettoyage et une uniformisation des noms de traits récoltés, certains étant formulés de manière inhabituelle ou inconsistante selon les sources. Ce traitement a été réalisé grâce au logiciel *R*, utilisant notamment le package *dplyr* pour le filtrage, la sélection et la transformation des données. Par exemple, pour les traits relatifs aux feuilles ("leaf", "lamina"), un script a permis d'isoler les valeurs maximales et minimales recensées pour la longueur et/ou la largeur des différents organes (feuilles, fruits, graines, tige) par espèce. Les noms des traits ont ensuite été renommés de manière standardisée ("LeafLength_High", "LeafLength_Low", etc...). Cette étape de standardisation a permis d'harmoniser les données et de faciliter leur intégration dans la base de données. Nous avons également calculé la moyenne des valeurs numériques pour les traits sélectionnés, afin d'obtenir une représentation synthétique et plus exploitable de certaines mesures continues.

Concernant l'extraction des formes de vie, comme lors des précédentes, nous avons rencontré des problèmes de cohérence dans les valeurs retournées par l'IA : certaines cellules étaient vides, d'autres contenaient des phrases complètes, des notes explicatives ou encore des valeurs mal formatées (TRUE/FALSE au lieu de yes/no, par exemple). Un nettoyage de ces résultats a donc été nécessaire avant leur intégration dans la base, afin d'assurer la standardisation des réponses et la compatibilité avec les autres tableaux. Ce processus, bien qu'ayant nécessité plusieurs étapes d'ajustement, a permis d'obtenir un jeu de données plus cohérent, homogène et exploitable pour les analyses futures.

3.4 Structuration et stockage des données

Afin de centraliser et d'organiser les données extraites de manière durable, nous avons mis en place deux bases de données distinctes à l'aide du logiciel *DBeaver*, connecté à un moteur *SQLite*. L'objectif était de pouvoir comparer, croiser ou exploiter séparément les résultats issus des deux méthodes d'extraction utilisées. La première base de données regroupe les tables générées à partir de *FloraTraitier*, tandis que la seconde contient celles obtenues via l'IA *Le Chat* de *Mistral AI*. Chaque base stocke les traits sous forme de tables

organisées par source, en conservant une structure de colonnes identique : ID, nom de l'espèce, nom du trait, valeur du trait et unité.

La création de ces bases de données a été facilitée par le travail de nettoyage et de standardisation effectué en amont. Tous les tableaux ayant été harmonisés au préalable (colonnes renommées, types de données unifiés, valeurs manquantes filtrées), l'importation s'est faite simplement et sans nécessiter de traitement supplémentaire. Cette structuration permet désormais d'interroger les données efficacement à l'aide de requêtes SQL, d'automatiser des extractions ou comparaisons, et de préparer d'éventuelles visualisations ou analyses statistiques à plus grande échelle.

3.5 Analyse des données

L'ensemble des analyses repose sur deux tableaux de données (un provenant de l'IA et un de *FloraTraiter*) que nous avons construit de manière itérative suivant une hiérarchisation des sources. Pour chaque espèce et pour chaque trait nous avons ainsi intégré préférentiellement des données issues de *Wiktrop*, puis nous avons progressivement complété ces tableaux avec les informations manquantes en suivant un ordre de priorité : d'abord *World Flora Online*, ensuite Wikipedia, puis Tropical et enfin Feedipedia. Ce choix nous a permis de compiler un maximum de traits pour chaque espèce tout en conservant une certaine cohérence dans les sources utilisées.

Pour évaluer la fréquence et la distribution des traits extraits, nous avons généré plusieurs histogrammes empilés représentant la proportion relative des modalités pour certains traits qualitatifs, notamment la forme de vie, la durée de vie, le type de support et le mode de dispersion. Nous avons produit ces graphiques séparément pour les résultats extraits avec *FloraTraiter* (NLP) et ceux issus de l'IA *Le Chat* (LLM). Ces visualisations reposent sur des fonctions en R qui utilisent le package *dplyr* pour compter les occurrences de chaque modalité, calculer leur fréquence relative (en %) et regrouper les modalités rares dans une catégorie "autre" afin de rendre les graphiques plus lisibles. Une fois les fréquences calculées, nous les avons représentées grâce à des histogrammes empilés correspondant aux données extraites via *NLP* d'une part et aux données extraites via LLM d'autre part. Ces représentations nous ont permis de comparer les tendances globales entre les deux méthodes, d'évaluer leur couverture des différents traits, et de mieux comprendre la répartition des stratégies écologiques parmi les adventices.

Ensuite, pour évaluer la part des données provenant de chacune des sources (*Wiktrop*, *WorldFlora*, *Wikipedia*, etc...), nous avons calculé la proportion relative des données extraites par source. Cette analyse nous a permis d'estimer la contribution relative de chaque site à la construction de la base de données finale. Nous avons ensuite affiné cette analyse en regardant, pour les quatre types de traits sélectionnés avant (forme de vie, durée de vie, support, dispersion), quelle source était la plus représentée. Nous avons calculé, pour chaque type de trait, la proportion d'occurrences issues de chaque source. Nous avons ensuite représenté cela via deux histogrammes empilés (un pour l'IA et l'autre pour *FloraTraiter*). Cela donne une vision synthétique des contributions de chaque source en fonction des différents traits.

Puis, pour évaluer la cohérence entre les deux méthodes d'extraction, nous avons réalisé un graphique représentant les 30 traits pour lesquels les correspondances exactes sont les plus fréquentes entre les résultats du LLM et ceux de *FloraTraiter*. Pour ce faire, nous avons utilisé la base de données issue de l'IA comme référence, puis comparé les résultats avec ceux obtenus par *FloraTraiter*. Pour chaque couple espèce/trait, nous avons comptabilisé les cas où la valeur extraite par *FloraTraiter* correspondait exactement à celle identifiée par le LLM.

Nous avons également calculé la proportion des valeurs de traits fonctionnels quantitatifs considérés comme les plus pertinents (comme la taille des feuilles, des fruits, des graines, ou l'aire foliaire) par rapport à l'ensemble des valeurs de traits présentes dans la base des traits sélectionnés. Nous avons également évalué la distribution de ces valeurs numériques, en représentant leur densité selon une échelle logarithmique (de manière à mieux visualiser leur distribution tout en limitant l'influence des valeurs extrêmes).

Enfin, pour évaluer la complétude de la base de donnée par rapport à la liste d'espèces utilisée dans notre étude, nous avons généré plusieurs graphiques complémentaires. Tout d'abord, un histogramme a été généré pour représenter le nombre d'espèces pour lesquelles des données ont été extraites, selon les différentes sources mobilisées. Cela permet d'illustrer les différences de richesse entre les bases de données. Nous avons ensuite analysé la distribution du nombre de traits par espèce, à travers un graphique montrant combien d'espèces possèdent un certain nombre de traits renseignés. Cette visualisation permet d'identifier la densité d'information disponible par espèce. Pour finir, afin d'évaluer la diversité des types de traits extraits selon les parties de la plante, nous avons effectué une recherche par mots-clés dans les noms des traits. Les mots ciblés correspondent aux principaux organes végétaux : *leaf* (feuille), *root* (racine), *stem* (tige), *fruit*, *seed* (graine), *flower* (fleur) et *plant* (plante entière). Deux histogrammes ont été générés à partir de cette recherche : l'un indique la fréquence totale d'apparition de chaque mot-clé dans l'ensemble des traits, l'autre le nombre de traits distincts associés à chaque organe. Cette analyse donne un aperçu de la couverture des données en fonction des organes végétaux, et permet d'identifier les parties les plus représentées dans les descriptions.

Nous avons terminé cette analyse par une évaluation de la distribution du nombre de traits extraits par espèce, afin d'évaluer leur couverture moyenne, et par un recensement des trente traits les plus fréquemment retrouvés dans la base (c'est-à-dire apparaissant au moins une fois chez une espèce), afin d'identifier rapidement les traits les plus représentés dans l'ensemble des données collectées.

Pour explorer la structuration des espèces en fonction de leurs traits fonctionnels, nous avons réalisé plusieurs analyses en composantes principales (ACP) à l'aide de *Python*, en utilisant principalement les bibliothèques *pandas*, *scikit-learn*, *numpy* et *plotly* pour l'affichage interactif. Avant d'effectuer l'ACP, nous avons procédé à plusieurs étapes de préparation. Les variables ont été transformées avec une transformation *Box-Cox* pour améliorer la normalité, puis centrées et réduites à l'aide de *StandardScaler* de *scikit-learn*. Enfin, nous avons utilisé l'ACP de *scikit-learn* pour réduire les dimensions à deux axes principaux (PC1 et PC2). Pour la visualisation, nous avons utilisé *plotly* afin de créer des graphiques interactifs où chaque point représente une espèce, positionnée selon ses coordonnées sur les axes PC1 et PC2; les points sont colorés selon un trait qualitatif comme la forme de vie, la durée de vie, ou le nombre de cotylédons (les monocotylédones et dicotylédones représentant deux grands groupes phylogénétiques associées à différentes morphologies et stratégies écologiques des plantes) et une ellipse de confiance est tracée autour de chaque groupe pour visualiser leur dispersion. Nous avons réalisé deux ACP parallèles pour chaque trait. Une première avec un ensemble de 4 traits fonctionnels (aire des feuilles, taille des plantes, des graines et des fruits) et une seconde intégrant également le "SLA" (Specific Leaf Area). Ce choix s'explique par la couverture incomplète du SLA dans la base de données, peu d'espèces possédaient une valeur renseignée pour ce trait. En proposant une version avec et une sans SLA, nous avons ainsi pu maintenir un nombre d'espèces suffisant pour permettre une analyse statistique robuste tout en évaluant l'apport spécifique de ce trait dans la structuration fonctionnelle des espèces. Les ellipses sont générées automatiquement via une fonction qui calcule la distribution spatiale des points par groupe. Les flèches des variables, quant à elles, sont dessinées à partir des coefficients des composantes principales.

Pour aller plus loin dans l'analyse des valeurs de traits des espèces, nous avons réalisé une ACP comparative entre les espèces étudiées dans ce stage et celles de la base de données *TRY*. L'objectif était de positionner les espèces d'adventices tropicales par rapport à un large éventail d'espèces végétales globales, sur la base de trois des quatre traits fonctionnels choisis pour les autres ACP. Nous avons pour cela réalisé une ACP basée sur les valeurs de traits de 3 traits à la fois intéressants pour évaluer les stratégies des espèces et pour lesquels des nombreuses valeurs étaient disponibles dans la base *TRY* (aire des feuilles, taille des plantes, taille des graines). L'ACP a été réalisée sur l'ensemble des valeurs de traits disponible dans *TRY* ainsi que celles récoltées dans ce travail (lorsque les deux bases de données présentaient des valeurs pour les mêmes espèces, seules les valeurs récoltées dans cette étude ont été prises en compte). Avant d'effectuer l'analyse, nous avons appliqué les mêmes prétraitements que précédemment (*Box-Cox* et *StandardScaler*).

Pour évaluer la performance des deux méthodes d'extraction, nous avons mis en place une démarche de validation à partir de données extraites visuellement sur un échantillon de 20 espèces tirées au hasard. Pour chacune d'elles, nous avons sélectionné cinq traits fonctionnels : la taille de la plante, la taille des graines, la taille des feuilles, le type de support et le mode de dispersion. Nous avons ensuite consulté manuellement les fiches de ces espèces sur *Wiktrop*, afin de collecter les vraies valeurs. Les résultats extraits automatiquement par les deux méthodes ont été comparés à ces valeurs de référence. En suivant la méthode et les formules décrites dans cet article (*Named-Entity evaluation metrics based on entity-level*), nous avons attribué à chaque résultat une catégorie d'évaluation (Correct, Incorrect, Partiel, Manquant et Invention), ce qui nous a permis de calculer la précision, le rappel et le F1-Score pour chaque méthode. Voici les formules pour calculer la précision et le rappel disponibles dans l'article :

$$\text{Precision} = \frac{COR}{ACT} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{COR}{POS} = \frac{TP}{TP + FN}$$

Avec :

- POS : POSSIBLE(*POS*) = *COR* + *INC* + *PAR* + *MIS* = *TP* + *FN*
- ACT : ACTUAL(*ACT*) = *COR* + *INC* + *PAR* + *SPU* = *TP* + *FP*
- COR (Correct) : la valeur extraite est identique à la valeur de référence
- INC (Incorrect) : la valeur est fausse
- PAR (Partiel) : la valeur est proche ou similaire, mais pas exactement la même
- MIS (Manquant) : aucun trait n'a été extrait
- SPU (Invention) : la méthode a généré une valeur inexistante dans les données réelles

Ces résultats ont été organisés dans des tableaux comparatifs, puis synthétisés sous forme de graphes : 3 histogrammes illustrent la proportion de chaque catégorie de réponse et 3 autres graphiques présentent les scores de précision, de rappel et de F1-score, permettant de visualiser de manière claire les différences de performance entre les deux approches.

L'ensemble des graphiques présentés dans cette section ont été réalisés à l'aide de la librairie *ggplot2* sous R.

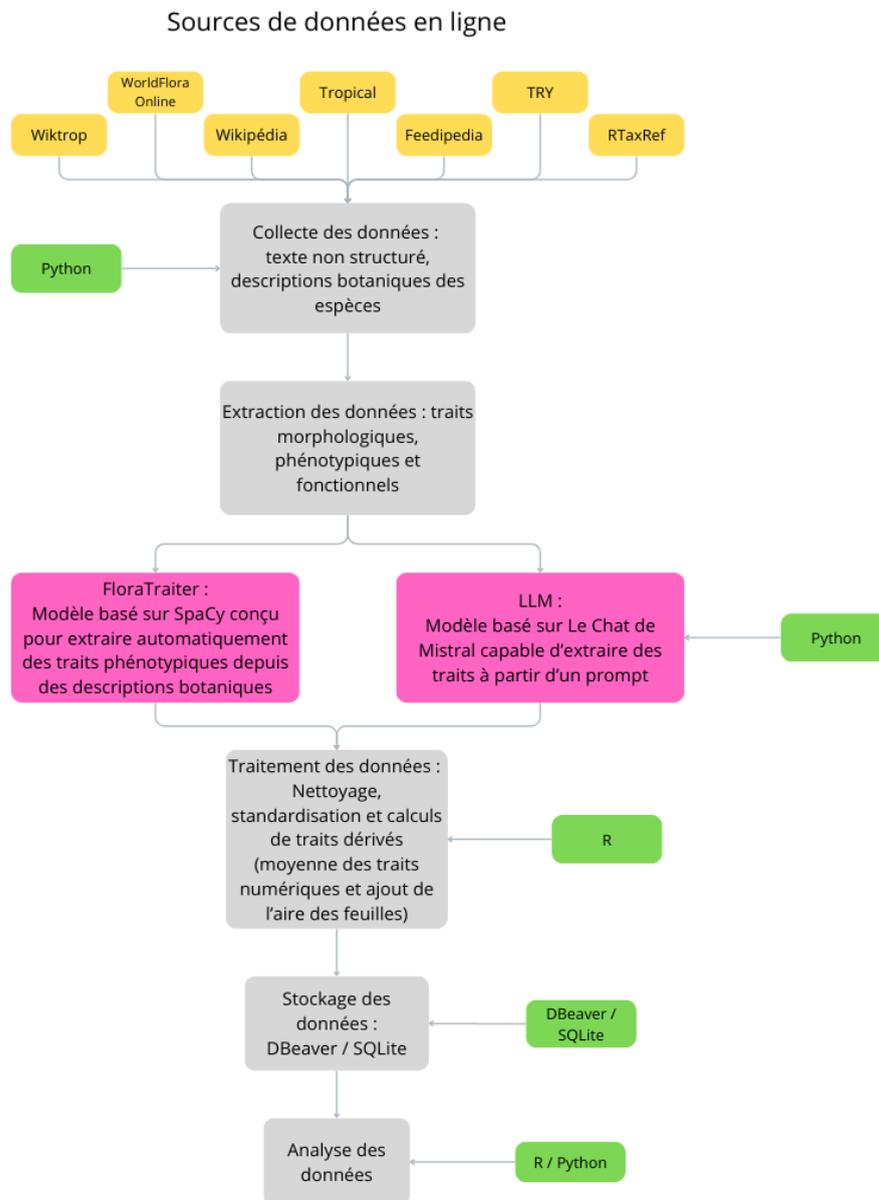


Figure 1 : Schéma récapitulatif du pipeline de collecte, traitement et analyse des traits fonctionnels des adventices tropicales

4. Résultats

4.1 Collecte et extraction des données

Les données ont été extraites à partir de six principales sources en ligne. Le nombre d'espèces pour lesquelles au moins une information a pu être récupérée varie selon les bases : 121 espèces pour *Feedipedia*, 394 pour *Wikipedia*, 463 pour *Tropical*, 824 pour la base *TRY*, 868 pour *Wiktrop*, et 878 pour *WorldFlora* (voir **Annexe 3**). Pour les extractions

réalisées via la méthode LLM, la grande majorité des données (environ 75%) provient de *WorldFlora*, tandis que les 25% restants sont issus de *Wiktrop*. Concernant *FloraTraiter*, les contributions sont plus équilibrées, avec 54% des données extraites à partir de *WorldFlora* et 46% depuis *Wiktrop* (voir **Annexe 4**). Ces proportions étaient attendues, d'une part en raison du nombre d'espèces disponibles dans chaque source, *Wiktrop* et *WorldFlora* étant largement majoritaires, et d'autre part en raison du tri opéré en amont, comme décrit dans la section *Matériel et Méthodes*. Cela reflète également des différences dans les performances des deux méthodes selon le type de descriptions et la richesse textuelle propre à chaque source.

