

Phase 5 – Analyse des données

Rapport exploratoire — Estimation de la biomasse forestière dans les DROM à partir de données LiDAR et d’inventaires terrain

Date de soumission	Version	Rédacteurs	Relecteurs	Commentaires
26/03/2026	1.0	Stevie Tickeng, ONF International	• Cédric Lardeux, ONF International • Aurore Oberlé, ONF International	Ces travaux ont été menés avec l'appui de Grégoire Vincent, UMR AMAP, IRD.

A	Résumé exécutif.....	3
B	Introduction : objectifs et démarche.....	6
C	Données utilisées.....	9
	C.1 Données tests (IRD UMR AMAP)	9
	C.2 Données LIDAR recueillies par GEOFIT	9
D	Méthodologie développée pour la Guyane	10
	D.1 Calcul de la biomasse de référence	11
	D.2 Segmentation individuelles	17
	D.3 Modélisation statistique (régression multivariée)	20
E	Analyse de faisabilité du déploiement à La Réunion : verrous identifiés et perspectives	24
	E.1 Le verrou technologique : La Dépendance au Canopy Constructor	24
	E.2 L'alternative à court terme pour La Réunion	25
	E.3 L'obstacle commun de l'échantillonnage (Guyane et Réunion)	26
F	Synthèse des outils évalués et recommandations	26
	F.1 Recommandations globales pour un passage à l'échelle opérationnelle.....	26
	F.2 Synthèse des outils évalués et recommandations	28
	F.3 Conclusion de synthèse	31
	RÉFÉRENCES BIBLIOGRAPHIQUES.....	32
	ANNEXES	34

A RESUME EXECUTIF

Objectif du travail

ONF International a conduit une phase exploratoire visant à évaluer la faisabilité technique de l'estimation de la biomasse aérienne (AGB) des forêts tropicales des DROMs à partir de données LiDAR et d'inventaires forestiers. Cette étude de R&D identifie les outils pertinents, leurs limites structurelles et les conditions nécessaires à un passage à l'échelle opérationnel. Elle répond au biais systématique des allométries pantropicales universelles qui, en ignorant les variations locales de hauteur, sous-estiment la biomasse réelle de 11 à 13 % ([Huertas et al.,2025](#)). L'objectif de ce travail n'est pas, à ce stade, de produire une cartographie opérationnelle de la biomasse à grande échelle, mais de tester, calibrer et qualifier des chaînes méthodologiques, d'en identifier les verrous techniques et statistiques, et de formuler des recommandations en vue d'un passage ultérieur à l'échelle.

Apports majeurs des données LiDAR très haute densité

Les tests menés reposent sur des nuages de points LiDAR aéroportés d'une densité exceptionnelle (~ 1300 points/m²), très supérieure aux standards usuels. Cette hyperdensité permet une restitution fine de la topographie sous couvert forestier et une meilleure caractérisation de la structure verticale de la canopée, condition indispensable à une estimation fiable des hauteurs d'arbres et, par conséquent, de la biomasse. Ce choix méthodologique est conforté par les travaux récents de [Huertas et al.,2025](#), qui montrent que les instruments classiques de mesure de hauteur (clinomètres, télémètres) présentent une précision insuffisante sous couvert dense et fermé, rendant les données terrain de hauteur sujettes à des biais importants. L'ALS (LiDAR aéroporté) constitue ainsi une alternative robuste : en fournissant des modèles de hauteur de canopée (MNHC) à haute résolution sur de grandes surfaces, il permet de dériver des allométries H-DBH locales et spécifiques aux espèces, sans les contraintes de coût et d'échelle inhérentes aux relevés terrain ou au LiDAR terrestre (TLS).

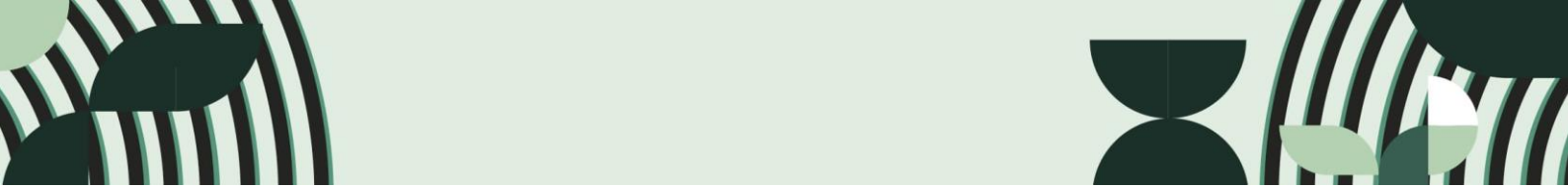
Calcul de la biomasse de référence : cas de la Guyane

En Guyane, la biomasse de référence a été estimée à partir des inventaires terrain en combinant des allométries locales récentes, calibrées par LiDAR, et l'équation pantropicale de référence de [Chave et al., 2014](#). Cette approche hybride est justifiée par les résultats de [Huertas et al., 2025](#), conduits sur le site de Paracou (Guyane française) : le remplacement d'une allométrie pantropicale universelle par une relation H-DBH locale dérivée de données ALS et intégrant l'identité spécifique des arbres a permis de réduire de près de moitié l'erreur moyenne de prédiction de hauteur, et a révélé que les allométries globales sous-estimaient la biomasse locale de 11 à 13 % (soit 40 à 54 t ha⁻¹). Cette approche hybride permet donc de réduire significativement les biais associés aux modèles globaux ne prenant pas en compte la hauteur réelle des arbres. Les analyses de propagation des incertitudes montrent que la biomasse estimée à l'échelle des placettes (UE : unité d'échantillonnage) est robuste, avec une variabilité limitée ($\pm 0,7$ à $\pm 1,1$ t/ha à l'échelle de la placette). Les résultats de biomasse par placette varient entre 370 et 444 t/ha.

Segmentation des arbres et extraction de métriques LiDAR

La deuxième étape du pipeline vise à segmenter les arbres individuels afin d'extraire des métriques structurelles (nombre de tiges, hauteur, diamètre de couronne) qui serviront de variables explicatives au modèle de régression (Random Forest) qui prédira simultanément à l'échelle placette : l'AGB, la densité de tiges, le diamètre quadratique moyen. Des méthodes géométriques non supervisées ont été retenues ([Li et al., 2012](#) ; [Silva et al., 2016](#)) en raison de leur rapidité de déploiement et de leur absence de besoin en données annotées sur les données LiDAR. A partir des orthophotos, les outils **Detectree2** et **SAM3** (Segment Anything Model, 2025) ont également été testés et combinés pour améliorer la détection des houppiers dominants.

Ces méthodes produisent des résultats physiquement cohérents mais structurellement limités : environ 97 houppiers sont détectés par placette sur le nuage de points ALS, alors que l'inventaire terrain recense environ 400 tiges/ha. Cette dissymétrie confirme que la segmentation actuelle ne capture que la strate dominante de la canopée, et reste aveugle au sous-étage. En conséquence, les variables explicatives LiDAR (X)



ne décrivent que partiellement la structure forestière que le modèle devra prédire, ce qui renforce la nécessité d'évaluer des outils de segmentation plus pénétrants, notamment **ComputeTree** (ONF).

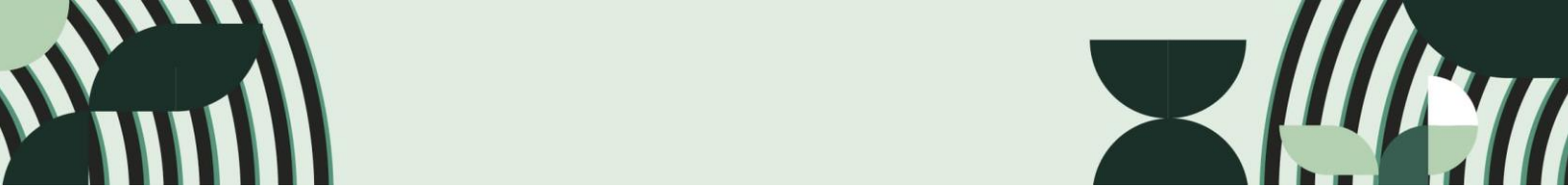
Calcul de la biomasse de référence à La Réunion : spécificités et verrou technologique

Contrairement au cas de la Guyane, le pipeline de traitement n'a pas pu être appliqué en l'état aux données de La Réunion.

- **Dépendance logicielle** : Le transfert se heurte à un verrou technologique lié au Canopy Constructor, qui nécessite une calibration spécifique avec des allométries locales réunionnaises, actuellement en phase de recherche.
- **Alternative à court terme** : Il reste possible de recourir à des allométries pantropicales ([Chave et al., 2014](#) ; [Feldpausch et al., 2011](#)) pour La Réunion. Toutefois, cette option expose l'estimation à des biais structurels. D'une part, l'omission de la hauteur réelle engendre des biais systématiques majeurs dans les bilans carbone ([Feldpausch et al., 2012](#)), et maintient une erreur relative d'estimation de la biomasse de l'arbre à 27,7 % (contre 15,3 % lorsque la variable de hauteur est incluse selon [Chave et al., 2014](#)). D'autre part, des travaux récents ont révélé que l'application de ces allométries globales sous-estime la biomasse locale de 11 à 13 % par rapport à un modèle calibré par LiDAR ([Huertas et al., 2025](#)).

Limites communes et recommandations

Bien que l'architecture logicielle de traitement soit techniquement fonctionnelle, l'entraînement d'un modèle prédictif par Machine Learning (Random Forest) est actuellement prématuré. Il se heurte au sous-échantillonnage spatial : en Guyane, seules 4 placettes disposaient simultanément de données LiDAR et d'inventaires terrains exploitables à cause du changement de protocole effectué en cours de collecte. La Réunion souffre du même obstacle opérationnel, rendant impossible l'entraînement d'un modèle sans risquer un surapprentissage. La littérature indique un minimum de 40 à 50 placettes pour stabiliser un modèle prédictif de biomasse ([Zolkos](#)



et al., 2013), voire plus de 60 en milieu tropical complexe pour garantir l'indépendance spatiale entre données d'entraînement et de validation (Ploton et al., 2020). Par ailleurs, la segmentation actuelle ne capturant que la strate dominante, les variables explicatives LiDAR ne décrivent pas la biomasse totale de la forêt. Ce second verrou structural doit être levé simultanément.

- **Recommandations** : Le passage à l'échelle opérationnelle nécessite impérativement une augmentation des placettes couplées. Il est également préconisé d'évaluer des outils de reconstruction 3D plus pénétrants, comme ComputeTree (ONF) ou des algorithmes de deep learning, pour mieux capter les strates inférieures.

Conclusion

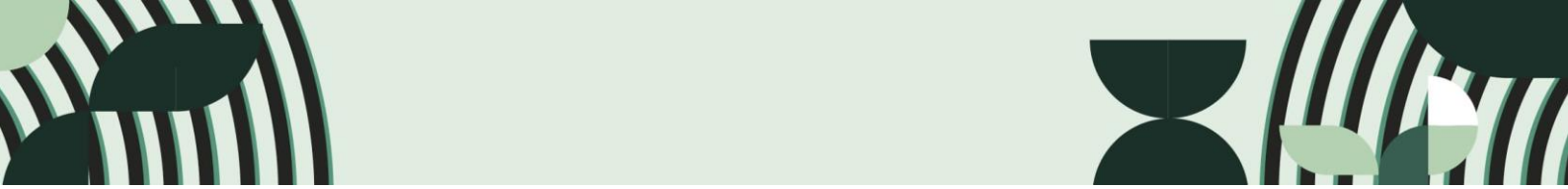
Cette phase exploratoire valide un pipeline de traitement complet et techniquement fonctionnel, de l'extraction des métriques LiDAR jusqu'à la production de la biomasse de référence en Guyane. En revanche, le déploiement opérationnel à grande échelle en Guyane comme à La Réunion reste conditionné à la levée de trois verrous identifiés:

- l'augmentation du nombre de placettes couplées (objectif : 60+ spatialement indépendantes) pour permettre des estimations de biomasse fiable
- l'amélioration de la segmentation pour couvrir l'ensemble du profil vertical
- et la calibration du Canopy Constructor pour La Réunion.

Ces leviers constituent la feuille de route prioritaire pour les prochaines phases.

B INTRODUCTION : OBJECTIFS ET DEMARCHE

L'estimation de la biomasse aérienne des forêts tropicales constitue un enjeu central pour le suivi des ressources forestières, l'évaluation des stocks de carbone et l'appui aux politiques publiques dans les Départements et Régions d'Outre-Mer (DROMs). En contexte tropical, cette estimation est rendue particulièrement complexe par la forte hétérogénéité structurelle des peuplements forestiers, caractérisés par des canopées multi-stratifiées, des sous-bois denses, une diversité floristique élevée et des reliefs

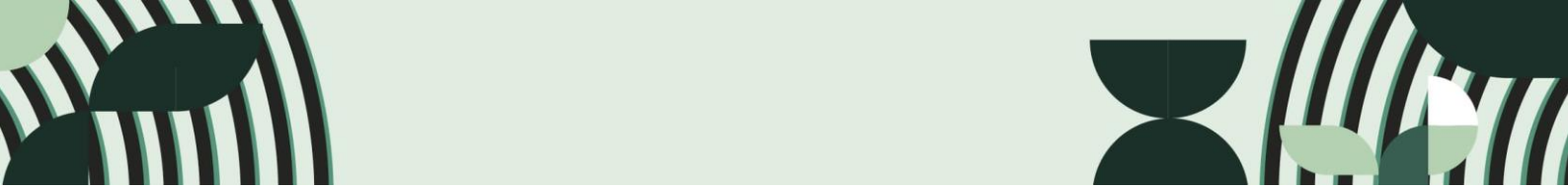


parfois marqués, qui limitent la transposabilité directe des approches développées en milieux tempérés.

Dans ce contexte, ONF International a engagé à partir de novembre 2025 une démarche exploratoire de recherche et développement visant à évaluer la capacité d'outils de télédétection, et en particulier du LiDAR aéroporté très haute densité, à améliorer et automatiser l'estimation de la biomasse aérienne (Above-Ground Biomass – AGB) à partir de données couplées LiDAR inventaires forestiers. L'objectif de ce travail n'est pas, à ce stade, de produire une cartographie opérationnelle de la biomasse à grande échelle, mais de tester, calibrer et qualifier des chaînes méthodologiques, d'en identifier les verrous techniques et statistiques, et de formuler des recommandations opérationnelles en vue d'un passage ultérieur à l'échelle.

La démarche adoptée repose sur deux phases distinctes et complémentaires :

- **Phase 1 (Tests et calibration)** : cette phase a consisté à ajuster les méthodes sur des jeux de données tests. Un travail préalable de recherche bibliographique a été mené afin d'identifier les approches pertinentes pour l'estimation de la biomasse, en élargissant l'analyse aux avancées récentes sur les techniques de segmentation des couronnes d'arbres en 2D et 3D, incluant notamment les algorithmes de Deep Learning et les méthodes classiques de détection. Par ailleurs, une étude approfondie des données disponibles et des protocoles envisageables a permis de sélectionner les outils les plus adaptés à l'objectif du projet. Cette première phase test s'est tenue sur une durée approximative allant de mi-novembre 2025 à janvier 2026.
- **Phase 2 (Application sur données finales)** : correspond à l'application du protocole calibré aux données LiDAR aéroportées et aux inventaires terrain définitifs, livrés mi-janvier 2026 pour la Guyane et La Réunion. Cette phase vise à évaluer la robustesse et la transférabilité des méthodes retenues dans des contextes forestiers contrastés, et à analyser de manière critique les performances obtenues lorsque l'on se confronte à des données de production réelles, avec leurs contraintes opérationnelles. Cette phase applicative du protocole s'est tenue de février 2026 à mi-mars 2026 incluant la rédaction de ce rapport.



Le présent document synthétise l'ensemble de cette démarche. Il décrit les données mobilisées, les méthodes testées, les résultats obtenus et les limites identifiées, en distinguant clairement ce qui relève de la validation technique du pipeline et ce qui constitue encore un verrou scientifique ou statistique. Il se conclut par une synthèse des outils évalués et par des recommandations destinées à orienter les futures phases d'acquisition de données et de déploiement opérationnel de l'estimation de la biomasse dans les DROMs.

C DONNEES UTILISEES

C.1 DONNÉES TESTS (IRD UMR AMAP)

Ces données ont servi exclusivement à la calibration des méthodes, avant réception des livrables finaux. Elles portent sur le site de Paracou (Guyane Française).

- **Données ForestScan (Paracou)** : collectées via l'archive CEDA, elles comprennent des données LiDAR drone (UAV-LS) d'une densité d'environ 100 points par mètre carré, couvrant certaines parcelles à Paracou ainsi que des produits dérivés (Modèle Numérique de Terrain, Modèle Numérique de Surface, Modèle Numérique de Canopée à 1 m de résolution, nuages de points classifiés). Des données de LiDAR terrestre (TLS) de 2022 sur 3 parcelles complètent ce jeu.
- **Données Guyafor** : base de données d'inventaire terrain fournie par l'ONF (Guyane), le CNRS et le CIRAD. Il s'agit d'une base de données de plus de 2 millions d'arbres en Guyane. Nous avons sélectionné des données d'inventaire terrains pour le site Paracou de 2023, sur lesquelles nous avons testé le protocole de calcul de la biomasse.
- **Couronnes délimitées terrain (2023)**¹ : données de délimitation manuelle de couronnes, utilisées comme référence de validation pour la segmentation automatique d'orthophoto. Ces données ont servi à calibrer le protocole exploratoire mis en place afin de le préparer à l'appliquer sur le jeu de données finales.

C.2 DONNÉES LIDAR RECUEILLIES PAR GEOFIT

Il s'agit des données finales de production sur lesquelles le protocole exploratoire est appliqué en Phase 2.

- **Guyane** : Nuage de point LiDAR aérien haute densité, nuage de point LiDAR terrestre et orthophotos haute résolution, livrés mi-janvier 2026, couplés aux relevés d'inventaire terrain finaux.

¹ Données fournies par le chercheur Grégoire VINCENT : IRD AMAP

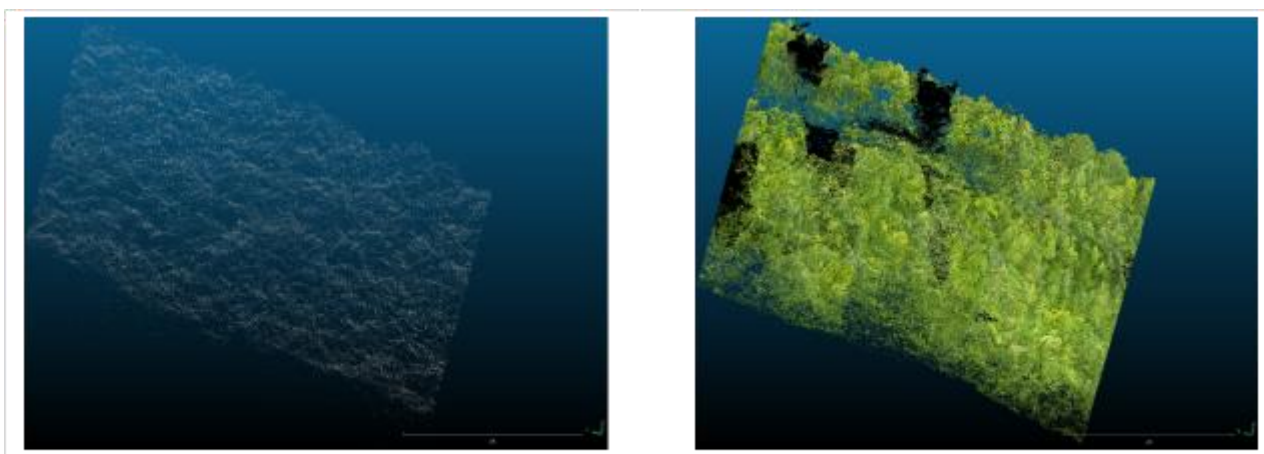
- **La Réunion** : même type de livrables, pour le déploiement en contexte insulaire tropical.