En termes de diversité des traits extraits, *FloraTraiter* a permis d'identifier 1342 traits différents, contre 1019 pour la méthode LLM. Cet écart s'explique principalement par les différences de traitement appliquées aux deux jeux de données : dans le cas du LLM, une sélection manuelle des traits les plus pertinents a été opérée pour faciliter leur interprétation, tandis que *FloraTraiter* extrait de manière exhaustive l'ensemble des éléments textuels correspondant à ses patrons de reconnaissance.

Malgré l'objectif commun des deux méthodes, la comparaison directe entre les résultats de LLM et *FloraTraiter* révèle que seulement 5,8% des traits extraits sont identiques entre les deux approches. Ce faible taux de correspondance souligne des divergences importantes dans la manière dont chaque méthode identifie, interprète et structure l'information. Cela peut s'expliquer par les différences de fonctionnement entre une approche fondée sur des règles linguistiques strictes et une autre basée sur l'interprétation contextuelle via un modèle de langage.

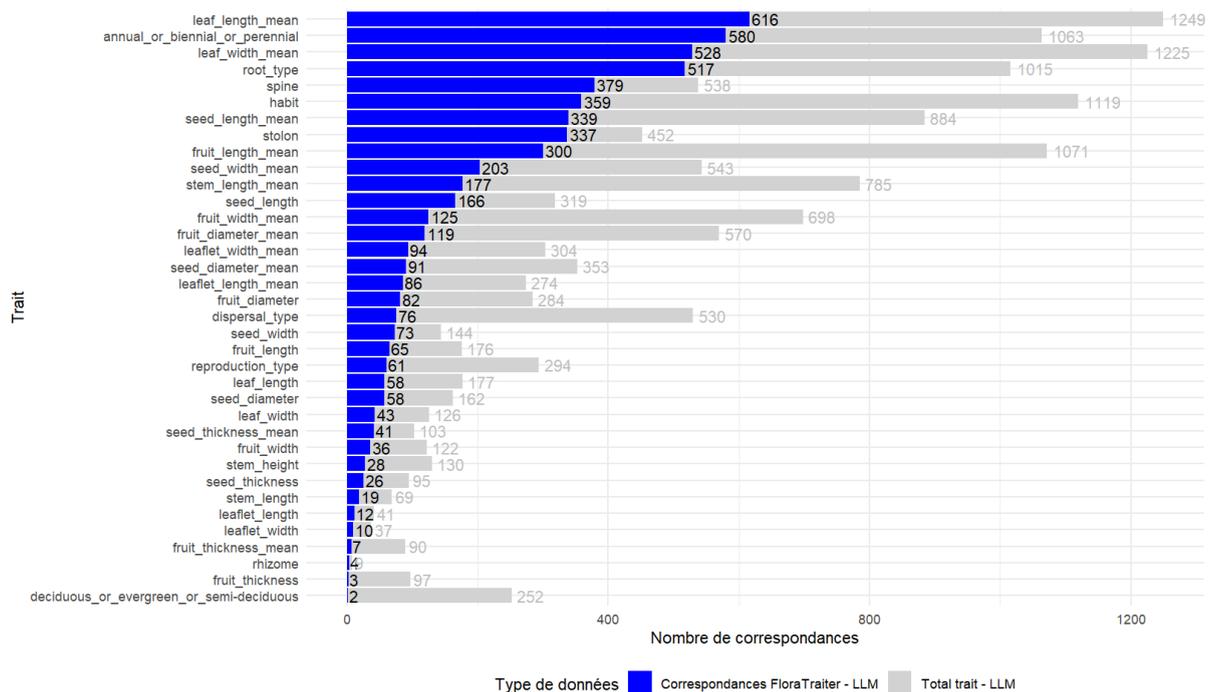


Figure 2 : Les 30 traits présentant le plus grand nombre de correspondances exactes dans leurs valeurs entre celles extraites via LLM et via *FloraTraiter*, comparé au nombre total de valeurs de traits extraits par LLM (ex: pour le premier trait : leaf_length_mean, 1249 valeurs de traits ont été extraites via LLM, dont 616 correspondent exactement à celles extraites via *FloraTraiter*).

La figure 2 met en évidence les traits pour lesquels les deux méthodes d'extraction, LLM et *FloraTraiter*, aboutissent à des résultats identiques. Sur les 5.8% de traits identiques, nous observons que les plus fréquemment identifiés de manière concordante concernent majoritairement certains traits fonctionnels : les dimensions des feuilles (8 traits différents),

graines et fruits (8 traits différents aussi pour chaque), ainsi que les formes de vie et les types de support des plantes. Ces résultats confirment que, malgré leurs différences, les deux approches tendent à converger sur l'extraction des traits les plus structurés et les plus fréquemment décrits dans les sources botaniques. Ces traits constitueront d'ailleurs la base des analyses fonctionnelles présentées dans la suite du rapport. Par ailleurs, les traits fonctionnels identifiés représentent environ 7,2% de l'ensemble des traits extraits, ce qui met en évidence à la fois leur importance pour l'analyse fonctionnelle, et la richesse en traits morphologiques variés que nous retrouvons dans les descriptions issues des différentes sources.

Enfin, en se concentrant sur quatre traits fonctionnels, le mode de dispersion, la durée de vie, la forme de vie et le type de support, nous observons que pour les deux méthodes, la majorité des données relatives à ces traits provient de *Wiktrop*. Une exception est toutefois notable pour le trait "forme de vie" avec *FloraTraiter*, pour lequel la source dominante n'est pas *Wiktrop* mais *WorldFlora* (voir **Annexe 5**).

4.2 Validation et qualité des données

Les figures 3 et 4 présentées ci-après sont des représentations visuelles issues de l'évaluation manuelle réalisée pour comparer les performances des deux méthodes d'extraction, LLM et *FloraTraiter*. Cette évaluation repose sur une vérification manuelle des valeurs de traits de 20 espèces sélectionnées de manière aléatoire, en consultant directement les descriptions sur *Wiktrop* (voir **Annexe 6**). Les traits examinés sont : la taille de la plante, la taille des graines, la longueur des feuilles, le type de support et le mode de dispersion. Les résultats extraits par les deux méthodes ont ensuite été comparés à ces valeurs de référence extraites manuellement (voir **Annexe 7** pour LLM et **Annexe 8** pour *FloraTraiter*). À partir de cette comparaison, nous avons classé chaque valeur dans une catégorie (correct, incorrect, partiellement correcte, manquant et invention), ce qui a permis de calculer les scores de précision, de rappel et de F1-Score (voir **Annexe 9**).

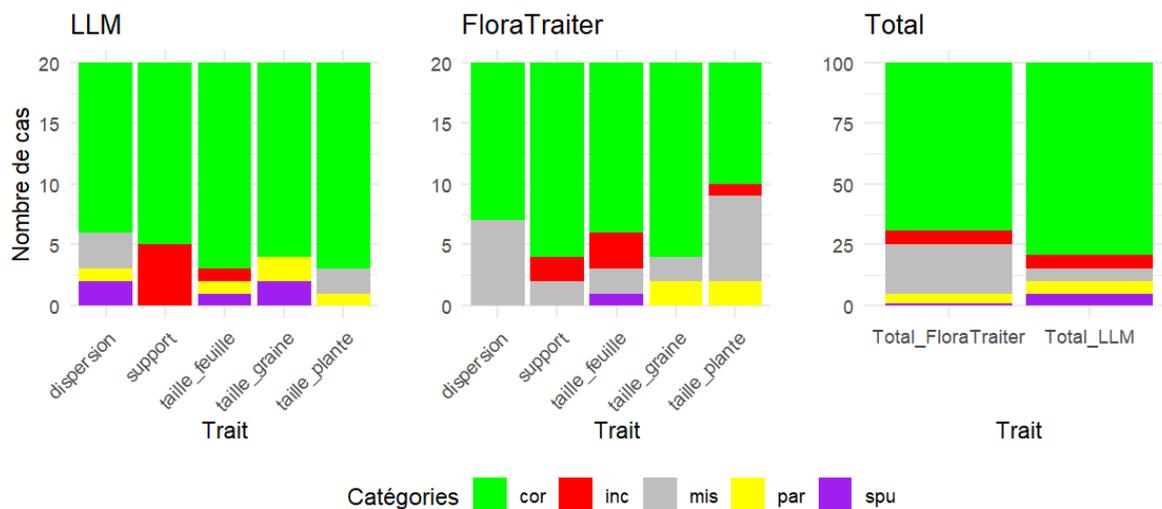


Figure 3 : Comparaison visuelle des vérifications manuelles avec les prédictions des deux méthodes (NLP et LLM). Cor: correct, inc: incorrect, mis: missing (valeur manquante), par: partiel, spu: spurious (invention, i.e. valeur générée par le model en l'absence de valeur réelle)

La figure 3 présente la répartition des résultats d'extraction pour les cinq traits fonctionnels par les deux méthodes d'extraction (LLM à gauche et *FloraTraiter* au centre). Les barres empilées indiquent, pour chaque trait, le nombre de cas classés selon les 5 catégories définies en légende.

Nous observons que pour certains traits, notamment la taille des feuilles et la taille de la plante, la méthode LLM surpasse nettement *FloraTraiter* en produisant davantage de valeurs correctes (figure 3). En revanche, pour le type de support, *FloraTraiter* obtient de meilleurs résultats, avec un nombre réduit de valeurs incorrectes. Le graphe de droite présente un résumé global, comparant les performances des deux méthodes les cinq catégories de traits confondues (figure 3). Nous observons que la méthode LLM présente une plus grande proportion de valeurs correctes par rapport à *FloraTraiter*, avec également moins de cas manquants. En revanche, le LLM produit un peu plus d'inventions et de valeurs partielles que *FloraTraiter*. Pour approfondir cette comparaison, nous allons à présent nous intéresser au F1-score, afin d'évaluer plus précisément les performances respectives des deux méthodes.

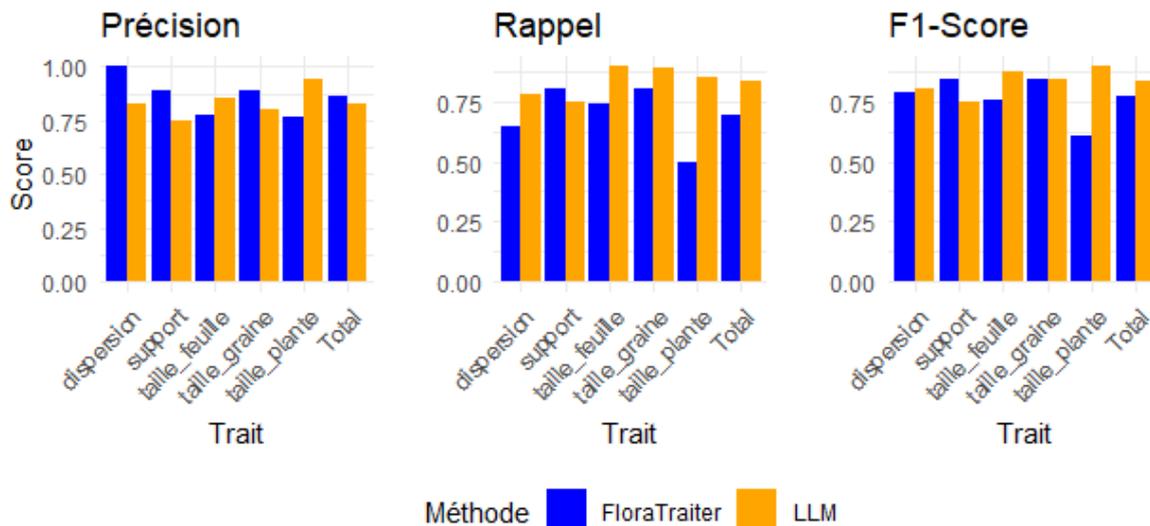


Figure 4 : Précision, Rappel et F1-Score des prédictions par trait pour LLM et *FloraTraiter*

La figure 4 permet de comparer les performances des méthodes *FloraTraiter* et LLM et met en évidence certaines différences en termes de précision, rappel et F1-score. Nous observons que *FloraTraiter* obtient un score de précision plus élevé. Cela signifie qu'une plus grande proportion des valeurs de traits extraites grâce à cette méthode sont correctes, autrement dit cette méthode génère moins de faux positifs (moins de valeurs générées par le modèle en l'absence de valeur réelle). En revanche, les scores de rappel et de F1-score obtenus avec *FloraTraiter* sont plus faibles que ceux du LLM, ce qui traduit une capacité moindre à détecter l'ensemble des traits pertinents présents dans les textes (c'est à dire plus de faux négatifs). À l'inverse, la méthode basée sur le LLM présente un meilleur score de rappel, ce qui suggère qu'elle détecte une plus grande diversité de traits présents dans les descriptions, même si cela se fait parfois au prix d'une précision légèrement inférieure. Son F1-score est aussi plus élevé, ce qui indique un meilleur compromis entre la qualité des traits extraits et la quantité d'informations retrouvées. Ces résultats montrent que *FloraTraiter* est plus conservateur et précis dans ses extractions, mais risque de ne pas détecter certaines informations, tandis que l'approche par LLM, plus souple, capte davantage d'informations mais peut introduire plus d'erreurs. Le choix de l'un ou l'autre dépend donc de l'objectif : prioriser la fiabilité ou la quantité des traits extraits.