Pour la suite du processus, nous avons été limités dans la mise en place complète du protocole par les données. En effet, nous avons fait une analyse des données reçues par Geofit et avons conclu que seules 4 placettes sur 8 en Guyane seraient exploitables, car sur ces placettes nous avons une couverture LiDAR et des relevés terrains avec des arbres géoréférencés. Ceci est dû au changement de protocole effectué afin de couvrir plus de placettes (positionnement par sous-unité versus positionnement individuel des arbres).

D Méthodologie développée pour la Guyane

Dans le cadre du déploiement de cette méthodologie, il est indispensable de quantifier l'apport technique des données LiDAR. Le jeu de données LiDAR utilisé pour l'ensemble des tests présente une densité d'environ 1300 pts/m², largement supérieure aux standards nationaux (ex : LiDAR IGN HD ~10-15pts/m²). Le passage d'une résolution standard à une hyper-densité modifie fondamentalement la nature des traitements algorithmiques possibles et la précision de l'estimation de la biomasse, avec un impact sur l'aspect suivant :

Pénétration du signal et précision du Modèle Numérique de Terrain (MNT) : le calcul de la hauteur de chaque arbre (Modèle Numérique de Hauteur Canopée, MNHC) repose sur la soustraction entre le sommet de la végétation et le sol (MNHC = MNS - MNT). En forêt tropicale dense, et particulièrement sur les reliefs accidentés (cirques, ravines fortement végétalisées), l'effet d'occlusion est extrême. Avec 10 pts/m², seule une infime fraction des impulsions laser parvient à traverser le feuillage pour toucher le sol, générant un MNT interpolé incertain. L'hyper-densité du nuage de point ALS (1300 pts/m²) garantit une probabilité de pénétration interstitielle maximale. La topographie sous-jacente est ainsi modélisée avec une précision centimétrique, ce qui supprime le biais d'estimation de la hauteur totale de l'arbre, variable la plus sensible dans le calcul de la biomasse (AGB). Ci-dessous une illustration de la différence de densité de bois entre deux nuages de points distinct sur une même zone (à La Réunion) :



Nuage de point LiDAR HD-IGN (10pts/m²) à gauche et nuage de point Geofit (1300 pts/m²) à droite

FIGURE 1 : COMPARAISON DE LA DENSITE DE NUAGES DE POINTS LIDAR SUR UNE MEME ZONE A LA REUNION

La différence de pénétration du signal sous la canopée est visible, avec une restitution nettement plus précise du sol et des strates végétales inférieures dans le cas hyper-dense, on parvient même à voir la canopée en sous-bois.

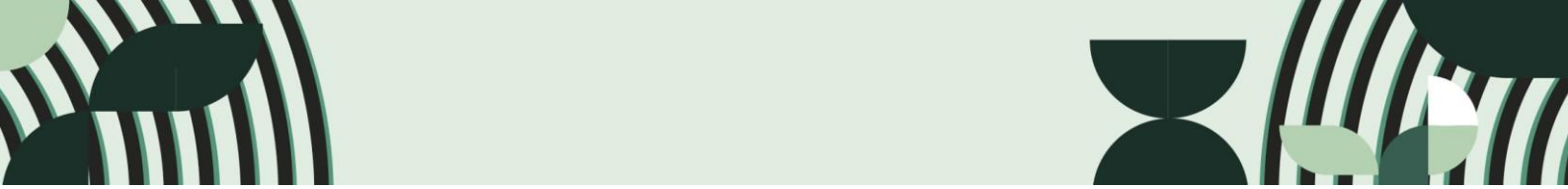
La méthodologie développée repose sur une succession de trois étapes : le calcul d'une biomasse de référence, qui servira de variable cible au modèle de régression ; la segmentation des arbres individuels pour extraire des métriques LiDAR; et enfin l'entraînement du modèle de régression.

D.1 Calcul de la biomasse de référence

Objectif : estimer une valeur de biomasse de référence par placette, qui avec la densité de bois, et DBH moyen serviront de variable cible (Y) pour entraîner un modèle de régression. Le modèle apprendra à prédire ces trois variables simultanément en se basant exclusivement sur les trois métriques extraites de la segmentation LiDAR (X : densité, hauteur moyenne, et diamètre moyen des couronnes).

Prérequis sur les données d'inventaire : la méthodologie développée pour calculer la biomasse de référence nécessite des données terrain dont chaque arbre devrait avoir au minimum comme information :

- Des noms scientifiques (famille, espèce, genre) dans la mesure du possible ou sinon un nom vernaculaire : les données collectées à la Réunion avaient tous des noms scientifiques, contrairement à celles collectées en Guyane. La base



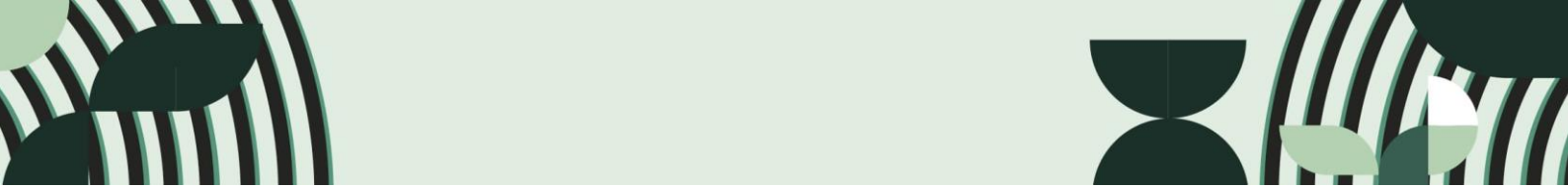
de données de Guyafor a dans ce cas aidé à compléter les noms scientifiques des données collectées.

- La densité du bois : si cette donnée n'est pas disponible, on peut utiliser la GWDD (**Global Wood Density Database**) à travers le package **R BIOMASS** pour retrouver les densités associées, mais pour cela, les données doivent absolument avoir des noms scientifiques.
- Le DBH (diamètre à hauteur de poitrine): on obtient généralement cette valeur grâce à la circonférence de l'arbre mesurée sur le terrain.

Notre approche repose sur un flux de travail en trois étapes successives :

1. **Calcul de la biomasse aérienne (AGB) de référence** à partir des données d'inventaires et des équations allométriques, afin de disposer d'une variable cible pour la calibration.
2. **Segmentation d'arbres/couronnes** pour extraire des métriques structurales issues du LiDAR et/ou de l'optique (variables explicatives).
3. **Modélisation statistique** (Random Forest) pour prédire simultanément AGB, densité et diamètre à l'échelle placette.

Pour l'estimation de la biomasse aérienne de référence (Y_{AGB}), la méthodologie repose sur une approche complémentaire intégrant le standard pantropical de [Chave et al.,2014](#) et les allométries locales récentes issues des travaux de [Huertas et al.,2025](#). Nous les utilisons en synergie pour réduire l'incertitude liée à la hauteur des arbres, qui constitue l'un des principaux biais des modèles globaux. Bien qu'il aurait été possible de calculer une biomasse de référence basée uniquement sur les inventaires terrain (DBH et taxonomie) en utilisant le paramètre de stress environnemental E ([Chave et al., 2014](#)), nous avons écarté cette approche classique. En effet, [Feldpausch et al.,2012](#) ont démontré que l'omission de la hauteur réelle engendre des biais systématiques majeurs dans les bilans carbones tropicaux. De plus, [Chave et al.,2014](#) concluent dans leur étude que l'inclusion de la variable H réduit l'erreur relative d'estimation de la biomasse de l'arbre de 27,7 % à 15,3 %. L'utilisation des modèles locaux de [Huertas et al.,2025](#), qui injectent des hauteurs calibrées par



LiDAR, permet donc de s'affranchir du lissage climatique des modèles pantropicaux et d'obtenir une vérité terrain (Y_{AGB}) qui se rapprocherait mieux de la structure réelle de la forêt tropicale.

Cette étape vise donc à produire une donnée de validation pour un modèle de régression qui prédira sur d'autres placettes la biomasse totale, et la structure de la forêt dans la placette (la densité de tige et le diamètre moyen).

Protocole : Nous avons appliqué le protocole de l'article [Huertas et al.,2025](#) sur les données tests premièrement (MNHC, inventaires terrains) et sur les données finales comme suit :

1. **Estimation des hauteurs** : dans les inventaires forestiers, la hauteur totale des arbres est rarement mesurée sur le terrain de manière exhaustive. Pour pallier ce manque, nous avons mobilisé les allométries locales développées par [Huertas et al.,2025](#) calibrées spécifiquement pour la région guyanaise grâce aux données LiDAR(ALS). La relation entre le diamètre (DBH) et la hauteur (H) est formalisée par une équation asymptotique de type Michaelis-Menten ([Huertas et al.,2025](#)). Cette forme mathématique est particulièrement adaptée aux peuplements tropicaux denses, car elle traduit le plafonnement de la croissance en hauteur lorsque l'arbre atteint la canopée, en tenant compte de la compétition locale. Nous avons appliqué le Modèle 4 de [Huertas et al.,2025](#), le plus complet, qui intègre deux sources de variation : l'identité taxonomique de l'espèce (paramètres α et β spécifiques à chaque espèce) et la hauteur de canopée locale (HC), définie comme la médiane du MNHC dans un rayon de 30 m autour de l'arbre. La formule est : $H = \alpha(\text{espèce}, HC) \times DBH / (\beta(\text{espèce}) + DBH)$.
2. **Traitement des palmiers** : l'architecture des palmiers nécessite un traitement distinct. Les palmiers ne présentent pas de croissance radiale secondaire (leur diamètre reste relativement constant au cours de leur vie). L'application de l'équation de Michaelis-Menten y est donc caduque. Pour ces espèces, la biomasse dépend quasi-exclusivement de la hauteur. Nous avons donc appliqué une allométrie spécifique aux palmiers amazoniens ([Goodman et al.,](#)

2013), où la biomasse est une fonction directe et souvent linéaire de la hauteur totale.

3. **Calcul de l'AGB** : application de l'équation pantropicale de référence [Chave et al., 2014](#), : $AGB = 0.0673 \times (WD \times H \times DBH^2)^{0.976}$, où WD est la densité de bois attribuée selon la taxonomie de chaque arbre, H la hauteur et DBH le diamètre à hauteur de poitrine.

Résultats :

- Production d'une base de données géoréférencée (GeoPackage + CSV) contenant les estimations de biomasse par placette (unité d'échantillonnage), avec le rapport de production associé.

Le calcul de la biomasse repose sur le protocole d'inventaire terrain, qui s'articule autour d'un échantillonnage stratifié en trois sous-placettes : une première sous placette circulaire de rayon 15m , une deuxième sous placette rectangulaire allant de 0 à 60 m de longueur sur 20 m de largeur, où tous les arbres de diamètre (DBH) > 10 cm ont été mesurés et géoréférencés ; et une troisième sous placette rectangulaire zone s'étendant de 60 à 250 m sur 20 m de largeur (le reste de la placette rectangulaire), où seuls les arbres de DBH > 40 cm ont été relevés.

Dans un premier temps, la biomasse a été calculée pour chaque arbre individuel. Cette valeur a ensuite été pondérée par un coefficient d'extrapolation spatiale (mise à l'hectare), inversement proportionnel à la surface de la sous-placette sur laquelle la classe de l'arbre était éligible à la mesure. Enfin, la biomasse totale et la densité de tiges de la placette ont été obtenues en sommant l'ensemble de ces valeurs extrapolées. Ci-dessous, une représentation de l'unité d'échantillonnage ayant servi de base à ces calculs :

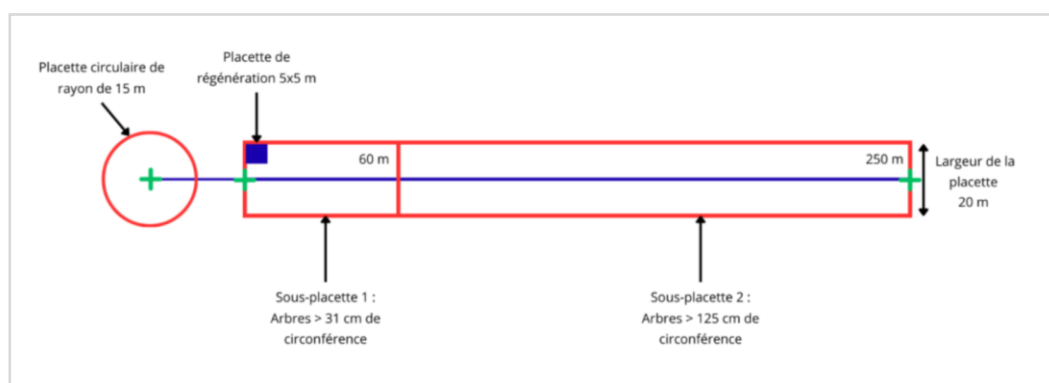


FIGURE 2 : SCHÉMA D'UNE UNITÉ D'ÉCHANTILLONNAGE

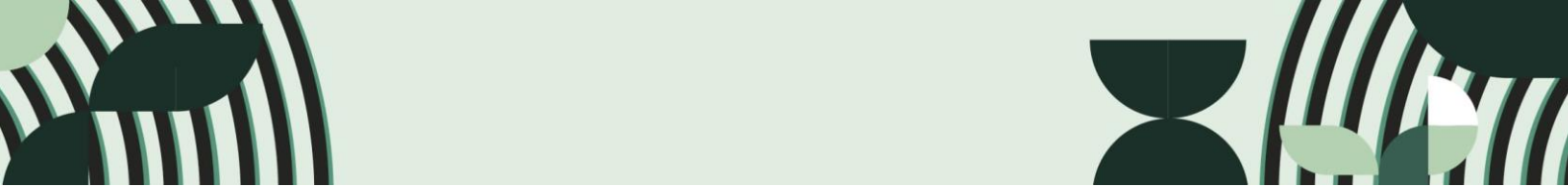
Les résultats de calcul de la biomasse, par UE sont les suivants :

Placettes	Y_Biomasse totale (t/ha)	Y_Nb_tige(ties/ha)	Surface Inventoriée (en ha)	Nb_arbre terrain
P2	443,80	744	0,57	169
P1	370,54	500	0,57	122
P4	434,43	744	0,57	168
P3	442,24	622	0,57	146

TABLEAU 1: BIOMASSE PAR UNITÉ D'ÉCHANTILLONNAGE D'APRÈS HUERTAS ET AL. (2025)

Il est à noter que les placettes P2 et P4 devraient théoriquement présenter des valeurs de biomasse très proches vue leur densité d'arbre identiques. Cette différence s'explique par la variabilité structurelle inhérente au peuplement (la taille individuelle des arbres différant d'une placette à l'autre, ce qui impacte directement l'AGB calculée). Environ 61 arbres ont été écartés pour l'estimation de la biomasse de référence, car nous n'avons pas pu établir leur correspondance avec un nom scientifique. En effet, pour le cas de la Guyane, les inventaires comportaient des noms vernaculaires (dont certains en dialectes locaux). Or, l'identification du taxon exact est indispensable pour leur attribuer une densité de bois via la base GWDD. Nous avons rencontré des cas particuliers où ces noms vernaculaires ne possédaient aucune correspondance scientifique (61 individus au total).

En l'absence du Canopy Constructor qui a guidé ce calcul de biomasse, il reste possible de recourir à des allométries pantropicales (Chave et al., 2014; Feldpausch et al., 2011). Cette option constitue une solution pragmatique à court terme, mais elle expose à des biais d'estimation potentiels, dépendant notamment de la composition

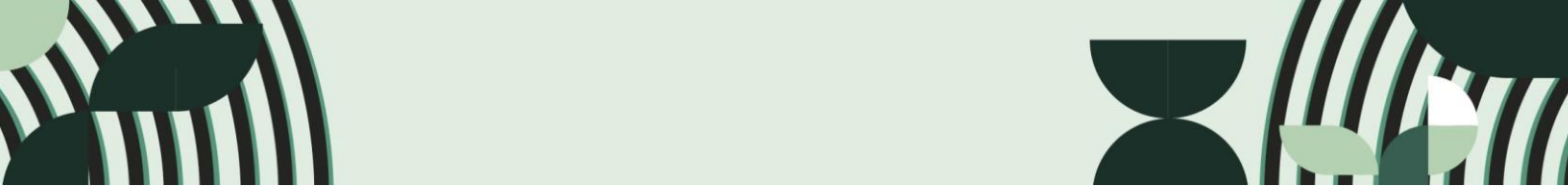


spécifique du peuplement. Si cette voie est retenue, ces biais devront être explicitement quantifiés et documentés dans les livrables, afin de préserver la traçabilité et l'interprétabilité des résultats.

Afin d'évaluer la robustesse de la biomasse calculée (Y_{AGB}) face aux incertitudes de modélisation, une approche de propagation des erreurs par méthode bayésienne a été mise en œuvre. Nous avons exploité la distribution postérieure des paramètres allométriques a et b , issue de l'ajustement bayésien réalisé avec le package **brms** de **R**. Pour chaque arbre, 1000 combinaisons de paramètres (a , b) ont été tirées aléatoirement depuis cette distribution postérieure. Chaque tirage produit une hauteur H différente via le modèle de Michaelis-Menten, qui est ensuite propagée dans l'équation de biomasse de [Chave et al., 2014](#), générant ainsi une distribution de 1000 valeurs d'AGB par arbre. Ces distributions sont sommées à l'échelle de la placette pour produire une distribution d'AGB totale, dont on extrait la médiane et l'intervalle de crédibilité à 95 %. Cette approche propage l'incertitude paramétrique c'est-à-dire l'incertitude sur les coefficients allométriques eux-mêmes et non uniquement l'erreur résiduelle autour de la hauteur prédite, ce qui en fait une estimation plus conservative et plus rigoureuse de l'incertitude.

Il convient néanmoins de souligner que cette propagation reste partielle : elle ne couvre que l'une des quatre sources d'erreurs identifiées dans la chaîne de calcul. L'incertitude sur la densité du bois (w_d), assignée par correspondance taxonomique avec la base de données mondiale GWDD, n'est pas propagée. L'incertitude sur la mesure terrain du DBH (erreurs de lecture, position de mesure sur les contreforts) n'est pas non plus prise en compte. L'incertitude sur la hauteur de canopée locale (LCH), extraite du MNHC à une résolution de 1 m et sensible à la qualité du modèle numérique de terrain, n'est pas propagée. Enfin, l'incertitude inhérente à la formule de [Chave et al., 2014](#) elle-même dont l'erreur résiduelle standard est de l'ordre de 0,357 en log n'est pas incluse. En conséquence, les intervalles de crédibilité obtenus ($CV < 1\%$ à l'échelle de la placette) constituent une borne inférieure de l'incertitude réelle, et non une quantification exhaustive de celle-ci.