4.3 Analyse des traits extraits

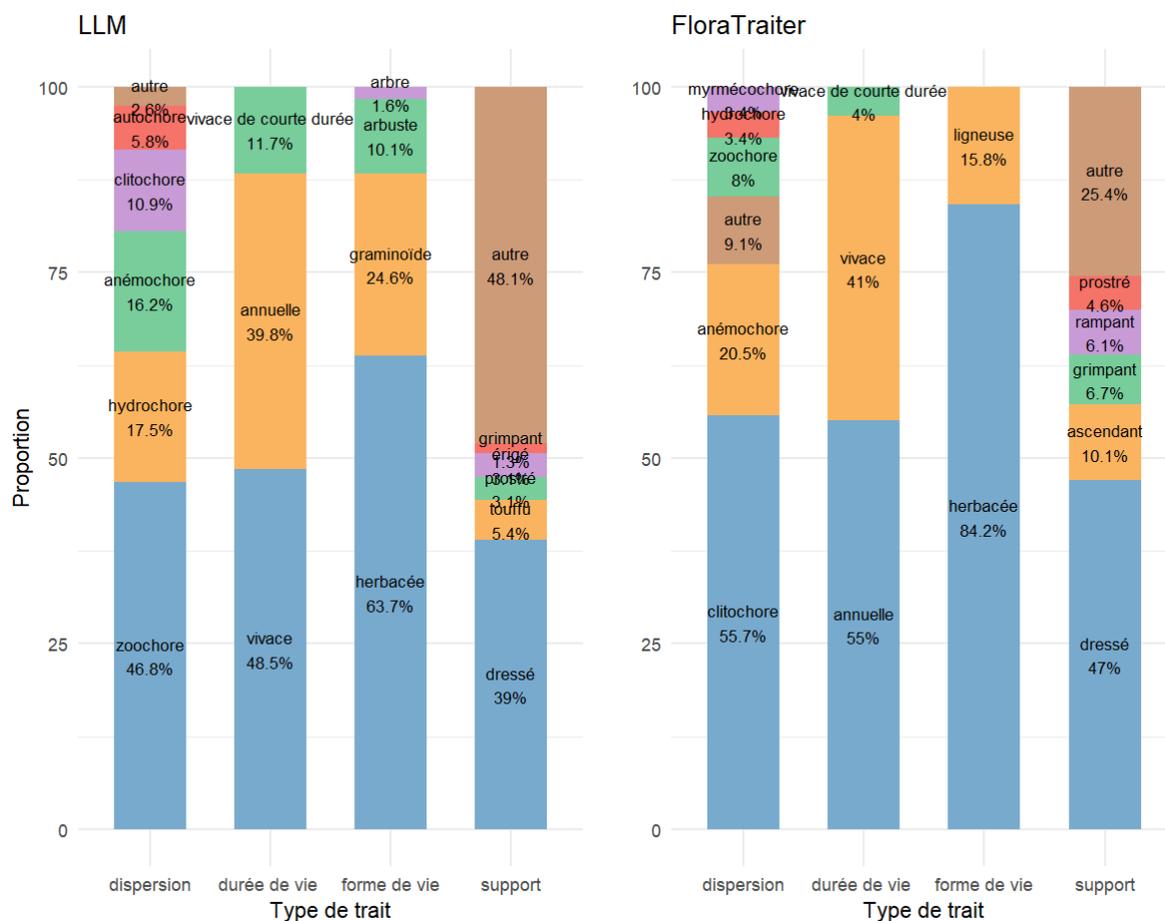


Figure 5 : Répartition des valeurs pour les traits de dispersion, durée de vie, forme de vie et support selon les 2 méthodes d'extraction (LLM et *FloraTraiter*).

En nous concentrant à nouveau sur les quatre traits fonctionnels cités précédemment (le mode de dispersion, la durée de vie, la forme de vie et le type de support), nous avons comparé la répartition des différentes valeurs observées pour chacun de ces traits selon les deux méthodes d'extraction (LLM et *FloraTraiter*, figure 5). Globalement, les deux approches s'accordent sur certaines valeurs dominantes, mais nous observons aussi des différences importantes, notamment dans la diversité des valeurs extraites. Pour le mode de dispersion, la méthode basée sur le LLM présente une répartition plus diversifiée des types de dispersion, incluant six modalités différentes, avec une dominance des "zoochore" (48.8%), "hydrochore" (17.5%) et "anémochore" (16.2%). De son côté, *FloraTraiter* identifie moins de modalités, avec une prédominance de "clitochore" (55.7%) et "anémochore" (20.5%). Cela reflète une extraction moins variée pour *FloraTraiter*. Concernant la durée de vie, les deux méthodes extraient principalement les catégories "annuelle" et "vivace", mais avec des proportions légèrement différentes. LLM identifie davantage d'espèces "vivaces" (48.5%) que d'"annuelles" (39.8%) contrairement à *FloraTraiter* qui identifie 55% d'espèces "annuelles" et 41% d'espèces "vivace". Cela reflète encore une différence entre les 2 méthodes. Concernant les formes de vie, la méthode LLM restitue une gamme plus étendue de modalités (quatre au total), dont "herbacée" (63.7%), "graminoïde" (24.6%), "arbuste" (10.1%) et "arbre" (1.6%). À l'inverse, *FloraTraiter* n'identifie que deux modalités : "herbacée" (84.2%) et "ligneuse" (15.8%), ce qui montre une extraction plus simplifiée. Enfin, pour le type de support, les différences entre les deux approches sont particulièrement marquées. La méthode LLM présente une répartition très hétérogène avec

près de 48% de modalités classées comme autres, et une fragmentation sur des modalités comme “grim pant”, “prostré”, “touffu”, “érigé” mais avec une majorité de “dressé”. Les extractions réalisées avec *FloraTraiter*, forment des catégories plus claires et mieux structurées, avec une modalité dominante “dressé” (47%), suivie de “ascendant” (10.1%), “grim pant” (6.7%), “rampant” et “prostré”, et seulement 25.4% classés en “autre”. En résumé, la méthode LLM permet une extraction plus riche et variée, ce qui est intéressant pour explorer la diversité fonctionnelle, mais au prix d’une plus grande hétérogénéité des résultats. *FloraTraiter*, en revanche, fournit des extractions plus homogènes et structurées, mais avec moins de variété dans les modalités identifiées.

Nous avons ensuite choisi de faire un focus sur des traits extraits à l’aide de la méthode LLM, celle-ci ayant montré de meilleures performances, notamment pour l’extraction de traits quantitatifs, comme démontré précédemment dans la section *Validation et qualité des données*.

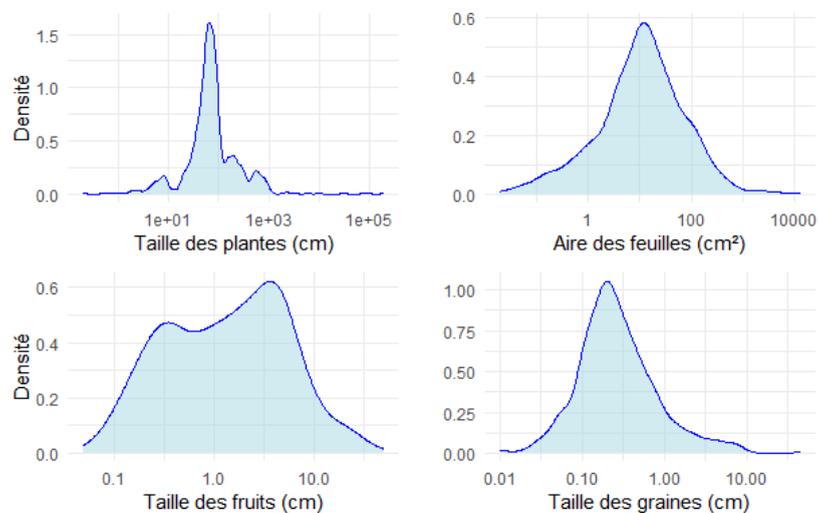


Figure 6 : Distribution logarithmique des valeurs de quatre traits fonctionnels extraits via la méthode basée sur LLM (L’utilisation de l’échelle logarithmique permet ici de mieux visualiser la large variabilité des valeurs)

La figure 6 présente la distribution logarithmique de quatre traits fonctionnels quantitatifs extraits par la méthode LLM : la taille des plantes, l’aire des feuilles, la taille des fruits et la taille des graines. Nous observons que la taille des plantes montre une forte concentration autour d’une centaine de centimètres, ce qui reflète une grande hétérogénéité structurelle entre espèces. L’aire des feuilles suit une distribution unimodale centrée autour de 10 cm², ce qui correspond à des tailles moyennes typiques des plantes herbacées tropicales. Nous notons cependant une longue queue à gauche indiquant la présence d’espèces à petites feuilles. La taille des fruits montre une distribution bimodale, ce qui pourrait refléter la coexistence de deux grands groupes fonctionnels ou morphologiques au sein des espèces étudiées : des espèces à petits fruits et d’autres à fruits plus larges. Enfin, la taille des graines suit une courbe en cloche relativement symétrique en échelle log, avec une concentration autour de 0.5 cm. Cette diversité de taille reflète la variabilité des stratégies de dispersion. Cette diversité de formes et de tailles montre bien la richesse fonctionnelle des adventices tropicales, et confirme la pertinence de l’approche par traits pour mieux comprendre leur écologie.

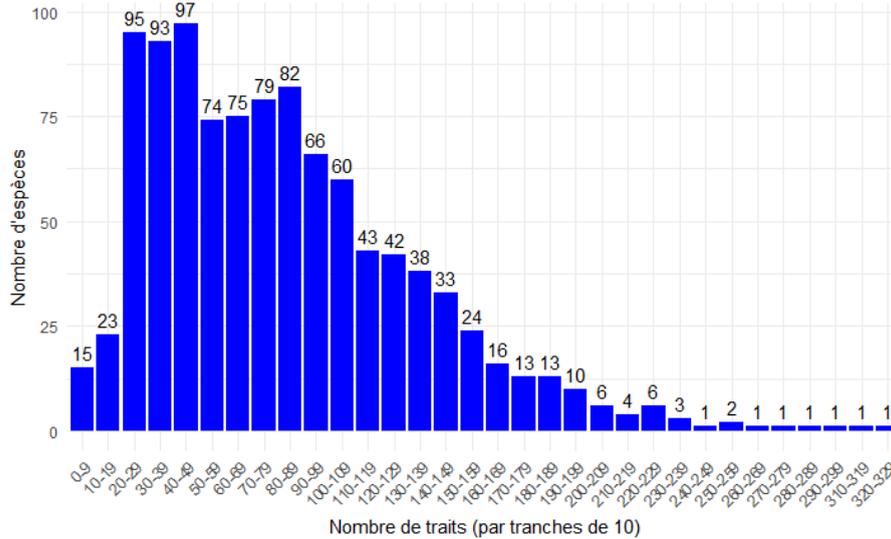


Figure 7 : Nombre de traits différents extraits par espèce - LLM

La figure 7 représente la distribution du nombre de traits fonctionnels extraits par espèce avec la méthode LLM. L'axe des abscisses indique le nombre de traits, regroupés par tranches de 10, tandis que l'axe des ordonnées montre le nombre d'espèces correspondant à chaque tranche. Nous observons que la majorité des espèces possèdent entre 20 et 100 traits, avec un pic dans la tranche 40-49 (97 espèces). Le nombre d'espèces décroît progressivement au-delà de cette plage. Très peu d'espèces ont plus de 200 traits renseignés, ce qui montre une certaine rareté des espèces très bien documentées. Inversement, seules quelques espèces ont moins de 20 traits. Cette distribution met en évidence une concentration autour d'un nombre moyen de traits extraits par espèce, ce qui témoigne à la fois de la capacité du LLM à récupérer un grand nombre de traits pour un large éventail d'espèces et des limites rencontrées pour certaines d'entre elles.

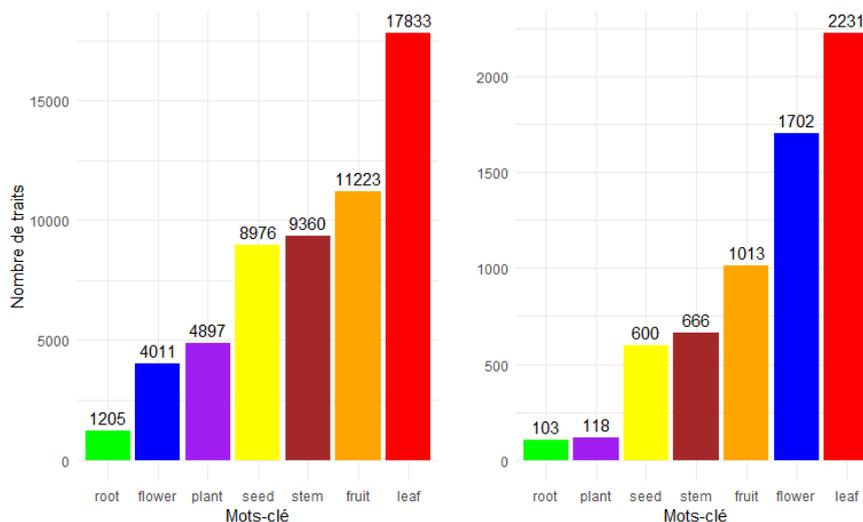


Figure 8 : Nombre de traits extraits par mot-clé (à gauche) et nombre de traits différents (chaque trait comptés qu'une seule fois) par mot-clé (à droite) - méthode basée sur le LLM.

La figure 8 présente deux histogrammes comparant, pour chaque organe de la plante (racine, fleur, plante, graine, tige, fruit et feuille), le nombre total de traits extraits (à gauche) et le nombre de traits unique (chaque nom de trait n'est compté qu'une seule fois, quelle que soit sa fréquence dans les données) (à droite) par la méthode LLM. Le graphique de gauche

montre que les traits en rapport avec les feuilles, les fruits, la tige et les graines sont les plus nombreux, avec un maximum pour les feuilles (17833 traits), suivi des fruits (11223) et de la tige (9360). À l'inverse, les mots-clés "root" (1205) et "plant" (4011) sont ceux pour lesquels le moins de traits ont été extraits. Le graphique de droite montre que le mot-clé "leaf" domine (2231 traits), suivi par "flower" (1702) et "fruit" (1013). Cela montre que les feuilles sont non seulement très représentées, mais aussi très diversifiées en termes de traits fonctionnels. Cette figure met en évidence une disparité dans la richesse et la fréquence des traits extraits selon les organes végétaux, ce qui peut refléter une abondance de données disponibles dans la littérature pour certains organes (notamment des feuilles). Cette observation est confirmée par le fait que les traits les plus similaires entre les deux méthodes d'extraction sont ceux associés aux feuilles, aux graines et aux fruits, comme le montre la figure 2.