L'application de la méthodologie décrite dans l'article de [Huertas et al., 2025](#) requiert



en amont la calibration d'un modèle forestier individu-centré appelé Canopy Constructor. Pour Paracou, ce modèle avait déjà été calibré par les équipes de recherche, ce qui a permis son utilisation directe. En revanche, son déploiement sur un nouveau site (comme La Réunion) nécessite une calibration spécifique, laquelle est encore en phase de recherche et validation.

D.2 SEGMENTATION INDIVIDUELLES

Objectif : la segmentation a pour but d'isoler des arbres individuels ou leurs couronnes afin d'extraire des métriques structurelles (surface/diamètre de couronne, hauteur, etc.) qui serviront de variables explicatives du modèle de régression. En forêt tropicale, la densité et l'entremêlement des canopées rendent cet exercice particulièrement difficile.

Contexte et État de l'art

La forêt tropicale constitue un environnement particulièrement complexe pour la télédétection. Elle se caractérise par une forte densité de tiges, la présence de lianes, des sous-bois denses et des canopées stratifiées qui s'entremêlent, rendant la délimitation des couronnes individuelles ardue. L'objectif de la segmentation ici est d'isoler les arbres individuels pour calculer des métriques structurelles (surface de couronne, hauteur...) qui serviront de variables explicatives (X) dans le modèle de régression final.

Avant d'arrêter notre choix méthodologique, une revue bibliographique a été menée sur la segmentation 3D et 2D en milieu tropical, distinguant deux grandes familles d'approches :

- **Les méthodes géométriques et statistiques** : Elles opèrent soit sur le Modèle Numérique de Canopée (MNC/MNHC) et le nuage de point, soit directement sur le nuage de points 3D. Elles ne nécessitent pas de données annotées et sont rapidement déployables.
- **Les méthodes par Deep Learning (État de l'art)** : Ces architectures combinent l'extraction de caractéristiques profondes avec des algorithmes de clustering avancés pour une segmentation plus précise.

- Modèles étudiés : PointGroup ([Jiang et al., 2020](#)), MinkowskiEngine ([Xiang et al., 2023](#)), ainsi que des modèles récents agnostiques aux capteurs comme ITS-Net ([Li et al., 2026](#)) et SegmentAnyTree ([Wielgosz et al., 2024](#)).
- Limite principale pour ce projet : ces méthodes nécessitent généralement des milliers d'arbres annotés manuellement pour l'entraînement, représentant un coût et un délai incompatible avec les contraintes du projet.

Protocole mis en œuvre :

Compte tenu de l'analyse bibliographique, nous avons opté pour des méthodes assez fiables et rapides à déployer :

- **Approche 2D (Optique)** : segmentation de couronnes effectuées sur des orthophotos avec des modèles de vision par ordinateur. Nous avons testé :
 - **Detectree2** : qui a un bon taux de détection global des houppiers.
 - **SAM3 (Segment Anything Model, 2025)** : précision de délimitation des contours de couronnes supérieure à celle de Detectree2, mais taux de détection plus faible.
 - Limite : ces deux méthodes sont aveugles du sous-bois, ne détectant que la canopée dominante.
- **Approche 3D-2D,5 – Segmentation sur nuage de points LiDAR** : Application de méthodes non supervisées :
 - **Li et al., 2012** : segmentation dans le nuage de points. Choisie car elle ne nécessite pas de base d'entraînement, offrant un gain de temps considérable.
 - **Silva et al., 2016** : segmentation basée sur le Modèle Numérique de Canopée (MNHC).

- **Test complémentaire** : lidaRtRee (INRAE), écarté à ce stade car principalement optimisé pour des structures tempérées et nécessitant des ajustements pour le tropical.

Résultats :

- Production d'un fichier CSV contenant des identifiants d'arbres segmentés et leurs métriques structurelles associées (la hauteur, diamètre de couronne...).

Nous avons décidé d'utiliser principalement les résultats issus de la segmentation géométrique du nuage de points, car cette méthode permet d'identifier un plus grand nombre d'arbres que la segmentation optique. Par ailleurs, l'approche optique reste en cours d'amélioration : nous avons combiné les deux méthodes de détection de couronnes (SAM3 et Detectree2) afin d'optimiser la précision et d'augmenter le nombre de houppiers segmentés.

Ci-dessous quelques résultats de la segmentation des nuages de points :

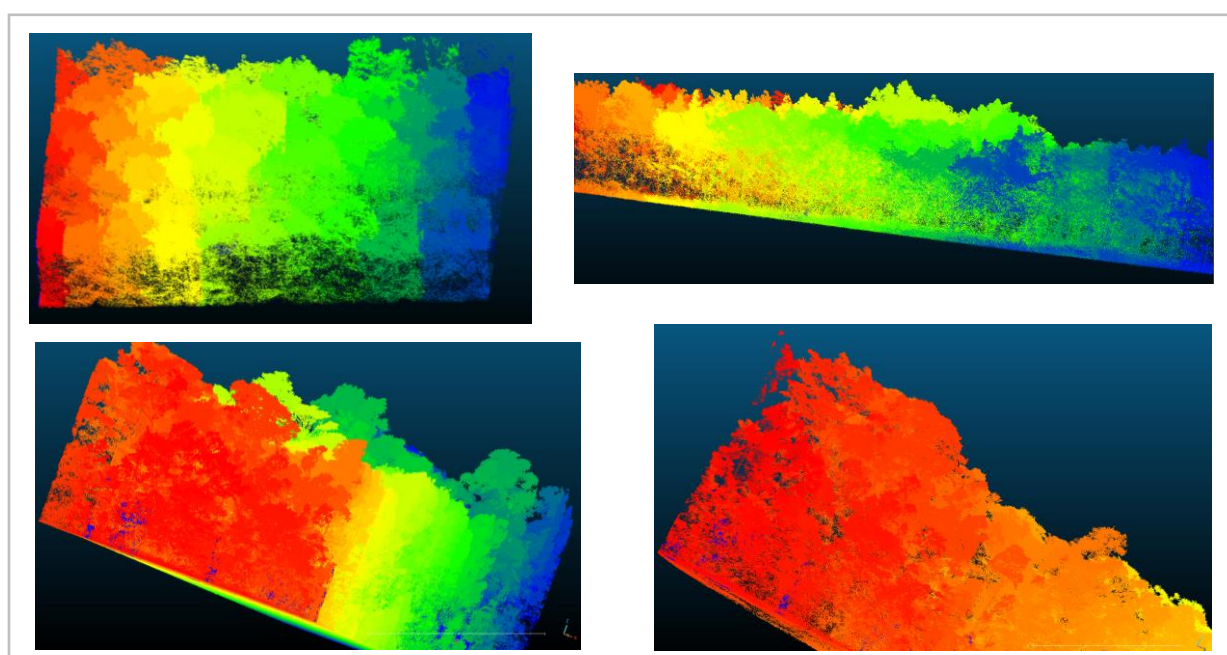
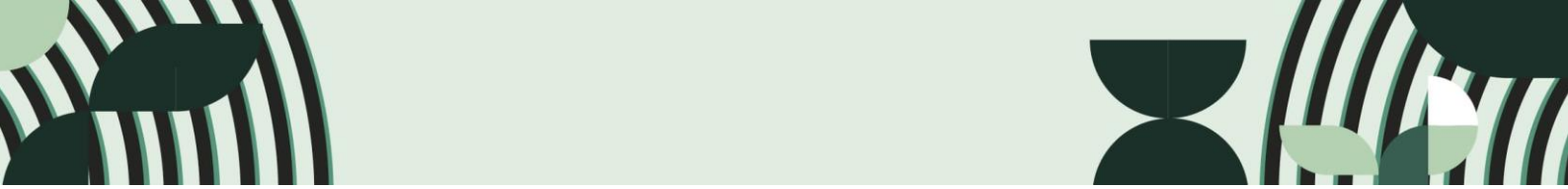


FIGURE 3 : RESULTATS DE SEGMENTATION INDIVIDUELLE D'ARBRES SUR NUAGE DE POINTS LIDAR AÉRIEN (DONNEES GEOFIT)

Chaque couleur représente un arbre segmenté individuellement par l'algorithme géométrique. Les deux premières vues illustrent la segmentation en vue de dessus (projection 2D de la canopée) et de profil, permettant d'apprécier la densité de détection des houppiers. Les deux vues suivantes montrent une coupe oblique du



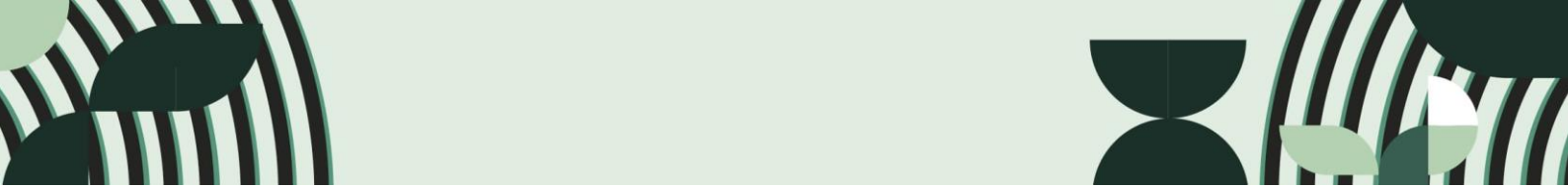
nuage de points segmenté, mettant en évidence la stratification verticale de la forêt et la capacité de l'algorithme à isoler des individus dans les strates dominantes. Cette densité de points (~ 1300 pts/m²) permet de détourner efficacement les couronnes de la strate supérieure, et de capter partiellement les arbres du sous-étage.

Les méthodes géométriques retenues ici ont été choisies pour leur rapidité de déploiement dans le cadre des contraintes du projet. D'autres algorithmes de segmentation sur nuage de points mériteraient d'être testés dans une phase ultérieure, notamment **ComputeTree**, un outil développé par l'ONF spécifiquement pour la reconstruction 3D d'arbres individuels à partir de nuages de points haute densité. Sa capacité à modéliser la structure complète de l'arbre, y compris les branches et le tronc, pourrait permettre une meilleure pénétration des strates inférieures et une réduction des erreurs de sous-segmentation identifiées dans cette étude. Son évaluation constitue une perspective pour les prochaines phases du projet.

Si le Deep Learning offre des perspectives prometteuses, il nécessite cependant un volume important d'arbres annotés manuellement pour l'entraînement, ce qui n'est pas compatible avec les contraintes de ce projet. Cependant, des alternatives existent pour avoir un jeu de données d'entraînement : inférer un modèle sur un nuage de point et récupérer les arbres segmentés (y faire un nettoyage) puis ensuite entraîner un autre modèle. Cette approche n'a pas pu être testée dans ce protocole, mais fera l'objet de tests futurs. L'approche géométrique proposée par [Li et al., 2012](#) et [Silva et al., 2016](#), qui fonctionne sans supervision, associée à une calibration statistique via le modèle Random Forest, constitue ainsi l'un des compromis entre efficacité et maîtrise des coûts.

D.3 MODELISATION STATISTIQUE (REGRESSION MULTIVARIEE)

La troisième étape du pipeline consiste à entraîner un modèle de régression (Random Forest) pour prédire la biomasse et la structure forestière à partir des métriques LiDAR extraites en D.2. Nous décrivons ici l'architecture complète de cette étape, c'est à dire le choix de l'algorithme, la structure du jeu de données, l'analyse de cohérence ; car le pipeline est techniquement fonctionnel et prêt à être déployé. Cependant, l'entraînement effectif du modèle n'a pas pu être réalisé à ce stade : le nombre de



placettes disponibles (4 en Guyane) est insuffisant pour produire un modèle généralisable. Cette section documente donc à la fois ce qui a été construit et ce qui reste à lever pour franchir cette dernière étape.

Objectif : Fusionner (i) la biomasse de référence (étape 1) et (ii) les métriques de segmentation (étape 2) pour calibrer un modèle capable de prédire simultanément à l'échelle de la placette : AGB, densité de tiges, diamètre quadratique moyen.

Protocole :

L'algorithme de Random Forest (RF) a été sélectionné pour sa capacité à traiter des variables d'entrée hétérogènes, il nécessite un volume important de données d'entraînement pour éviter le phénomène de surapprentissage (overfitting : le modèle apprend par cœur les données d'entraînement au lieu de généraliser) et stabiliser la variance du modèle.

En Guyane française, la constitution de ce jeu de données (dataset) se heurte à des contraintes logistiques fortes : l'inaccessibilité des sites forestiers et les coûts associés aux acquisitions LiDAR aéroportées couplées aux inventaires dendrométriques limitent les zones de recouvrement. Dans le cadre de cette étude, seules 4 placettes répondaient aux exigences de qualité requises. Dans ces conditions, l'entraînement et l'évaluation statistique d'un modèle de régression s'avèrent prématurés et produiraient des métriques de performance (R^2) biaisées.

Un autre biais potentiel identifié a été la dispersion géographique des placettes : comme l'illustre la figure suivante, les placettes sélectionnées sont regroupées par paires rapprochées mais avec un écart géographique notable entre les deux paires, pouvant induire des différences de composition floristique et de structure forestière défavorables à la généralisation du modèle.



Figure 4 : DSIPERSION DES PLACETTES (DONNEES COLLECTEES PAR ONFI et GEOFIT)

Analyse de la cohérence physique du jeu de données

Le tableau ci-dessous présente le jeu de données d'apprentissage généré par notre chaîne de traitement pour les 4 placettes (variables ramenées à l'hectare et inventaire filtré pour DBH > 10 cm) :

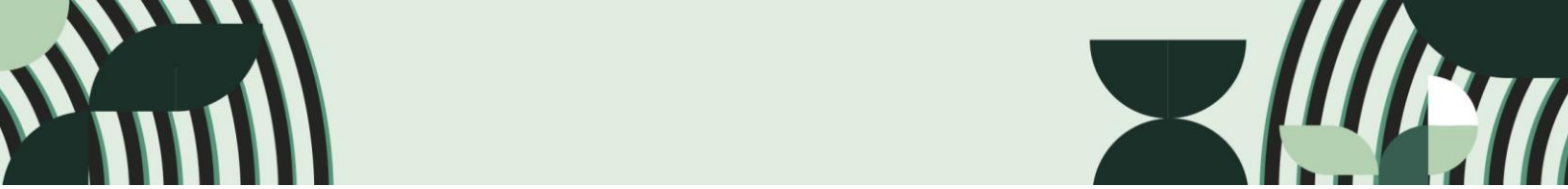
Placettes	Y_Biomasse totale (t/ha)	Y_DBH (cm)	Y_Nb_tige(ties/ha)	X_H_moy_segmentation	X_D_moy_couronne	X_Nb_tige_segmentation
P2	443,80	30,74	744	30,55	10,42	102
P1	370,54	36,85	499	31,52	10,70	96
P4	434,43	29,10	744	34,94	10,55	98,7
P3	442,24	33,69	621	35,13	10,90	92,7

Tableau 2 : Entrées du modèle de régression

L'analyse croisée des variables prédictives LiDAR (X) et des variables cibles terrain (Y) s'aligne avec la précision du pipeline de préparation des données :

a. Détection de la strate dominante (N)

L'algorithme de segmentation LiDAR extrait en moyenne environ 97 houppiers sur le nuage de point (ALS) couvrant toute l'UE (X_Nb_tiges_segmentation). Or, l'inventaire terrain démontre que la forêt compte réellement entre 500 et 750 tiges à l'hectare (Y_Densité extrapolée). Cet écart confirme visuellement et mathématiquement que la segmentation géométrique actuelle ne détecte que le toit de la canopée (les arbres



dominants et co-dominants) et reste majoritairement aveugle au sous-étage dominé, même avec une densité LiDAR de 1300 pts/m².

b. Relation allométrique de la strate dominante (Dg)

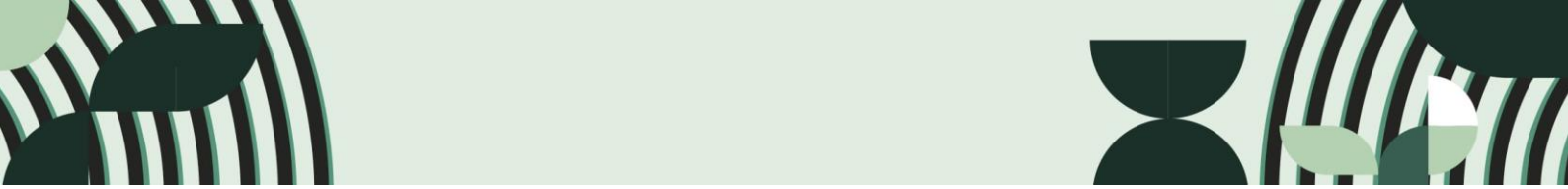
Le diamètre moyen des couronnes segmentées ($X_D_moy_cour \sim 10,5$) et le diamètre quadratique moyen au sol ($Y_Dg \sim 30-35$ cm) témoignent d'un peuplement composé d'individus matures et dominants. Cette relation confirme que les objets volumétriques segmentés par l'algorithme correspondent bien aux houppiers de la strate supérieure mesurée in situ.

c. Interprétation de la biomasse de référence (Y_AGB)

Les valeurs de biomasse (370 à 444 t/ha) reflètent un biais structurel pour l'apprentissage du modèle Random Forest : notre variable cible (Y_AGB) représente bien la forêt entière, mais nos variables explicatives LiDAR (X) ne décrivent physiquement que la strate dominante. Le modèle devra donc apprendre à déduire une biomasse totale de 400 t/ha en ne regardant que les ~100 arbres (moyenne du nombre de tiges segmentés dans le nuage de point) du toit de la canopée. Cela renforce notre recommandation de tester d'autres méthodes de segmentation, plus pénétrantes, pour capturer une plus grande proportion de la canopée du sous-étage et réduire cet écart de perception.

De manière générale, le pipeline de traitement automatisé est fonctionnel techniquement : il extrait des descripteurs LiDAR physiquement cohérents et formatés pour l'apprentissage automatique. Cependant, l'analyse critique de ce premier jeu de données met en exergue l'importance capitale de l'échantillonnage terrain. Pour calibrer un modèle prédictif final capable d'estimer la biomasse totale, l'architecture logicielle développée devra être alimentée par de futures placettes d'étalonnage, permettant d'évaluer la détection LiDAR sur l'ensemble du profil vertical de la forêt.

L'approche par Machine Learning (Random Forest) nécessite une taille d'échantillon suffisante pour stabiliser la variance des arbres de décision et éviter le surapprentissage (overfitting). Selon la littérature spécialisée dans l'estimation de biomasse par télédétection ([Zolkos et al., 2013](#)), les modèles prédictifs à l'échelle du



peuplement nécessitent un minimum de 40 à 50 placettes d'inventaire. De plus, en milieu tropical complexe, [Ploton et al., 2020](#) soulignent que les algorithmes non paramétriques requièrent une forte représentativité spatiale (souvent > 60 placettes) pour généraliser correctement les prédictions.