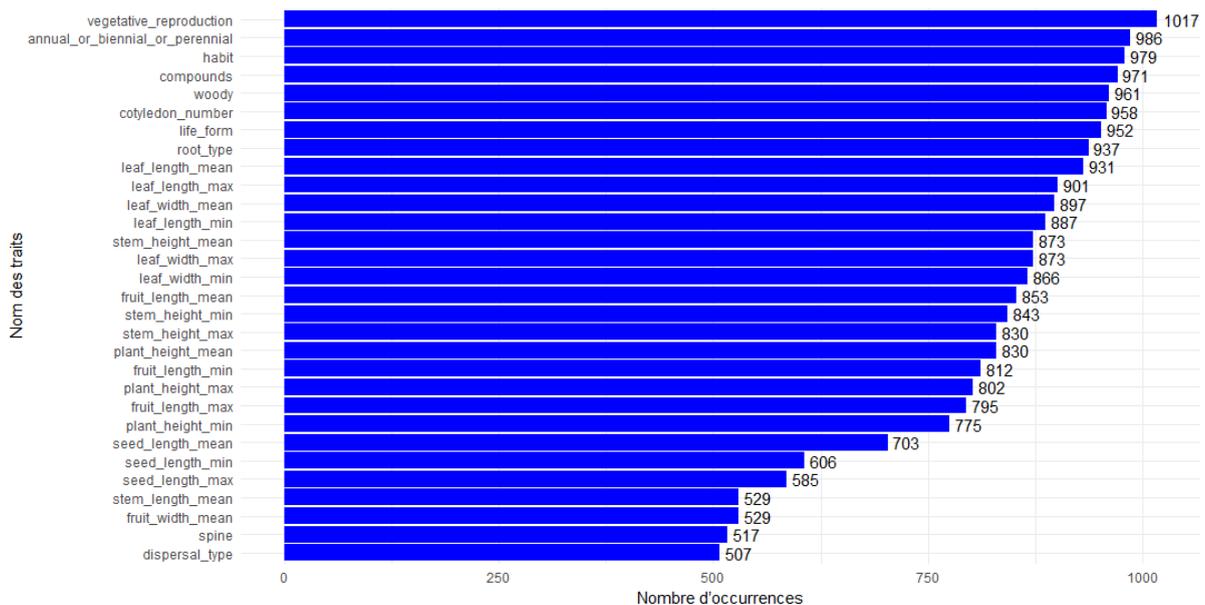


Figure 9 : 30 traits les plus fréquemment trouvés au moins une fois par espèce - LLM

La figure 9 présente les 30 traits les plus fréquemment extraits par la méthode LLM. Nous observons que les traits les plus représentés sont liés aux caractéristiques générales des plantes, comme la reproduction végétative, le type de durée de vie (annual_or_biennial_or_perennial), le type de support (habit) ou encore la forme de vie (life_form). Ces traits sont bien renseignés par plusieurs sources, en particulier *Wiktrop* et *WorldFlora*, et leur fréquence élevée confirme leur rôle central dans les descriptions botaniques disponibles en ligne. Un second ensemble très représenté concerne les dimensions foliaires, avec une forte occurrence des mesures de longueur et de largeur des feuilles (min, max, moyenne). Cette abondance vient renforcer ce qui avait déjà été mis en évidence dans les analyses précédentes : les traits associés aux feuilles constituent l'un des groupes les plus abondants et les mieux documentés dans les bases de données consultées. Leur fréquence peut s'expliquer par leur importance écologique et leur caractère facilement mesurable. De la même manière, la forte représentation des traits liés à la durée de vie et au type de support confirme que ces deux dimensions sont également bien documentées. Enfin, des traits relatifs aux graines, fruits et tiges apparaissent également dans ce classement, indiquant que ces traits sont également souvent documentés et peuvent être facilement extraits, bien que de manière un peu plus lacunaire selon les espèces et les sources.

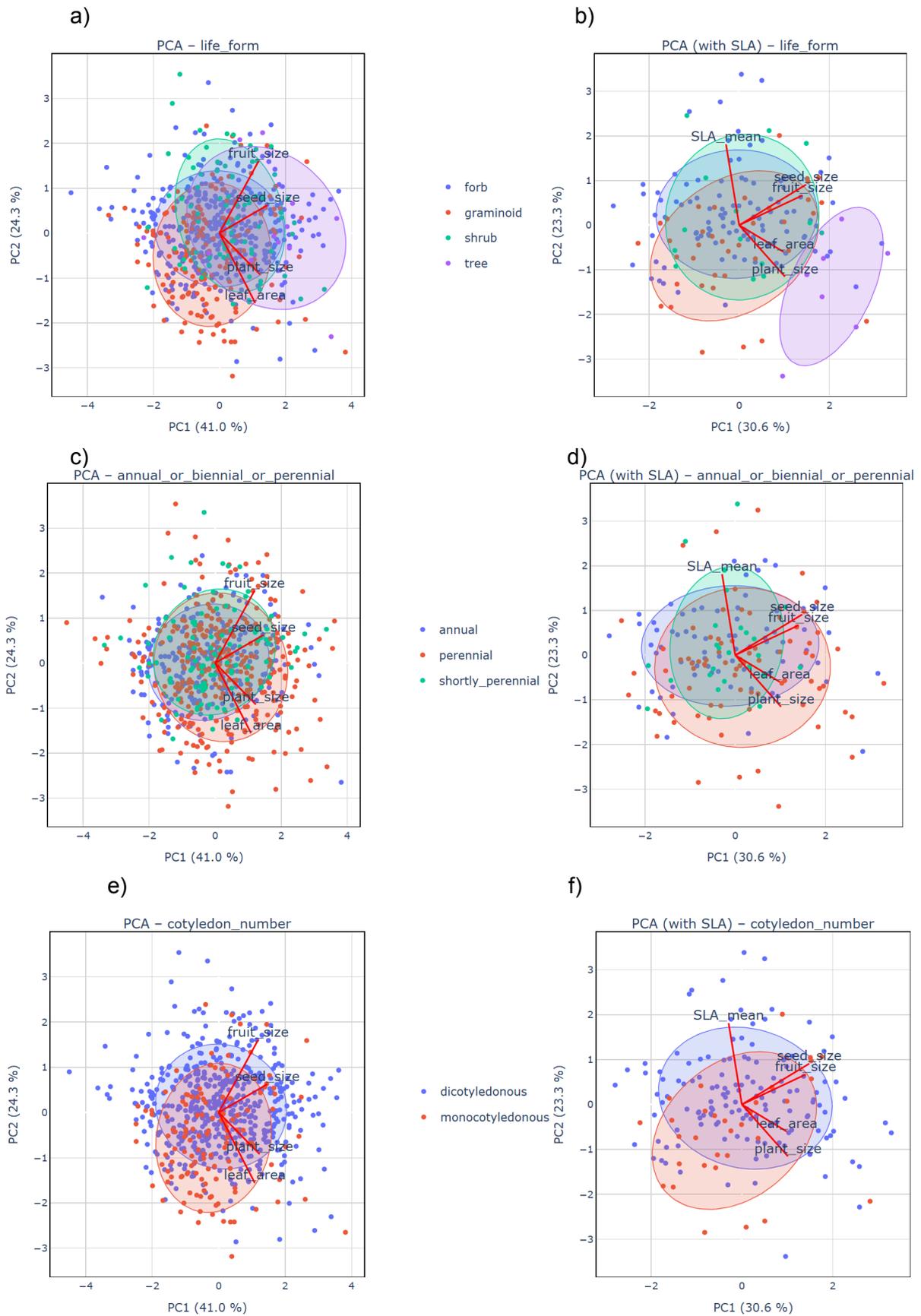


Figure 10 : Analyse en composantes principales (ACP) des valeurs moyennes de différents traits fonctionnels quantitatifs des espèces (tailles des fruits, des plantes, des feuilles et des graines) selon les formes de vie végétale (a, b), la durée de vie (c, d) et le nombre de cotylédons (e, f), à gauche

sans le SLA (Surface Foliaire Spécifique), à droite avec le SLA inclus (ce trait n'étant disponible que pour certaines espèces, réduisant le nombre d'espèces analysées). Les ellipses visibles sur les graphiques délimitent les regroupements d'espèces partageant une même modalité pour le trait catégoriel analysé. Elles permettent de visualiser la dispersion et le positionnement fonctionnel des groupes dans l'espace réduit des deux premières composantes principales. - Données issues de la méthode basée sur le LLM.

Sur la figure 10, nous remarquons que l'ajout du SLA (colonnes de droite) change légèrement la structuration de l'espace défini par les composantes principales. L'axe PC1 explique un peu moins de variance avec le SLA (30,6%) qu'en son absence (41%), ce qui indique que ce trait capte une partie de la variabilité fonctionnelle des espèces qui n'est pas représentée par les autres traits quantitatifs retenus ici. Nous voyons aussi que le vecteur correspondant au "SLA_mean" est plus orienté vers l'axe PC2, ce qui montre qu'il contribue davantage à la séparation des groupes sur cette dimension. Du point de vue de la forme de vie (figures a et b), nous notons une séparation relativement claire des arbres par rapport aux autres groupes, surtout lorsqu'on ajoute le SLA. Les valeurs des arbustes et graminoides restent plus dispersées, tandis que les herbacées forment un groupe plus dense et central. Cela reflète des différences marquées de taille de toutes les parties de ces plantes. Les arbres (tree) sont plus grands donc ont des feuilles, des fruits et des graines plus grandes tandis que les herbacées (forb) sont souvent petites donc auront des feuilles, des fruits et des petites graines. Pour la durée de vie (figures c et d), les trois modalités (annuelle, vivace, vivace de courte durée) présentent des regroupements partiels mais significatifs, avec une légère séparation entre les plantes vivaces et les annuelles. Nous remarquons que les plantes vivaces ont tendance à être plus grandes que les plantes annuelles et les vivaces de courte durée, ce qui s'explique par leur durée de vie plus longue, leur permettant donc de se développer plus. Elles investissent aussi plus dans des structures pérennes, contrairement aux annuelles qui privilégient la reproduction rapide. L'ajout du SLA renforce cette distinction, ce qui suggère que ce trait permet de mieux différencier les stratégies de croissance associées à la longévité des plantes. Enfin, pour le nombre de cotylédons (figures e et f), la distinction entre monocotylédones et dicotylédones est relativement bien marquée le long des deux axes. Les dicotylédones semblent davantage associées à des valeurs plus élevées du SLA, des tailles de graines et de fruits que les monocotylédones. La différence de la taille des feuilles et des plantes n'est pas très marquée entre ces deux groupes. Cela peut s'expliquer par une grande diversité morphologique et écologique de ces deux caractéristiques. Ces résultats montrent que les ACP permettent d'identifier des groupes fonctionnels cohérents au sein des données extraites, avec des regroupements pertinents selon les grandes catégories biologiques et écologiques de plantes. L'ajout du SLA améliore la capacité de discrimination de certaines modalités, en particulier pour la forme de vie et la durée de vie, et enrichit l'interprétation écologique globale des traits analysés.

PCA of DF_interest and DF_TRY based on plant_size, leaf_area, seed_size

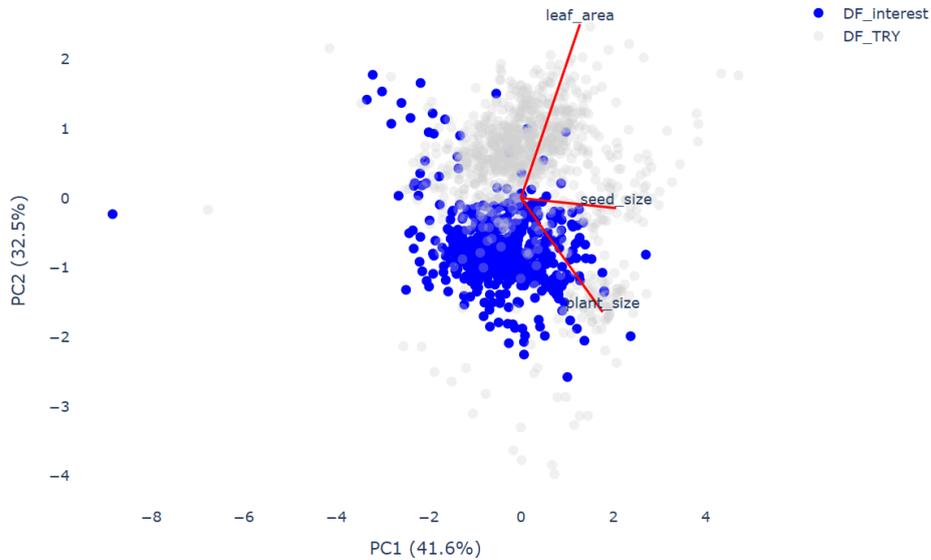


Figure 11 : Représentation en ACP des espèces d'intérêts et des espèces de la base de donnée TRY selon les traits fonctionnels - Données extraites par la méthode basée sur le LLM

L'ACP présentée sur la figure 7 compare la position des espèces d'intérêt (en bleu), extraites dans ce projet, avec celles issues de la base de données *TRY* (en gris), en se basant sur trois traits fonctionnels communs : la taille de la plante, l'aire foliaire et la taille des graines. Les flèches rouges indiquent la contribution de chaque trait aux axes principaux. L'axe PC1 (41,6% de la variance) est fortement influencé par la taille des plantes et des graines, tandis que l'aire foliaire contribue davantage à l'axe PC2 (32,5%). Les espèces d'intérêt forment un groupe relativement concentré, positionné dans la partie centrale et inférieure du nuage, ce qui suggère une gamme de valeurs plus restreinte ou spécifique par rapport à l'ensemble plus dispersé des espèces de *TRY*. Cette distribution illustre que les espèces tropicales adventices analysées ici occupent une niche fonctionnelle distincte, différente de la majorité des espèces présentes dans *TRY*. Elles tendent notamment à avoir des valeurs plus faibles pour les trois traits analysés, ce qui peut refléter des stratégies écologiques particulières associées à leur statut d'adventices en milieu tropical. Ce résultat renforce la pertinence de constituer une base de données ciblée sur ces espèces, qui sont sous-représentées dans les grandes bases généralistes.

5. Discussion

Ce travail avait pour objectif d'extraire, de stocker et d'analyser automatiquement des traits fonctionnels des adventices tropicales à partir de descriptions textuelles issues de différentes sources en ligne. Pour cela, deux méthodes ont été comparées. La première était un pipeline basé sur du traitement automatique de langage naturel (*FloraTraiter*). La seconde était basée sur un grand modèle de langage (LLM) via l'IA *Le Chat de Mistral*. Les résultats montrent qu'il est possible d'extraire un volume significatif de traits à partir de textes non structurés, avec toutefois des niveaux de performance variables selon la méthode employée et la qualité des sources. Environ 7,2 % des traits extraits relèvent des catégories fonctionnelles ciblées (dimensions, durée de vie, support, dispersion), ce qui confirme à la fois les limites de l'approche dans sa capacité à identifier des informations écologiquement pertinentes mais aussi la capacité des méthodes à extraire une grande diversité de données.

L'utilisation des LLM permet ainsi de gagner en flexibilité et en productivité en automatisant des tâches longues et complexes. Cependant, leur usage soulève également des enjeux importants, notamment sur le plan éthique et juridique : la question des droits d'auteurs concernant les textes analysés reste floue, et le recours à ces technologies pourrait à terme remettre en cause certains emplois. Ces aspects devront être pris en compte dans les projets futurs exploitant l'intelligence artificielle dans le domaine scientifique. (Weaver et al., 2023)

5.1 Comparaison, avantages et limites des méthodes

D'un point de vue méthodologique, les deux approches présentent des avantages et des limites. *FloraTraiter* repose sur une extraction basée sur des règles syntaxiques et lexicales, en s'appuyant sur *SpaCy*. Cette approche est particulièrement adaptée à des descriptions formalisées, mais elle s'avère peu robuste face à la diversité des styles rédactionnels rencontrés dans les bases consultées (*Wiktrop*, *WorldFlora*, *Wikipedia*...). Elle détecte des traits structurés avec rigueur, mais rate souvent des valeurs exprimées de manière plus souple ou intégrées dans un long paragraphe. Les performances limitées de *FloraTraiter* peuvent donc être expliquées en grande partie par la complexité et l'hétérogénéité des textes, souvent trop narratifs, structurés de manière différente, ou riches en informations non standardisées.

En particulier, certains traits comme le type de dispersion posent des difficultés spécifiques. L'information est souvent exprimée sans être explicitement liée à un nom de trait dans le texte (ex : "cette espèce est anémochore"), ou bien formulée de manière non standardisée (ex : "dispersée par le vent"). Cela limite fortement la capacité de *FloraTraiter* à les détecter automatiquement, ce qui se traduit par des données très lacunaires (seulement 91 espèces avec une valeur extraite pour ce trait) et des distributions biaisées, comme la surreprésentation des modalités comme "clitochore" (voir **Figure 5**).

Face à ces limites, l'approche par LLM s'est révélée plus flexible et mieux adaptée à la diversité des descriptions. Grâce à des prompts guidés, cette approche a permis d'extraire des valeurs numériques, des unités et des modalités qualitatives dans un format structuré. Cette méthode permet une meilleure couverture de l'information, mais au prix d'une certaine hétérogénéité dans les résultats, avec des erreurs de format ou des valeurs imprécises. L'étude des scores de précision-rappel montre par ailleurs que cette approche atteint de meilleurs scores de rappel et de F1-score que *FloraTraiter*, au prix d'une précision parfois inférieure. Ces résultats rejoignent les constats faits dans d'autres travaux récents sur l'usage des LLM en extraction de traits, notamment dans *AutomaticTraitExtraction* (Marcos et al., 2025). Toutefois, les performances de *FloraTraiter* restent satisfaisantes et cohérentes avec celles rapportées dans l'étude de Folk et al. (Folk et al., 2024).

Cependant, certaines limites subsistent, notamment lorsque plusieurs valeurs sont extraites pour un même trait. Les méthodes utilisées, qu'il s'agisse de *FloraTraiter* ou du LLM, ne disposent pas de mécanisme de validation contextuelle leur permettant de trancher entre des valeurs contradictoires ou multiples. Par exemple, l'espèce *Phragmites australis* qui est un roseau commun mais qui peut atteindre 8 mètres de haut, est référencée en temps qu'arbuste dans *WorldFloraOnline*. Cette ambiguïté a conduit à son classement automatique dans la catégorie "tree" lors de l'extraction, alors que cette espèce, bien que de grande taille, est plutôt classée dans la catégorie des herbacées par les botanistes. Ce cas illustre bien le fait que, sans intervention humaine, il est difficile pour les modèles d'arbitrer correctement en cas d'hétérogénéité ou de variabilité des informations pour une même espèce. Cela souligne la nécessité, dans les projets futurs, d'envisager une étape de validation manuelle ciblée, ou d'une étape supplémentaire d'arbitrage et de sélection des données basée sur des LLM ou sur des modèles statistiques (ex: sélection de la valeur la

plus probable lorsque plusieurs valeurs sont extraites pour un même trait), notamment pour les espèces à fort enjeu agronomique ou écologique.

L'utilisation de modèles plus récents et plus puissants, comme GPT-4, Mistral Medium 3 ou Llama 4 Maverick, pourrait encore améliorer la qualité des extractions, notamment en ce qui concerne la précision des unités, des valeurs numériques, ou la catégorisation des traits. Ces modèles, bien qu'encore peu accessibles, représentent une perspective intéressante pour les projets futurs. Leur utilisation reste toutefois limitée par leur coût, à la fois en termes d'accès (modèles payants) et de ressources nécessaires (puissance de calcul élevée).

Une piste intéressante à explorer en perspective serait de combiner les deux approches d'extraction, en tirant parti des atouts complémentaires des LLM et des méthodes NLP. Concrètement, une première étape pourrait consister à utiliser un LLM pour transformer les descriptions botaniques brutes en textes ou tableaux plus structurés, avec une syntaxe simplifiée et des formats homogènes. Ces sorties intermédiaires, standardisées, pourraient ensuite être analysées de manière plus efficace par des outils NLP comme *FloraTraiter* ou d'autres modèles comme les modèles BERT de Google (Domazetoski et al., 2023), dont les performances dépendent fortement de la régularité du langage. En élargissant la réflexion, certains travaux proposent également des approches plus "low-tech" (Coleman et al., 2023), reposant sur l'extraction ciblée de termes ou de patrons lexicaux spécifiques, sans avoir recours à des modèles d'IA complexes. Ces méthodes, bien que plus simples et moins flexibles, peuvent être très efficaces pour extraire rapidement un petit nombre de traits bien définis. Elles offrent une solution rapide et légère pour des usages ciblés, notamment dans des contextes où les ressources techniques ou informatiques sont limitées, ou lorsque l'on cherche à valider manuellement un sous-ensemble de données. Une telle hybridation des méthodes permettrait potentiellement d'améliorer la précision, la complétude et la cohérence des extractions, tout en facilitant l'automatisation à grande échelle.

Une autre piste prometteuse pour réduire les fausses informations et construire des modèles capables de suivre l'évolution rapide de la littérature est la capacité des LLM à interagir avec des bases de données grâce à la génération augmentée par récupération (RAG), une méthode qui consiste à interroger une base de donnée pour enrichir les réponses du modèle avec des informations à jour. (Farrell et al., 2024).

Enfin, il serait aussi utile de constituer une base de données standardisée, ouverte, et enrichie de traits fonctionnels issus de multiples sources textuelles. À l'image de *TRY*, mais étendue à toutes les espèces végétales, une telle ressource faciliterait l'analyse comparative, la modélisation fonctionnelle et la gestion agroécologique à grande échelle. Cela pourrait être fait en utilisant par exemple les données des herbiers numériques qui peuvent enrichir de façon significative le nombre de données récoltées (Heberling, 2022) (Pearson et al., 2020).

Dans cette perspective, il serait également essentiel d'œuvrer vers une harmonisation des termes employés et de leurs définitions dans les travaux en écologie (Garnier et al., 2017). L'utilisation cohérente de concepts et de traits, avec des références communes partagées entre études, est en effet indispensable pour garantir la comparabilité des jeux de données, éviter les ambiguïtés sémantiques, et renforcer la robustesse des analyses écologiques à grande échelle.

5.2 Intérêt pour l'agroécologie tropicale et la compréhension de l'écologie des adventices des cultures

Au-delà des considérations techniques, ce projet contribue à une meilleure compréhension des stratégies fonctionnelles des adventices tropicales. Ces plantes,

souvent perçues uniquement comme nuisibles, jouent un rôle crucial dans les agroécosystèmes tropicaux, que ce soit en tant que concurrentes des cultures ou pour leur capacité à fournir certains services écosystémiques (refuges aux ennemis naturels des ravageurs, amélioration de la fertilité des sols, etc.) (Gaba et al., 2017).

Bien que les traits fonctionnels ne représentent que 7,2 % des données extraites par la méthode LLM, leur qualité et leur structuration ont permis de réaliser des analyses en composantes principales (ACP) sur un nombre d'espèces suffisant pour dégager des tendances robustes. L'identification et la structuration de ces traits permettent en effet d'étudier le positionnement des adventices tropicales le long d'axes fonctionnels élémentaires représentant les stratégies écologiques des espèces. Nos résultats mettent en évidence des regroupements d'espèces selon leur forme de vie, leur durée de vie, ou leur nombre de cotylédon, ce qui peut par exemple être interprété à l'aide du cadre théorique CSR (Compétitives, Stress-tolérantes, Rudérales) (Pierce et al., 2013). En effet, ce cadre peut permettre de distinguer, à partir des valeurs de traits des espèces, les espèces compétitives, tolérantes au stress, ou rudérales (c'est à dire ayant une stratégie pionnière), offrant ainsi un outil pertinent pour mieux comprendre les stratégies écologiques des adventices et anticiper leur comportement dans les agroécosystèmes tropicaux.

Parmi les espèces identifiées, certaines comme *Oryza sativa* (graminoïde), *Spathodea campanulata* (arbre) ou *Dracontium polyphyllum* (herbacée) présentent des tailles importantes et de grandes surfaces foliaires, des caractéristiques typiques des espèces compétitives. Dans un contexte tropical tel que celui étudié en Guadeloupe, ce type d'espèce adventices peuvent avoir un impact particulièrement négatif sur les grandes cultures comme la banane, en entrant en compétition directe pour les ressources (lumière, nutriments, eau). À l'inverse, d'autres espèces comme *Eleocharis retroflexa* et *Lipocarpa chinensis*, toutes deux graminoïdes de petite taille avec de faibles surfaces foliaires, illustrent plutôt des stratégies rudérales. Ce type d'espèces se caractérise par une croissance rapide, une production élevée de petites graines, et une forte capacité de colonisation, ce qui peut les rendre plus problématiques dans des systèmes à cycle court ou à cultures basses, comme les parcelles de maraîchage ou les cultures d'ananas. Cette typologie fonctionnelle, basée sur des traits morphologiques simples mais informatifs, permet donc d'envisager une gestion différenciée des adventices, adaptée au contexte agronomique local et aux cultures en place.

6. Conclusion

En conclusion, ce stage nous a permis d'explorer des méthodes innovantes pour enrichir les connaissances sur les traits fonctionnels des adventices tropicales à partir de sources textuelles en ligne. En combinant des techniques de traitement automatique du langage naturel (NLP), de web scraping et en utilisant des grands modèles de langage (LLM), nous avons pu collecter, structurer et analyser des données issues de bases variées, souvent non harmonisées. L'utilisation conjointe de deux approches complémentaires, *FloraTraiter*, basé sur des règles linguistiques, et un LLM (*Le Chat*), a mis en lumière les avantages et limites de chacune : d'un côté la rigueur, de l'autre la flexibilité et la richesse sémantique. Si le faible taux de correspondance exacte entre les deux méthodes souligne les divergences d'interprétation, il révèle aussi leur potentiel combinatoire pour couvrir une plus grande diversité de traits.

Les résultats obtenus montrent que certains traits, comme ceux liés aux feuilles, aux graines ou à la durée de vie, sont particulièrement présents dans les descriptions botaniques. Ils confirment également l'intérêt croissant d'utiliser les traits fonctionnels pour comprendre les rôles écologiques des adventices, notamment dans un contexte tropical où leur prolifération, accentuée par le changement climatique, représente un enjeu fort pour la

durabilité des systèmes agricoles. Par ailleurs, cette analyse met aussi en lumière des catégories de traits moins fréquemment renseignés, comme ceux liés aux racines, qui pourraient faire l'objet de travaux futurs afin d'enrichir les bases de données existantes et mieux couvrir l'ensemble des dimensions fonctionnelles des plantes.

Ce travail ouvre des perspectives prometteuses pour la construction de bases de données de traits adaptées aux spécificités des systèmes tropicaux, encore largement sous-documentés. Il constitue une première étape vers une meilleure caractérisation fonctionnelle des adventices tropicales, en mobilisant des outils d'extraction semi-automatisés basés sur le NLP et les LLM. Les données extraites peuvent ainsi contribuer à enrichir la connaissance de ces espèces, alimenter des outils d'aide à la décision, ou encore soutenir des stratégies de gestion agroécologique plus ciblées. Des améliorations restent à envisager, notamment en ce qui concerne la qualité et la diversité des sources utilisées, ainsi que la précision des modèles d'extraction.

Enfin, une conclusion importante à retenir est que plusieurs méthodes plus ou moins sophistiquées existent pour extraire des traits fonctionnels, mais que leur pertinence dépend avant tout des objectifs spécifiques du projet : le nombre d'espèces ciblées, l'hétérogénéité des sources, ou encore la nature et le niveau de précision attendu pour les traits à extraire. Ce projet pose ainsi les bases d'une réflexion plus large sur l'intégration raisonnée des approches d'intelligence artificielle dans l'agronomie tropicale.

7. Bibliographie

An R client for TaxRef the French Taxonomical Database. Consulté 7 mai 2025, à l'adresse

<https://rekyt.github.io/rtaxref/>

Anwar, M. P., Islam, A. K. M. M., Yeasmin, S., Rashid, M. H., Juraimi, A. S., Ahmed, S., &

Shrestha, A. (2021). Weeds and Their Responses to Management Efforts in A

Changing Climate. *Agronomy*, 11(10), Article 10.

<https://doi.org/10.3390/agronomy11101921>

Cirad, la recherche agronomique pour le développement durable des régions tropicales et

méditerranéennes. | *Cirad*. Consulté 7 mai 2025, à l'adresse <https://www.cirad.fr/>

Coleman, D., Gallagher, R. V., Falster, D., Sauquet, H., & Wenk, E. (2023). A workflow to

create trait databases from collections of textual taxonomic descriptions. *Ecological*

Informatics, 78, 102312. <https://doi.org/10.1016/j.ecoinf.2023.102312>

Côte, F.-X. (ed), Le Bellec, F. (ed), Martin, T. (ed), Temple, L. (ed), Blouin, A. (ed),

Loeillet, D. (ed), Baufume, S. (ed), & Ghneim Herrera, T. (ed). (2025). *Rapport final*

de l'initiative " Pesticide Reduction for Tropical Agricultures " (Pretag) 2023-2024

[Monograph]. CIRAD. <https://doi.org/10.18167/agritrop/00835>

Díaz, S., Kattge, J., Cornelissen, J. H. C., Wright, I. J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C., Colin Prentice, I., Garnier, E., Bönisch, G., Westoby, M., Poorter, H., Reich, P. B., Moles, A. T., Dickie, J., Gillison, A. N., Zanne, A. E., ... Gorné, L. D. (2016). The global spectrum of plant form and function. *Nature*, 529(7585), 167-171. <https://doi.org/10.1038/nature16489>

Domazetoski, V., Kreft, H., Bestova, H., Wieder, P., Koynov, R., Zarei, A., & Weigelt, P. (2023). *Using natural language processing to extract plant functional traits from unstructured text* (p. 2023.11.06.565787). bioRxiv. <https://doi.org/10.1101/2023.11.06.565787>

Farrell, M. J., Brierley, L., Willoughby, A., Yates, A., & Mideo, N. (2022). Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 289(1975), 20212721. <https://doi.org/10.1098/rspb.2021.2721>

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., Yates, A., & Mideo, N. (2024). The changing landscape of text mining : A review of approaches for ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 291(2027), 20240423. <https://doi.org/10.1098/rspb.2024.0423>

Folk, R. A., Guralnick, R. P., & LaFrance, R. T. (2024). FloraTraiter : Automated parsing of traits from descriptive biodiversity literature. *Applications in Plant Sciences*, 12(1), e11563. <https://doi.org/10.1002/aps3.11563>

Gaba, S., Perronne, R., Fried, G., Gardarin, A., Bretagnolle, F., Biju-Duval, L., Colbach, N., Cordeau, S., Fernández-Aparicio, M., Gauvrit, C., Gibot-Leclerc, S., Guillemin, J.-P., Moreau, D., Munier-Jolain, N., Strbik, F., & Reboud, X. (2017). Response and effect traits of arable weeds in agro-ecosystems : A review of current knowledge. *Weed Research*, 57(3), 123-147. <https://doi.org/10.1111/wre.12245>

Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougnot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F. S., Bönisch, G., Bunker, D. E., Cornelissen, J. H. C., Díaz, S.,