E ANALYSE DE FAISABILITE DU DEPLOIEMENT A LA REUNION : VEROUS IDENTIFIES ET PERSPECTIVES

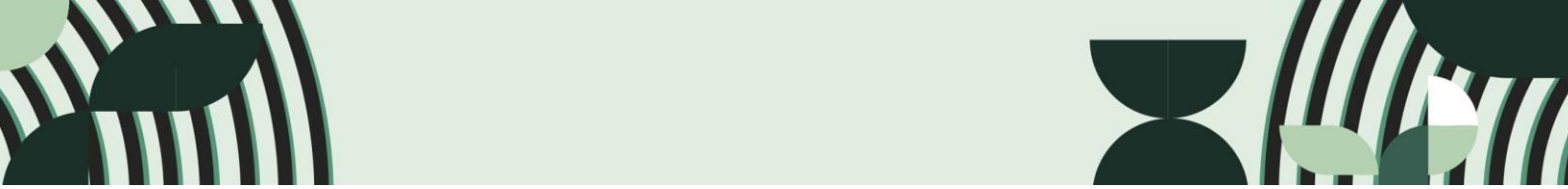
Contrairement à la Guyane où le pipeline a pu être appliqué sur des données réelles, La Réunion constitue dans ce rapport un cas d'analyse de faisabilité. L'objectif de cette section n'est pas de présenter des résultats de calcul de biomasse, mais d'identifier précisément les verrous algorithmiques et statistiques qui conditionnent le transfert de la méthodologie, et de proposer des alternatives ou des prérequis au protocole de la Guyane applicable à La Réunion.

E.1 LE VEROU TECHNOLOGIQUE : LA DÉPENDANCE AU CANOPY CONSTRUCTOR

Le protocole appliqué en Guyane bénéficie de travaux de recherche très avancés menés sur la station de Paracou, notamment l'utilisation du Canopy Constructor ([Huertas et al., 2025](#)) pour définir la biomasse de référence.

Le fonctionnement du modèle : le Canopy Constructor est un modèle forestier centré sur l'individu (Individual-Based Model). Son objectif est de générer une forêt virtuelle 3D qui correspond le plus fidèlement possible à la canopée réelle observée par le LiDAR. Via une procédure d'optimisation itérative, le logiciel ajuste les dimensions et les positions des arbres pour minimiser la différence entre le Modèle Numérique de Canopée (MNHC) simulé et le MNHC réel. Pour fonctionner, il nécessite :

- Des **inventaires au sol** (diamètres des arbres) pour générer le peuplement virtuel initial ;
- Des **allométries locales** (relations diamètre-hauteur et diamètre-couronne) issues de mesures terrain pour contraindre l'optimisation.



Cette optimisation s'appuie sur le modèle asymptotique de Michaelis-Menten pour la relation hauteur-diamètre : $H = (\alpha * DBH) / (\beta + DBH)$. Dans ce modèle, le paramètre alpha représente la hauteur maximale asymptotique, et beta le diamètre pour lequel l'arbre atteint la moitié de cette hauteur maximale.

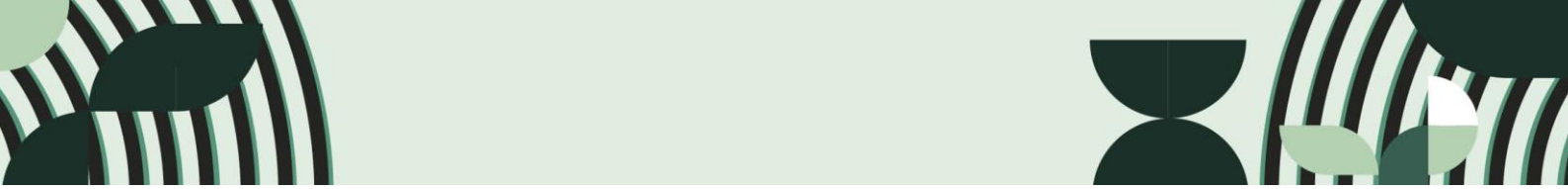
Performances et contraintes :

- **Gains** : le modèle génère des estimations de biomasse cohérentes, souvent supérieures de 11 à 13 % aux modèles pantropicaux classiques ([Huertas et al., 2025](#)).
- **Limites pour la transposabilité** : ce logiciel de recherche n'est pas une solution clé en main. Pour contraindre l'optimisation des paramètres alpha et beta, il nécessite impérativement une allométrie de base issue de mesures terrain rigoureuses. Bien qu'une utilisation sans données sol soit théoriquement possible dans l'architecture du code, cette approche à l'aveugle n'est pas validée par la communauté scientifique et est considérée comme risquée. De plus, à l'échelle d'un large paysage, le gain de précision apporté sur la biomasse totale par rapport à des méthodes surfaciques plus simples ne compense pas nécessairement le temps de paramétrage, de calcul et de débogage exigé par l'outil.

Il faudrait engager un échange avec les équipes ONF, CIRAD-IRD/AMAP pour préciser le calendrier de disponibilité du Canopy Constructor calibré pour La Réunion (prendre connaissance des allométries locales disponibles), et décider d'un traitement intermédiaire avec allométries pantropicales si le délai est incompatible avec le calendrier du projet.

E.2 L'ALTERNATIVE A COURT TERME POUR LA REUNION

L'alternative la plus directe consiste à utiliser des allométries régionales ou pantropicales, telles que celles de [Feldpausch et al., 2011](#) pour estimer la hauteur, couplées à l'équation de biomasse de [Chave et al., 2014](#). Ces modèles utilisent le diamètre mesuré au sol (DBH), la densité du bois des espèces et un paramètre de stress environnemental local (E).



Biais attendus : il convient néanmoins de souligner que cette méthode expose l'estimation à des biais structurels. En s'affranchissant de la hauteur réelle mesurée par le LiDAR (MNCH), l'utilisation d'équations pantropicales a tendance à lisser la complexité de la canopée et peine à capturer la variabilité architecturale induite par la compétition lumineuse, la topographie ou les spécificités des espèces endémiques de l'île.

E.3 L'OBSTACLE COMMUN DE L'ECHANTILLONNAGE (GUYANE ET REUNION)

Indépendamment du calcul de la biomasse de référence, La Réunion souffre de la même contrainte opérationnelle que la Guyane : la couverture LiDAR aéroportée (Geofit) ne se superpose qu'à une fraction des placettes d'inventaire terrain. Entraîner un modèle de Machine Learning (Random Forest) sur un si faible nombre d'échantillons appariés produirait inévitablement un algorithme biaisé et incapable de généraliser ses prédictions (surapprentissage).

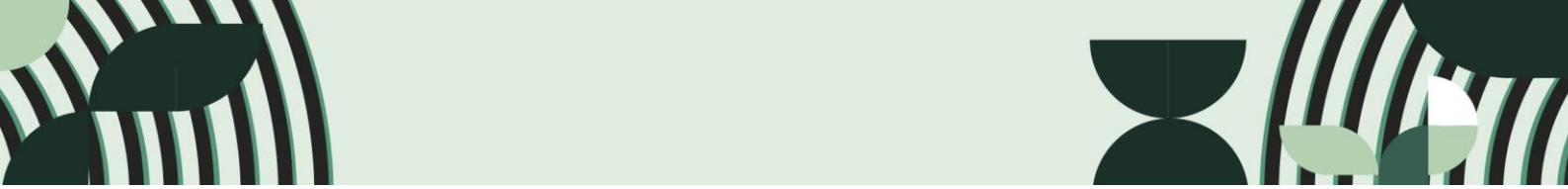
F SYNTHÈSE DES OUTILS ÉVALUÉS ET RECOMMANDATIONS

F.1 RECOMMANDATIONS GLOBALES POUR UN PASSAGE À L'ÉCHELLE OPÉRATIONNELLE

À l'issue de cette preuve de concept, le pipeline d'extraction des métriques LiDAR (X) et de préparation des données est validé informatiquement. En revanche, la transformation de ce socle technologique en un outil de prédiction de biomasse déployable à grande échelle (en Guyane comme à La Réunion) nécessite de lever des verrous statistiques et méthodologiques identifiés au cours de l'étude.

(1) Réviser la stratégie d'acquisition – le prérequis statistique

Les modèles non paramétriques (Machine Learning) exigent une forte représentativité spatiale. La littérature montre qu'un modèle prédictif de biomasse à l'échelle du peuplement nécessite une base d'entraînement d'au moins 40 à 50 placettes d'inventaire (Zolkos et al., 2013). En milieu forestier tropical complexe, ce besoin est



encore plus contraignant, pour une raison méthodologique souvent sous-estimée : l'autocorrélation spatiale de l'AGB due à la variabilité des peuplements dans les placettes.

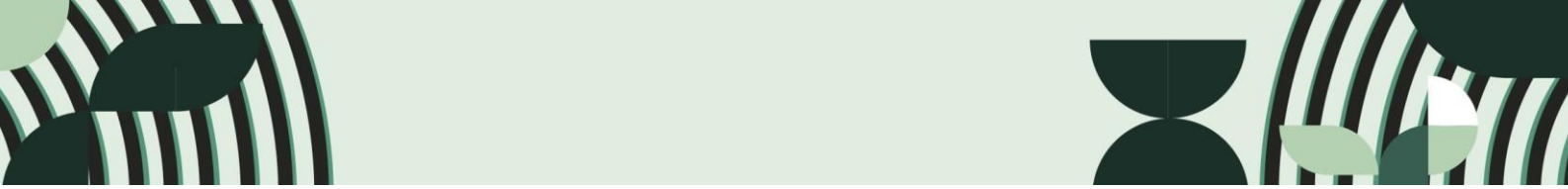
[Ploton et al., 2020](#) ont en effet démontré, sur un jeu de données massif de 11,8 millions d'arbres en Afrique centrale, qu'un modèle Random Forest validé par une méthode classique aléatoire (random K-fold CV: cross validation) affichait un R^2 apparent de 0,53, suggérant une bonne capacité prédictive. Or, dès lors qu'une validation spatiale rigoureuse était appliquée, en s'assurant que les points de test étaient géographiquement indépendants des points d'entraînement, ce R^2 s'effondrait à 0,14, puis devenait quasi nul au-delà de 100 km de distance. Ce résultat fondamental illustre que les performances statistiques habituellement reportées pour ce type de modèle sont trompeuses : le modèle ne prédit pas réellement la biomasse, il exploite la simple proximité géographique entre les points d'entraînement et de test.