- Enquist, B. J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., ... Klotz, S. (2017). Towards a thesaurus of plant characteristics : An ecological contribution. *Journal of Ecology*, *105*(2), 298-309.
<https://doi.org/10.1111/1365-2745.12698>
- GECO - Fonctionnement écologique et gestion durable des agrosystèmes bananiers et ananas*. Consulté 7 mai 2025, à l'adresse <https://ur-geco.cirad.fr/>
- Heberling, J. M. (2022). Herbaria as Big Data Sources of Plant Traits. *International Journal of Plant Sciences*, *183*(2), 87-118. <https://doi.org/10.1086/717623>
- WFO. Consulté 9 mai 2025, à l'adresse <https://www.worldfloraonline.org/>
- Le Chat—Mistral AI*. Consulté 2 juin 2025, à l'adresse <https://chat.mistral.ai/chat>
- List of feeds | Feedipedia*. (s. d.). Consulté 9 mai 2025, à l'adresse <https://www.feedipedia.org/content/feeds?category=All>
- Marcos, D., Vlasakker, R. van de, Athanasiadis, I. N., Bonnet, P., Goeau, H., Joly, A., Kissling, W. D., Leblanc, C., Proosdij, A. S. J. van, & Panousis, K. P. (2025). *Fully automatic extraction of morphological traits from the Web : Utopia or reality?* (No. arXiv:2409.17179; Version 2). arXiv. <https://doi.org/10.48550/arXiv.2409.17179>
- Named-Entity evaluation metrics based on entity-level*. (s. d.). Consulté 6 mai 2025, à l'adresse https://www.davidsbatista.net//blog/2018/05/09/Named_Entity_Evaluation/
- Oerke, E.-C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, *144*(1), 31-43. <https://doi.org/10.1017/S0021859605005708>
- Oikonomou, E. D., Karvelis, P., Giannakeas, N., Vrachatis, A., Glavas, E., & Tzallas, A. T. (2024). How natural language processing derived techniques are used on biological data : A systematic review. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *13*(1), 23. <https://doi.org/10.1007/s13721-024-00458-1>
- Palma, E., Vesk, P. A., White, M., Baumgartner, J. B., & Catford, J. A. (2021). Plant functional traits reflect different dimensions of species invasiveness. *Ecology*, *102*(5), e03317. <https://doi.org/10.1002/ecy.3317>
- Pearson, K. D., Nelson, G., Aronson, M. F. J., Bonnet, P., Brenskelle, L., Davis, C. C.,

- Denny, E. G., Ellwood, E. R., Goëau, H., Heberling, J. M., Joly, A., Lorieul, T., Mazer, S. J., Meineke, E. K., Stucky, B. J., Sweeney, P., White, A. E., & Soltis, P. S. (2020). Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience*, 70(7), 610-620. <https://doi.org/10.1093/biosci/biaa044>
- Pierce, S., Brusa, G., Vagge, I., & Cerabolini, B. E. L. (2013). Allocating CSR plant functional types : The use of leaf economics and size traits to classify woody and herbaceous vascular plants. *Functional Ecology*, 27(4), 1002-1010. <https://doi.org/10.1111/1365-2435.12095>
- Schrader, J., Shi, P., Royer, D. L., Peppe, D. J., Gallagher, R. V., Li, Y., Wang, R., & Wright, I. J. (2021). Leaf size estimation based on leaf length, width and shape. *Annals of Botany*, 128(4), 395-406. <https://doi.org/10.1093/aob/mcab078>
- Thessen, A. E., Cui, H., & Mozzherin, D. (2012). Applications of Natural Language Processing in Biodiversity Science. *Advances in Bioinformatics*, 2012(1), 391574. <https://doi.org/10.1155/2012/391574>
- TRY Proposal Submission*. Consulté 9 mai 2025, à l'adresse <https://www.try-db.org/TryWeb/Prop0.php>
- Useful Tropical Plants*. Consulté 9 mai 2025, à l'adresse <https://tropical.theferns.info/>
- Violle, C., Navas, M., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, 116(5), 882-892. <https://doi.org/10.1111/j.0030-1299.2007.15559.x>
- Weaver, W. N., Ruhfel, B. R., Lough, K. J., & Smith, S. A. (2023). Herbarium specimen label transcription reimaged with large language models : Capabilities, productivity, and risks. *American Journal of Botany*, 110(12), e16256. <https://doi.org/10.1002/ajb2.16256>
- Wikipédia:Accueil principal. (2024). In *Wikipedia, the free encyclopedia*. https://fr.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Accueil_principal&oldid=221036809
- WIKTROP - Weed Identification and Knowledge in the Tropical and Mediterranean areas*.

8. Annexes

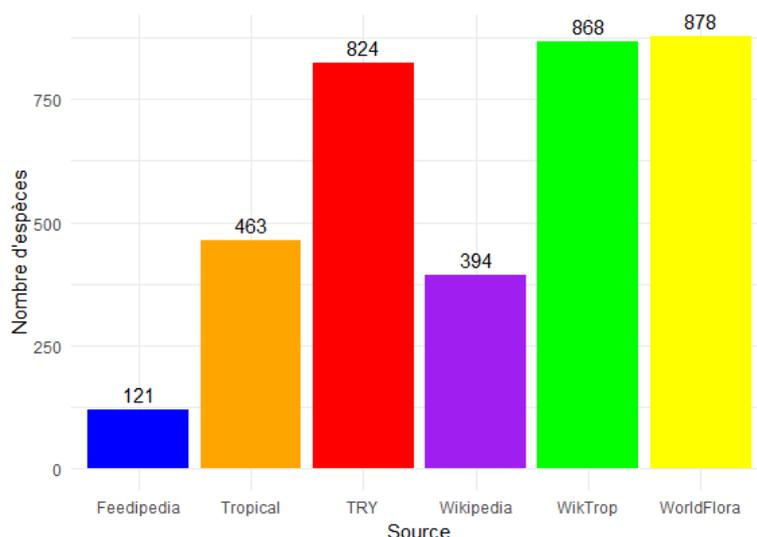
Annexe 1 : Prompt utilisé pour extraire les traits avec un LLM (Le Chat, Mistral AI)

```
prompt = (
    f"You are an expert in botanical data extraction. Extract and structure the requested traits as accurately as possible.\n"
    f"\n"
    f"### Instructions:\n"
    f"1. Analyze the following text and extract traits in CSV format : {text} \n"
    f"2. Use these column headers: TraitName, TraitValue, Units.\n"
    f"3. Follow these strict rules:\n"
    f"   - Convert all trait names to lowercase with '_' as a separator.\n"
    f"   - Replace spaces, hyphens, and other separators with '_'.\n"
    f"   - If a trait has min and max values, store them separately as 'trait_name_min' and 'trait_name_max'.\n"
    f"   - Units must be in cm.\n"
    f"   - If a trait is mentioned but has no value, replace with 'NA'.\n"
    f"   - Exclude notes, ambiguous data, or irrelevant text.\n"
    f"   - If 'TraitValue' and 'Units' are both missing, exclude the row.\n"
    f"   - Do not infer values—extract only explicit data.\n"
    f"\n"
    f"### Traits to Extract:\n"
    f'{', '.join(traits)}\n'
    f"\n"
)
```

Annexe 2 : Définition et fonction écologique de certains traits d'intérêt

Traits	Définition	Fonctions écologiques
Dimensions des parties de la plante	caractéristiques mesurables telles que la longueur, la largeur ou le diamètre des organes végétaux (feuilles, tiges, graines, fruits, etc.).	Vitesse de croissance (plante entière), compétition (plante entière, feuilles), dispersion (fruits, graines), utilisation des ressources (plante entière, feuilles)
Type de support	manière dont une plante se maintient ou se développe en hauteur (ex : dressée, grimpante, rampante).	Compétition
Mode de reproduction	stratégie de reproduction d'une plante, incluant la reproduction sexuée (via graines) ou asexuée (par stolons, rhizomes, etc.).	Compétition, vitesse d'établissement dans un nouvel environnement
Mode de dispersion	mécanisme par lequel les graines ou propagules sont disséminées (ex : par le vent, l'eau, les animaux...).	Vitesse d'établissement dans un nouvel environnement, vitesse et dispersion depuis un environnement vers un autre
Durée de vie	cycle de vie d'une plante, déterminant si elle est annuelle, vivace ou vivace de courte durée.	Compétition, résistance au stress, résilience des populations face aux perturbations

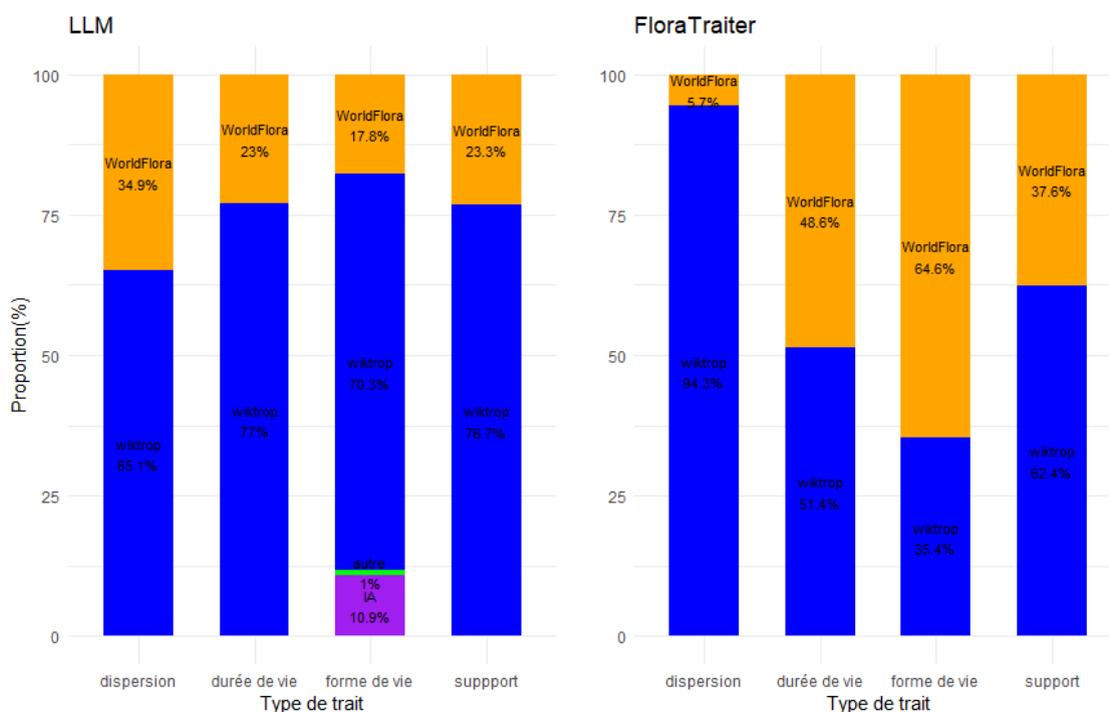
Annexe 3 : Nombre d'espèce pour lesquelles des données ont été extraites selon les différentes sources



Annexe 4 : Proportion des traits extraits par sources

Méthode \ Source	WorldFlora	Wiktrop	Wikipedia	Tropical	IA	Other
LLM (en %)	75.2	24.6	0	0	0.1	0
FloraTraiter (en %)	54.2	45.8	0	0	0	0

Annexe 5 : Proportion du nombre de valeurs des traits de dispersion, durée de vie, forme de vie et support extraites selon la source



Annexe 6 : Vérifications manuelles sur *Wiktrop* de la taille des plantes, de la taille des graines, de la longueur des feuilles, du support et du type de dispersion de 20 espèces tirées au sort

Species	plant_size	seed_size	leaf_length	habit	dispersal_type
Ambrosia_maritima	30-100	X	3-8	erect/ramulous	sd by wind, water, animals and tillage equipment
Bacopa_floribunda	8-65	0.03	1-5	erect/glabrous	X
Bolboschoenus_maritimus	30-180	fruit (0.25-4)	0.2-1.2	erect/creeping	X
Dianella_ensifolia	60-150	0.35	25-60	erect	dispersed by water and man
Emex_spinosa	?-50	fruit (0.5-0.9)	X	prostrate	dispersed by wind, machines, animal
Eragrostis_unioloides	?-30	X	X	upright	X
Foeniculum_vulgare	100-200	fruit (0.5-0.7)	X	upright/glabrous/erect	clithochorous
Galinsoga_parviflora	15-60	X	3-7	erect	X
Heterotis_rotundifolia	?-40	0.06	1.5-6	upright/ascending	X
Lithospermum_arvense	?-50	fruit (0.3)	X	upright/ascending	clithochorous
Lolium_arundinaceum	60-100	0.33-0.35	10-60	upright	dispersed by animal
Ludwigia_abyssinica	100-200	0.05-0.08	3-12	erect	X
Malva_sylvestris	?-150	X	3-5	erect	clithochorous
Oryza_longistaminata	200-300	0.75-0.85	20-45	creeping/erect	X
Ottelia_ulvifolia	?	fruit (4-6)	?-23 (15 average)	sub-acauline aquatic grass	dispersed by water
Paederia_farinosa	?-several meters	0.55-0.8	6.5-12.5	climbing/creeping	X
Pycnus_flavescens	10-30	X	X	clumps	X
Rhynchospora_rugosa	30-150	fruit (0.11-0.15)	10-40	clumping	X
Rhynchospora_trispicata	30-150	fruit (0.5-0.6)	20-100	cespitose	X
Sicyos_angulatus	?-1000	fruit (1)	20-25	climbing	dispersed by animal or machines

Annexe 7 : Comparaison des valeurs de traits entre LLM et les vérifications faites manuellement

Species	plant_size	seed_size	leaf_length	habit	dispersal_type
Ambrosia_maritima	30-100	X	3-8	erect/ramulous	sd by wind, water, animals and tillage equipment
Bacopa_floribunda	8-65	0.03	1-5	erect/glabrous	X
Bolboschoenus_maritimus	30-180	X	0.2-1.2	erect/creeping	X
Dianella_ensifolia	60-150	0.35	25-60	erect	dispersed by water and man
Emex_spinosa	X	X	X	prostrate	dispersed by wind, machines, animal
Eragrostis_unioloides	?-30	X	X	upright	X
Foeniculum_vulgare	100-200	X	X	upright/glabrous/erect	clithochorous
Galinsoga_parviflora	15-60	X	3-7	erect	X
Heterotis_rotundifolia	?-40	0.06	1.5-6	upright/ascending	X
Lithospermum_arvense	?-50	X	X	upright/ascending	clithochorous
Lolium_arundinaceum	60-100	0.33-0.35	10-60	upright	dispersed by animal
Ludwigia_abyssinica	100-200	0.05-0.08	3-12	erect	X
Malva_sylvestris	?-150	X	3-5	erect	clithochorous
Oryza_longistaminata	200-300	0.75-0.85	20-45	creeping/erect	X
Ottelia_ulvifolia	X	X	?-23 (15 average)	sub-acauline aquatic grass	dispersed by water
Paederia_farinosa	X	0.55-0.8	6.5-12.5	climbing/creeping	X
Pycnus_flavescens	10-30	X	X	clumps	X
Rhynchospora_rugosa	30-150	X	10-40	clumping	X
Rhynchospora_trispicata	30-150	X	20-100	X	X
Sicyos_angulatus	?-1000	X	20-25	climbing	dispersed by animal or machines

Annexe 8 : Comparaison des valeurs de traits entre FloraTraiter et les vérifications faites manuellement

Species	plant_size	seed_size	leaf_length	habit	dispersal_type
Ambrosia_maritima	30-100	X	3-8	erect/ramulous	dispersed by wind, water, animals and tillage equipment
Bacopa_floribunda	8-65	0.03	1-5	erect/glabrous	X
Bolboschoenus_maritimus	30-180	X	0.2-1.2	erect/creeping	X
Dianella_ensifolia	60-150	0.35	25-60	erect	dispersed by water and man
Emex_spinosa	7-50	X	X	prostrate	dispersed by wind, machines, animal
Eragrostis_unioloides	7-30	X	X	upright	X
Foeniculum_vulgare	100-200	X	X	upright/glabrous/erect	clithochorous
Galinsoga_parviflora	15-60	X	3-7	erect	X
Heterotis_rotundifolia	7-40	0.06	1.5-6	upright/ascending	X
Lithospermum_arvense	7-50	X	X	upright/ascending	clithochorous
Lolium_arundinaceum	60-100	0.33-0.35	10-60	upright	dispersed by animal
Ludwigia_abyssinica	100-200	0.05-0.08	3-12	erect	X
Malva_sylvestris	7-150	X	3-5	erect	clithochorous
Oryza_longistaminata	200-300	0.75-0.85	20-45	creeping/erect	X
Ottelia_ulvifolia	?	X	7-23 (15 average)	sub-acauline aquatic grass	dispersed by water
Paederia_farinosa	7-several meters	0.55-0.8	6.5-12.5	climbing/creeping	X
Pycnus_flavescens	10-30	X	X	clumps	X
Rhynchospora_rugosa	30-150	X	10-40	clumping	X
Rhynchospora_trispicata	30-150	X	20-100	cespitose	X
Sicyos_angulatus	7-1000	X	20-25	climbing	dispersed by animal or machines

Annexe 9 : Calcul du F1-Score

Calcul F1_Score LLM							
Catégories	plant_size	seed_size	leaf_length	habit	dispersal_type	TOTAL	
cor	17	16	17	15	14	79	
par	1	2	1	0	1	5	
spu	0	2	1	0	2	5	
inc	0	0	1	5	0	6	
mis	2	0	0	0	3	5	
pos	20	18	19	20	18	95	
act	18	20	20	20	17	95	
prec	0,9444444444	0,8	0,85	0,75	0,8235294118	0,8315789474	
rec	0,85	0,8888888889	0,8947368421	0,75	0,7777777778	0,8315789474	
F1	0,8947368421	0,8421052632	0,8717948718	0,75	0,8	0,8315789474	

Calcul F1_Score FloraTraiter							
Catégories	# plant_size	# seed_size	# leaf_length	# habit	# dispersal_type	TOTAL	
cor	10	16	14	14	16	69	
par	2	2	0	0	0	4	
spu	0	0	1	0	0	1	
inc	1	0	3	2	0	6	
mis	7	2	2	2	2	20	
pos	20	20	19	20	20	99	
act	13	18	18	18	18	80	
prec	0,7692307692	0,8888888889	0,7777777778	0,8888888889	0,8	0,8625	
rec	0,5	0,8	0,7368421053	0,8	0,65	0,696969697	
F1	0,6060606061	0,8421052632	0,7567567568	0,8421052632	0,7878787879	0,7709497207	

Abelmoschus ficulneus

Abrus precatorius
Abutilon indicum

Acacia farnesiana
Acacia nilotica

<i>Acalypha alopecuroides</i>	<i>Amaranthus retroflexus</i>	<i>Axonopus surinamensis</i>
<i>Acalypha crenata</i>	<i>Amaranthus spinosus</i>	<i>Bacopa crenata</i>
<i>Acalypha indica</i>	<i>Amaranthus viridis</i>	<i>Bacopa floribunda</i>
<i>Acalypha segetalis</i>	<i>Ambrosia maritima</i>	<i>Baldellia ranunculoides</i>
<i>Acanthocereus</i>	<i>Ambrosia tenuifolia</i>	<i>Barbarea verna</i>
<i>tetragonus</i>	<i>Ammannia auriculata</i>	<i>Basella alba</i>
<i>Acanthospermum</i>	<i>Ammannia baccifera</i>	<i>Basilicum polystachyon</i>
<i>australe</i>	<i>Ammannia coccinea</i>	<i>Bellardia trixago</i>
<i>Acanthospermum</i>	<i>Ammannia multiflora</i>	<i>Bellis annua</i>
<i>hispidum</i>	<i>Ammannia priureana</i>	<i>Beta vulgaris</i>
<i>Achyranthes aspera</i>	<i>Ammi majus</i>	<i>Bidens cynapiifolia</i>
<i>Acis autumnalis</i>	<i>Amorpha fruticosa</i>	<i>Bidens frondosa</i>
<i>Acmella caulirhiza</i>	<i>Amorphophallus</i>	<i>Bidens pilosa</i>
<i>Acmella uliginosa</i>	<i>paeoniifolius</i>	<i>Bifora testiculata</i>
<i>Acroceras amplexans</i>	<i>Anacyclus clavatus</i>	<i>Biophytum umbraculum</i>
<i>Acroceras hubbardii</i>	<i>Anagallis arvensis</i>	<i>Blyxa aubertii</i>
<i>Acroceras zizanioides</i>	<i>Anchusa italica</i>	<i>Boerhavia diffusa</i>
<i>Actinoscirpus grossus</i>	<i>Ancylobothrys petersiana</i>	<i>Boerhavia repens</i>
<i>Adonis microcarpa</i>	<i>Andropogon bicornis</i>	<i>Bolboschoenus maritimus</i>
<i>Aegilops triuncialis</i>	<i>Aneilema lanceolatum</i>	<i>Borago officinalis</i>
<i>Aegilops ventricosa</i>	<i>Anemone coronaria</i>	<i>Bothriospermum tenellum</i>
<i>Aerva javanica</i>	<i>Anredera cordifolia</i>	<i>Brachiaria deflexa</i>
<i>Aerva lanata</i>	<i>Anthoxanthum odoratum</i>	<i>Brachiaria dictyoneura</i>
<i>Aeschynomene afraspera</i>	<i>Antigonon leptopus</i>	<i>Brachiaria distachya</i>
<i>Aeschynomene</i>	<i>Arachis hypogaea</i>	<i>Brachiaria eruciformis</i>
<i>americana</i>	<i>Arachis pintoi</i>	<i>Brachiaria lata</i>
<i>Aeschynomene aspera</i>	<i>Argemone mexicana</i>	<i>Brachiaria mutica</i>
<i>Aeschynomene indica</i>	<i>Argyreia nervosa</i>	<i>Brachiaria nana</i>
<i>Aeschynomene sensitiva</i>	<i>Arisarum simorrhinum</i>	<i>Brachiaria reptans</i>
<i>Ageratina riparia</i>	<i>Aristida adscensionis</i>	<i>Brachiaria umbellata</i>
<i>Ageratum conyzoides</i>	<i>Aristida hordeacea</i>	<i>Brachiaria villosa</i>
<i>Ageratum houstonianum</i>	<i>Aristida kerstingii</i>	<i>Brassica carinata</i>
<i>Agrimonia eupatoria</i>	<i>Artemisia annua</i>	<i>Brassica juncea</i>
<i>Agrostis stolonifera</i>	<i>Artemisia verlotiorum</i>	<i>Brassica rapa</i>
<i>Albizia lebeck</i>	<i>Arundo donax</i>	<i>Brassica souliei</i>
<i>Alectra vogelii</i>	<i>Asclepias curassavica</i>	<i>Bromus catharticus</i>
<i>Alisma lanceolatum</i>	<i>Asemeia violacea</i>	<i>Bromus diandrus</i>
<i>Alisma plantago-aquatica</i>	<i>Aspilia kotschyi</i>	<i>Bromus hordeaceus</i>
<i>Allium neapolitanum</i>	<i>Aster squamatus</i>	<i>Bromus lanceolatus</i>
<i>Alternanthera</i>	<i>Asystasia gangetica</i>	<i>Bromus rigidus</i>
<i>philoxeroides</i>	<i>Asystasia gangetica</i>	<i>Bryonia dioica</i>
<i>Alternanthera pungens</i>	<i>Atractylis cancellata</i>	<i>Bulbostylis barbata</i>
<i>Alternanthera sessilis</i>	<i>Austroeupatorium</i>	<i>Bulbostylis hispidula</i>
<i>Alysicarpus ovalifolius</i>	<i>inulaefolium</i>	<i>Bupleurum lancifolium</i>
<i>Alysicarpus rugosus</i>	<i>Avena ludoviciana</i>	<i>Butomus umbellatus</i>
<i>Amaranthus blitum</i>	<i>Avena sativa</i>	<i>Caesalpinia bonduc</i>
<i>Amaranthus dubius</i>	<i>Avena sterilis</i>	<i>Caesulia axillaris</i>
<i>Amaranthus graecizans</i>	<i>Avena strigosa</i>	<i>Cajanus cajan</i>
<i>Amaranthus hybridus</i>	<i>Axonopus compressus</i>	<i>Cajanus scarabaeoides</i>
<i>Amaranthus interruptus</i>	<i>Axonopus purpusii</i>	<i>Calendula arvensis</i>

Calopogonium mucunoides	Chrysanthellum americanum	Cosmos caudatus
Canavalia	Chrysopogon aciculatus	Costus arabicus
Canavalia ensiformis	Chrysopogon nigritanus	Courtoisina cyperoides
Canavalia gladiata	Chrysopogon zizanioides	Crassocephalum crepidioides
Canna indica	Cichorium intybus	Crocoshmia crocosmiiflora
Caperonia palustris	Cinnamomum verum	Crotalaria brevidens
Capnophyllum peregrinum	Cirsium arvense	Crotalaria goreensis
Capsella bursa-pastoris	Cirsium vulgare	Crotalaria grahamiana
Cardamine hirsuta	Cissus erosa	Crotalaria incana
Cardiospermum microcarpum	Cissus quadrangularis	Crotalaria juncea
Carduus meoanthus	Cissus verticillata	Crotalaria ochroleuca
Carduus pycnocephalus	Citrullus lanatus	Crotalaria ononoides
Carex otrubae	Cladanthus arabicus	Crotalaria pallida
Carlina racemosa	Cladanthus mixtus	Crotalaria retusa
Cassytha filiformis	Cleome aculeata	Crotalaria senegalensis
Catharanthus roseus	Cleome coeruleorosea	Crotalaria spectabilis
Cecropia peltata	Cleome gynandra	Crotalaria trichotoma
Celosia argentea	Cleome hirta	Croton bonplandianus
Celosia trigyna	Cleome monophylla	Croton hirtus
Cenchrus americanus	Cleome rutidosperma	Croton trinitatis
Cenchrus biflorus	Cleome viscosa	Cryptostegia grandiflora
Cenchrus echinatus	Clibadium surinamense	Cucumis melo
Centaurea acaulis	Clidemia hirta	Cuphea carthagenensis
Centaurea benedicta	Clitoria arborescens	Curculigo pilosa
Centaurea eriophora	Clitoria heterophylla	Cuscuta campestris
Centaurea melitensis	Coccinia grandis	Cyanotis axillaris
Centella asiatica	Coix lacryma-jobi	Cyanotis lanata
Centrosema brasilianum	Colchicum autumnale	Cyanthillium cinereum
Centrosema pubescens	Coldenia procumbens	Cyathula prostrata
Cerastium glomeratum	Colocasia esculenta	Cyclosporum leptophyllum
Ceratophyllum demersum	Commelina africana	Cynanchum callialatum
Ceratopteris cornuta	Commelina benghalensis	Cynara cardunculus
Ceratopteris thalictroides	Commelina diffusa	Cynodon dactylon
Chamaecrista diphylla	Commelina diffusa	Cyperus aggregatus
Chamaecrista mimosoides	Commelina erecta	Cyperus amabilis
Chamaeleon gummifer	Commelina forskoolii	Cyperus articulatus
Chenopodium murale	Commelina nigritana	Cyperus articulatus
Chenopodium album	Commelina subulata	Cyperus balfourii
Chloris barbata	Conium maculatum	Cyperus compressus
Chloris gayana	Convolvulus althaeoides	Cyperus cyperoides
Chloris pilosa	Convolvulus arvensis	Cyperus difformis
Chloris pycnothrix	Convolvulus tricolor	Cyperus distans
Chloris virgata	Conyza sumatrensis	Cyperus eragrostis
Chromolaena odorata	Corchorus aestuans	Cyperus esculentus
Chrozophora tinctoria	Corchorus fascicularis	Cyperus exaltatus
	Corchorus olitorius	Cyperus fuscus
	Corchorus tridens	Cyperus giganteus
	Corchorus trilocularis	Cyperus glomeratus
	Coronilla scorpioides	Cyperus gracilis

Cyperus haspan	Digitaria horizontalis	Eragrostis cylindriflora
Cyperus iria	Digitaria humbertii	Eragrostis japonica
Cyperus laxus	Digitaria insularis	Eragrostis patula
Cyperus luzulae	Digitaria longiflora	Eragrostis tremula
Cyperus obtusatus	Digitaria radicata	Eragrostis turgida
Cyperus odoratus	Dinebra retroflexa	Eragrostis unioloidea
Cyperus papyrus	Diplachne fusca	Erigeron bonariensis
Cyperus podocarpus	Dracontium polyphyllum	Erigeron canadensis
Cyperus prolifer	Drosera burmanni	Erigeron karvinskianus
Cyperus pulcherrimus	Drymaria cordata	Erigeron sumatrensis
Cyperus pustulatus	Dysphania ambrosioides	Eriocaulon truncatum
Cyperus rotundus	Ecballium elaterium	Eriochloa polystachya
Cyperus sphacelatus	Echinochloa colonum	Eriochloa procera
Cyperus squarrosus	Echinochloa crus-galli	Erodium ciconium
Cyperus surinamensis	Echinochloa crus-pavonis	Erodium malacoides
Cyperus tenuispica	Echinochloa obtusiflora	Eruca vesicaria
Dactylis glomerata	Echinochloa oryzoides	Eryngium triquetrum
Dactyloctenium	Echinochloa polystachya	Euphorbia chamaesyce
aegyptium	Echinochloa pyramidalis	Euphorbia convolvuloides
Dalechampia scandens	Echinochloa stagnina	Euphorbia cyathophora
Datura innoxia	Echinolaena inflexa	Euphorbia forskalii
Datura stramonium	Echinops strigosus	Euphorbia heterophylla
Daucus carota	Eclipta prostrata	Euphorbia hirta
Delphinium halteratum	Eichhornia azurea	Euphorbia hypericifolia
Delphinium peregrinum	Eichhornia crassipes	Euphorbia hyssopifolia
Dentella repens	Elatine triandra	Euphorbia peplus
Desmanthus virgatus	Eleocharis acutangula	Euphorbia prostrata
Desmodium adscendens	Eleocharis complanata	Fagopyrum esculentum
Desmodium barbatum	Eleocharis dulcis	Fagopyrum tataricum
Desmodium dichotomum	Eleocharis filiculmis	Fedia cornucopiae
Desmodium	Eleocharis interstincta	Filago pyramidata
heterophyllum	Eleocharis limosa	Fimbristylis dichotoma
Desmodium hirtum	Eleocharis minuta	Fimbristylis ferruginea
Desmodium incanum	Eleocharis mutata	Fimbristylis littoralis
Desmodium intortum	Eleocharis retroflexa	Fimbristylis ovata
Desmodium	Elephantopus mollis	Fimbristylis
ramosissimum	Eleusine indica	quinquangularis
Desmodium tortuosum	Eleutheranthera ruderalis	Fimbristylis umbellaris
Desmodium triflorum	Emex spinosa	Flemingia macrophylla
Desmodium uncinatum	Emilia citrina	Flemingia strobilifera
Dianella ensifolia	Emilia fosbergii	Flueggea virosa
Dichanthium annulatum	Emilia humifusa	Foeniculum vulgare
Dichondra repens	Emilia sonchifolia	Fuirena ciliaris
Dichrostachys cinerea	Epilobium hirsutum	Fuirena stricta
Dicliptera paniculata	Epilobium tetragonum	Fuirena umbellata
Dicoma tomentosa	Eragrostis amabilis	Fumaria muralis
Digitaria argillacea	Eragrostis aspera	Fumaria officinalis
Digitaria bicornis	Eragrostis atrovirens	Fumaria parviflora
Digitaria ciliaris	Eragrostis cilianensis	Furcraea foetida
Digitaria didactyla	Eragrostis ciliaris	Galinsoga parviflora

Galium aparine	Hydrocharis	Ischaemum rugosum
Galium tricornutum	morsus-ranae	Ischaemum timorense
Geropogon hybridus	Hydrocotyle bonariensis	Jacquemontia tamnifolia
Gladiolus italicus	Hydrocotyle umbellata	Jatropha gossypiifolia
Glaucium corniculatum	Hydrolea glabra	Juncus acutus
Glebionis coronaria	Hydrolea zeylanica	Juncus maritimus
Glebionis segetum	Hymenachne	Justicia ciliata
Glinus lotoides	amplexicaulis	Kohautia tenuis
Glinus oppositifolius	Hypocoum procumbens	Kyllinga brevifolia
Gloriosa superba	Hypericum japonicum	Kyllinga bulbosa
Gnaphalium purpureum	Hypochaeris radicata	Kyllinga erecta
Gomphocarpus	Hyptis atrorubens	Kyllinga nemoralis
physocarpus	Hyptis lanceolata	Kyllinga pumila
Gomphrena celosioides	Hyptis pectinata	Kyllinga squamulata
Grona heterocarpos	Hyptis spicigera	Kyllinga tenuifolia
Guizotia abyssinica	Hyptis suaveolens	Lablab purpureus
Habranthus robustus	Impatiens auricoma	Lactuca indica
Hackelochloa granularis	Impatiens walleriana	Lagarosiphon
Hedychium flavescens	Imperata brasiliensis	madagascariensis
Hedychium gardnerianum	Imperata cylindrica	Laggera crispata
Hedypnois rhagadioloides	Indigofera dendroides	Lagurus ovatus
Helianthus annuus	Indigofera hendecaphylla	Lamarckia aurea
Heliconia psittacorum	Indigofera hirsuta	Lamium amplexicaule
Heliotropium	Indigofera linifolia	Lantana camara
amplexicaule	Indigofera pilosa	Lapsana communis
Heliotropium europaeum	Indigofera stenophylla	Lathyrus aphaca
Heliotropium indicum	Indigofera suffruticosa	Lathyrus cicera
Heliotropium ovalifolium	Indigofera tinctoria	Lathyrus inconspicuus
Heliotropium strigosum	Ipomoea acanthocarpa	Launaea cornuta
Helosciadium nodiflorum	Ipomoea alba	Leersia hexandra
Hemarthria altissima	Ipomoea aquatica	Leersia oryzoides
Herniaria hirsuta	Ipomoea asarifolia	Lemna aequinoctialis
Heteranthera callifolia	Ipomoea cairica	Leonotis nepetifolia
Heteranthera limosa	Ipomoea coptica	Leonurus sibiricus
Heteranthera reniformis	Ipomoea dichroa	Leopoldia comosa
Heterotis rotundifolia	Ipomoea eriocarpa	Lepidium africanum
Hexasepalum scandens	Ipomoea hederifolia	Lepidium didymum
Hibiscus asper	Ipomoea indica	Lepidium draba
Hibiscus bifurcatus	Ipomoea nil	Leptadenia
Hibiscus panduriformis	Ipomoea obscura	madagascariensis
Hibiscus sabdariffa	Ipomoea purpurea	Leptochloa caerulea
Hibiscus surattensis	Ipomoea quamoclit	Leptochloa scabra
Hippeastrum reginae	Ipomoea setifera	Lepturus radicans
Hippobroma longiflora	Ipomoea tiliacea	Leucaena leucocephala
Hirschfeldia incana	Ipomoea triloba	Leucas lavandulifolia
Holcus lanatus	Iris pseudacorus	Leucas martinicensis
Homolepis aturensis	Isachne globosa	Leucas zeylanica
Hordeum murinum	Ischaemum afrum	Limnocharis flava
Hydrilla verticillata	Ischaemum aristatum	Limnophila geoffrayi
	Ischaemum ciliare	Linaria reflexa

Linaria triphylla	Melia azedarach	Ocimum americanum
Lindernia antipoda	Melilotus indicus	Ocimum gratissimum
Lindernia ciliata	Melilotus sulcatus	Oenothera glazioviana
Lindernia crustacea	Melinis repens	Oenothera rosea
Lindernia dubia	Melochia corchorifolia	Oenothera stricta
Lindernia nummulariifolia	Melochia pyramidata	Oldenlandia brachypoda
Lindernia oblonga(Benth.)	Mentha aquatica	Oldenlandia corymbosa
Lindernia	Mentha suaveolens	Oldenlandia erecta
procumbens(Krock.)	Mercurialis ambigua	Oldenlandia lancifolia
Lindernia pusilla(Willd.)	Merremia aegyptia	Ononis spinosa
Lipocarpha chinensis	Merremia dissecta	Onopordum
Lipocarpha sphacelata	Merremia emarginata	macracanthum
Lithospermum arvense	Microglossa pyrifolia	Operculina turpethum
Litsea glutinosa	Micropus supinus	Oplismenus burmanni
Litsea monopetala	Microstachys corniculata	Oplismenus hirtellus
Lobelia cliffortiana	Microtea debilis	Opuntia stricta
Lolium arundinaceum	Mikania micrantha	Ornithogalum algeriense
Lolium rigidum	Mimosa diplotricha	Ornithogalum
Ludwigia abyssinica	Mimosa pigra	narbonense
Ludwigia adscendens	Mimosa pudica	Orobanche mutelii
Ludwigia erecta	Mirabilis jalapa	Oryza barthii
Ludwigia grandiflora	Misopates calycinum	Oryza glaberrima
Ludwigia hyssopifolia	Mitracarpus hirtus	Oryza longistaminata
Ludwigia leptocarpa	Mollugo nudicaulis	Oryza rufipogon
Ludwigia octovalvis	Mollugo pentaphylla	Oryza sativa
Ludwigia peploides	Mollugo verticillata	Ottelia ulvifolia
Lycopus europaeus	Momordica charantia	Oxalis barrelieri
Lysimachia	Monochoria cyanea	Oxalis corniculata
linum-stellatum	Monochoria vaginalis	Oxalis debilis
Lysimachia vulgaris	Montrichardia	Oxalis latifolia
Lythrum hyssopifolia	arborescens	Oxalis pes-caprae
Lythrum salicaria	Mucuna pruriens	Oxalis tetraphylla
Macroptilium	Murdannia nudiflora	Paederia farinosa
atropurpureum	Narcissus serotinus	Paederia foetida
Macroptilium lathyroides	Nelsonia canescens	Pallenis spinosa
Malvastrum	Neojeffreya decurrens	Panicum antidotale
coromandelianum	Neonotonia wightii	Panicum brevifolium
Malva olbia	Neptunia oleracea	Panicum humile
Malva parviflora	Neptunia plena	Panicum laetum
Malva sylvestris	Neslia apiculata	Panicum maximum
Malva trimestris	Neustanthus	Panicum mertensii
Mantiscalca salmantica	phaseoloides	Panicum pansum
Marrubium vulgare	Nicandra physalodes	Panicum pilosum
Marsilea minuta	Nigella arvensis	Panicum
Marsypianthes	Notobasis syriaca	pseudowoeltzkowii
chamaedrys	Nuphar lutea	Panicum repens
Matourea scutellarioides	Nymphaea maculata	Panicum subalbidum
Matthiola parviflora	Nymphaea nouchali	Papaver rhoeas
Medicago polymorpha	Nymphaea rudgeana	Paronychia argentea
Melaleuca quinquenervia	Nymphoides indica	

Parthenium	Phyllanthus tenellus	Rhynchosia malacophylla
hysterophorus	Phyllanthus urinaria	Rhynchosia minima
Paspalidium geminatum	Phyllanthus virgatus	Rhynchosia viscosa
Paspalum conjugatum	Physalis angulata	Rhynchospora
Paspalum dilatatum	Physalis lagascae	cephalotes
Paspalum distichum	Phytolacca americana	Rhynchospora
Paspalum paniculatum	Picnomon acarna	corymbosa
Paspalum plicatum	Picris cupuligera	Rhynchospora
Paspalum scrobiculatum	Pilea microphylla	holoschoenoides
Paspalum urvillei	Pistia stratiotes	Rhynchospora longisetis
Paspalum vaginatum	Pithecellobium dulce	Rhynchospora pubera
Paspalum virgatum	Plantago afra	Rhynchospora rubra
Passiflora foetida	Plantago coronopus	Rhynchospora rugosa
Passiflora suberosa	Plantago lagopus	Rhynchospora tenerrima
Passiflora subpeltata	Plantago lanceolata	Rhynchospora trispicata
Pavonia cancellata	Plantago major	Richardia scabra
Pennisetum cafferum	Poa annua	Ricinus communis
Pennisetum clandestinum	Polycarpha corymbosa	Ridolfia segetum
Pennisetum pedicellatum	Polydora serratuloides	Rivina humilis
Pennisetum polystachion	Polygala paniculata	Roemeria hispida
Pentanema indicum	Polygonum aviculare	Rolandra fruticosa
Pentas lanceolata	Polymeria pusilla	Rostraria cristata
Pentodon pentandrus	Polypogon monspeliensis	Rottboellia
Peperomia pellucida	Polypogon viridis	cochinchinensis
Perotis patens	Pombalia calceolaria	Rubus alceifolius
Persicaria acuminata	Portulaca grandiflora	Rubus fraxinifolius
Persicaria barbata	Portulaca oleracea	Rubus rosifolius
Persicaria chinensis	Portulaca quadrifida	Rumex abyssinicus
Persicaria ferruginea	Potamogeton nodosus	Rumex crispus
Persicaria lapathifolia	Potamogeton pusillus	Rumex dentatus
Persicaria maculosa	Priva lappulacea	Sacciolepis africana
Persicaria mitis	Prospero autumnale	Sacciolepis indica
Persicaria salicifolia	Psidium guajava	Sacciolepis striata
Persicaria senegalensis	Psychine stylosa	Sagittaria guayanensis
Phacelia tanacetifolia	Pteridium aquilinum	Salpichroa organifolia
Phalaris arundinacea	Pterolepis glomerata	Salsola kali
Phalaris brachystachys	Pycnus flavescens	Salvia algeriensis
Phalaris minor	Pycnus lanceolatus	Salvinia adnata
Phaseolus lunatus	Pycnus macrostachyos	Salvinia nymphellula
Phaulopsis imbricata	Pycnus mundtii	Sambucus ebulus
Phaulopsis verticillaris	Pycnus polystachyos	Samolus valerandi
Phenax sonneratii	Ranunculus arvensis	Sanguisorba minor
Phragmites australis	Ranunculus sceleratus	Sannantha leratii
Phthirusa pyrifolia	Raphanus raphanistrum	Sauvagesia erecta
Phyllanthus amarus	Raphanus sativus	Scandix pecten-veneris
Phyllanthus	Rapistrum rugosum	Schenkia spicata
maderaspatensis	Reichardia tingitana	Schinus terebinthifolia
Phyllanthus niruri	Reseda alba	Schoenoplectiella
Phyllanthus niruroides	Rhagadiolus stellatus	juncoides
Phyllanthus orbicularis	Rhamphicarpa fistulosa	

Schoenoplectiella mucronata	Silene rubella	Sporobolus africanus
Schoenoplectiella perrieri	Silene vulgaris	Sporobolus coromandelianus
Schoenoplectiella senegalensis	Silybum marianum	Sporobolus festivus
Schoenoplectiella supina	Sinapis	Sporobolus indicus
Schoenoplectus lacustris	Sinapis alba	Sporobolus pyramidalis
Schwenckia americana	Sinapis arvensis	Stachytarpheta cayennensis
Scirpoides holoschoenus	Sipanea pratensis	Stachytarpheta indica
Scleria depressa	Siphonochilus aethiopicus	Stachytarpheta jamaicensis
Scleria gaertneri	Sisymbrium irio	Stachytarpheta urticifolia
Scleria microcarpa	Sisyrinchium micranthum	Steinchisma laxum
Scleria tessellata	Solanum americanum	Stellaria media
Scleria verrucosa	Solanum elaeagnifolium	Stenotaphrum dimidiatum
Scleria vogelii	Solanum mauritianum	Stigmaphyllon convolvulifolium
Scolymus grandiflorus	Solanum nigrum	Striga asiatica
Scolymus hispanicus	Solanum rugosum	Striga aspera
Scoparia dulcis	Solanum seafortianum	Striga gesnerioides
Scorpiurus muricatus	Solanum stramonifolium	Striga hermonthica
Scorpiurus vermiculatus	Solanum subinerme	Stylosanthes guianensis
Senecio leucanthemifolius	Solanum torvum	Synedrella nodiflora
Senecio vulgaris	Solanum villosum	Tacca leontopetaloides
Senna alata	Solanum violaceum	Tecoma stans
Senna hirsuta	Solidago canadensis	Tephrosia bracteolata
Senna obtusifolia	Sonchus asper	Tephrosia linearis
Senna occidentalis	Sonchus maritimus	Tephrosia noctiflora
Senna tora	Sonchus oleraceus	Tephrosia purpurea
Sesamum radiatum	Sorghum arundinaceum	Teramnus labialis
Sesbania bispinosa	Sorghum halepense	Tetragonia tetragonioides
Sesbania madagascariensis	Spathodea campanulata	Thalia geniculata
Sesbania pachycarpa	Spergularia flaccida	Thelepogon elegans
Sesbania sesban	Spergula arvensis	Themeda quadrivalvis
Setaria barbata	Spermacoce alata	Thesium humile
Setaria intermedia	Spermacoce capitata	Thrasya granitica
Setaria parviflora	Spermacoce	Thunbergia alata
Setaria pumila	chaetocephala	Thunbergia grandiflora
Setaria punctata(Burm.f.)	Spermacoce latifolia	Thunbergia laevis
Setaria verticillata	Spermacoce ocymifolia	Tilesia baccata
Sherardia arvensis	Spermacoce pusilla	Tithonia diversifolia
Sicyos angulatus	Spermacoce radiata	Tonina fluviatilis
Sida acuta	Spermacoce stachydea	Torilis nodosa
Sida cordifolia	Spermacoce verticillata	Trianthema portulacastrum
Sida glabra	Sphaeranthus angustifolius	Tribulus cistoides
Sida rhombifolia	Sphaeranthus cotuloides	Tribulus terrestris
Sida rhombifolia	Sphaeranthus senegalensis	Trichodesma indicum
Sida urens	Sphaeranthus suaveolens	Trichodesma zeylanicum
Sigesbeckia orientalis	Sphagneticola trilobata	Tricliceras pilosum
Silene muscipula	Sphenoclea zeylanica	

Tridax procumbens	Vaccaria hispanica	Vigna unguiculata
Tristemma mauritianum	Valerianella discoidea	Vitex trifolia
Triumfetta pentandra	Varronia curassavica	Vossia cuspidata
Triumfetta rhomboidea	Verbena bonariensis	Waltheria indica
Tropaeolum majus	Verbena brasiliensis	Wikstroemia indica
Turgenia latifolia	Verbena officinalis	Xanthium orientale
Turnera subulata	Verbena rigida	Xanthium spinosum
Turnera ulmifolia	Vernoniastrum ambiguum	Xanthium strumarium
Typha domingensis	Vernonia chapmanii	Xenostegia medium
Typha latifolia	Veronica	Xyris indica
Typha laxmannii	anagallis-aquatica	Youngia japonica
Uraria lagopodioides(L.)	Veronica cymbalaria	Zaleya pentandra
Uraria picta	Vicia benghalensis	Zehneria thwaitesii
Urena lobata	Vicia hirsuta	Ziziphus lotus
Urochloa brizantha	Vicia sativa	Zornia glochidiata
Urochloa eminii	Vicia villosa	Zornia latifolia
Urtica urens	Vigna luteola	
Utricularia aurea	Vigna radiata	
Utricularia foliosa	Vigna unguiculata	

Annexe 10 : Liste des espèces de la flore adventice tropicale retenue pour cette étude (correspondant à la liste d'espèces référencées sur la plateforme wiktrop.org).