Cette démonstration impose deux exigences concrètes pour la constitution du jeu de données d'entraînement de notre pipeline. D'une part, une taille minimale de placette de 0,5 ha est recommandée, et idéalement de 1 ha, afin d'assurer une estimation de l'AGB de référence suffisamment représentative à l'échelle du pixel ([Ploton et al., 2020](#)). D'autre part, les placettes devront être distantes ou avoir des peuplements assez différents les unes des autres pour garantir l'indépendance spatiale entre les données d'entraînement et de validation.

Les futures campagnes de vol LiDAR devront donc être planifiées en visant à la fois un nombre suffisant de placettes (> 60) et une dispersion géographique maximale sur l'ensemble du territoire, couvrant une gamme représentative de contextes forestiers guyanais (gradient de composition floristique, de topographie et de régimes de perturbation).

(2) Encadrer l'alternative allométrique si le Canopy Constructor n'est pas disponible

En l'absence du Canopy Constructor, il reste possible de recourir à des allométries pantropicales ([Chave et al., 2014](#) ; [Feldpausch et al., 2011](#)). Cette option constitue une



solution pragmatique à court terme, mais elle expose à des biais d'estimation potentiels, dépendant notamment de la composition spécifique du peuplement. Si cette voie est retenue, ces biais devront être explicitement quantifiés et documentés dans les livrables, afin de préserver la traçabilité et l'interprétabilité des résultats

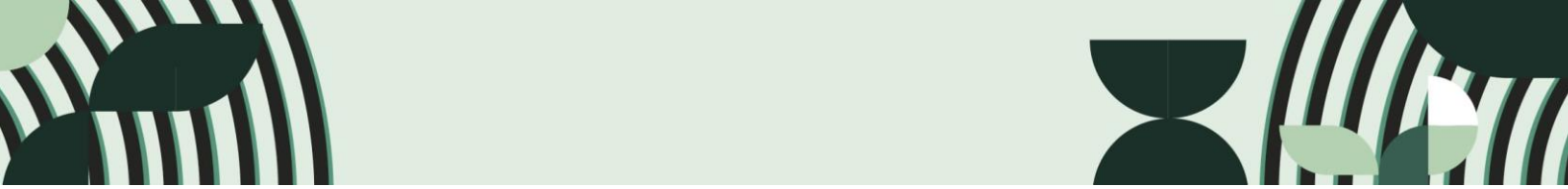
F.2 SYNTHÈSE DES OUTILS ÉVALUÉS ET RECOMMANDATIONS

Cette phase exploratoire a permis d'évaluer des outils et méthodes répartis en trois catégories : (i) segmentation LiDAR, (ii) segmentation optique 2D, et (iii) calcul de biomasse et modélisation. Pour chaque famille, un statut est attribué selon trois niveaux : opérationnel (déploable dès maintenant dans les conditions du projet), fort potentiel (résultats prometteurs mais nécessitant ajustements ou données complémentaires), et écarté à ce stade (incompatible avec les contraintes actuelles).

(A) Segmentation LiDAR — méthodes géométriques (fort potentiel, non encore opérationnel) :

Les algorithmes de segmentation géométrique sur nuage de points ([Li et al., 2012](#)) et sur Modèle Numérique de Canopée ([Silva et al., 2016](#)) présentent un fort potentiel mais ne peuvent pas encore être considérés comme pleinement opérationnels dans l'état actuel des données. Ces méthodes ne requièrent aucune donnée d'entraînement et sont rapides à déployer ; elles ont produit des résultats partiels par rapport à l'inventaire terrain extrapolé ~97 houppiers détectés par placette de nuage de point LiDAR. Cette valeur est très inférieure aux 400 tiges/ha (ou environ 150 arbres terrains par placette) caractéristiques d'une forêt guyanaise fermée, ce qui soulève une limite fondamentale. À ce titre, **ComputeTree** (ONF) constitue une piste à évaluer dans une phase suivante, en raison de sa conception spécifique pour la reconstruction individuelle d'arbres en 3D sur nuages de points haute densité.

La méthode **lidaRtRee** développée par l'INRAE a également été testée et est écartée à ce stade : robuste en forêt tempérée, elle nécessite des ajustements importants pour



s'adapter à la complexité structurelle de la forêt tropicale guyanaise et ne constitue pas une priorité dans l'état actuel du projet.

(B) Segmentation optique 2D (fort potentiel, limite structurelle) :

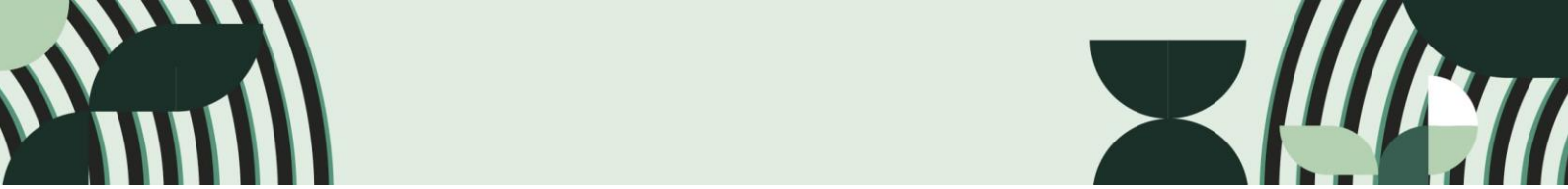
Les outils testés sur orthophotos, Detectree2 et SAM3 (Segment Anything Model, 2025), présentent chacun des atouts complémentaires. Detectree2 offre un bon taux de détection global des houppiers, tandis que SAM3 se distingue par une précision supérieure dans la délimitation des contours de couronnes, au prix d'un taux de détection plus faible. Leur combinaison est en cours d'optimisation et laisse entrevoir un fort potentiel pour améliorer la qualité de la segmentation 2D. Ces deux méthodes restent cependant aveugles au sous-bois et ne détectent que la canopée dominante, ce qui constitue une limite structurelle inhérente à l'approche optique.

(C) Deep Learning 3D (écarté à ce stade – contrainte de données annotées) :

Les architectures de Deep Learning appliquées aux nuages de points 3D (PointGroup, ITS-Net, SegmentAnyTree, MinkowskiEngine) représentent l'état de l'art en matière de segmentation individuelle d'arbres et présentent un potentiel scientifique, notamment les modèles comme ITS-Net (Li et al., 2026) et SegmentAnyTree (Wielgosz et al., 2024). Ces approches sont néanmoins écartées à ce stade : elles requièrent des milliers d'arbres annotés manuellement pour l'entraînement, ce qui représente un coût humain et un délai incompatible avec les contraintes du présent projet. Elles constituent en revanche une piste prioritaire à explorer dans le cadre d'un projet dédié disposant des ressources nécessaires à la constitution d'une base d'annotation.

(D) Calcul de la biomasse de référence (opérationnel en Guyane, non transposable en l'état à La Réunion) :

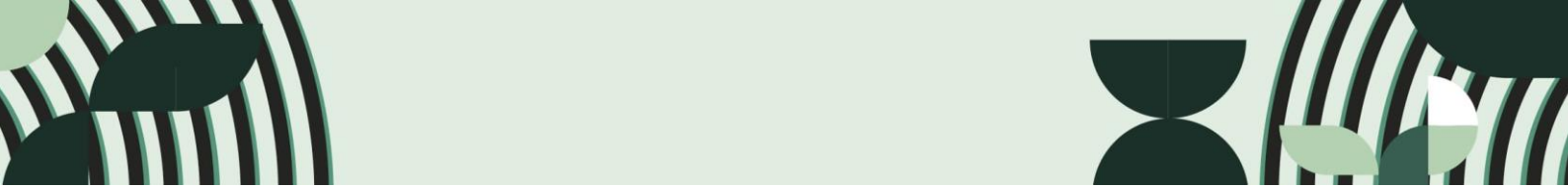
Le protocole hybride combinant les allométries locales de Huertas et al., 2025, calibrées par LiDAR sur la station de Paracou, et l'équation pantropicale de Chave et al., 2014 est jugé opérationnel pour la Guyane. Il permet de réduire l'erreur relative d'estimation de la biomasse de 27,7 % à 15,3 % par rapport à une approche sans mesure de hauteur (Chave et al., 2014), et la propagation des incertitudes par simulations de bayésienne (tirages MCMC) confirme la robustesse de la variable cible



ainsi produite ($\pm 0,7$ à $\pm 1,1$ t/ha à l'échelle de la placette). Ce protocole est en revanche non transposable à La Réunion en l'état, en raison de la dépendance au modèle Canopy Constructor, calibré uniquement pour Paracou. Pour La Réunion, le recours aux allométries pantropicales de [Chave et al., 2014](#) ou de [Feldpausch et al. \(2011, 2012\)](#) constitue une alternative à fort potentiel, sous réserve d'accepter des biais d'estimation potentiels << l'omission de la hauteur réelle engendre des biais systématiques majeurs dans les bilans carbone (Feldpausch et al., 2012) , et maintient une erreur relative d'estimation de la biomasse de l'arbre à 27,7 % (contre 15,3 % lorsque la variable de hauteur est incluse selon [Chave et al., 2014](#)). D'autre part, des travaux récents ont révélé que l'application de ces allométries globales sous-estime la biomasse locale de 11 à 13 % par rapport à un modèle calibré par LiDAR ([Huertas et al., 2025](#))>> selon la composition spécifique du peuplement.

(E) Modélisation par Random Forest (fort potentiel, non déclenchable à ce stade) :

L'algorithme de Random Forest, retenu pour la modélisation prédictive de l'AGB, de la densité de tiges et du diamètre quadratique moyen, présente un fort potentiel à deux conditions non encore réunies. D'une part, le pipeline de préparation des données est entièrement fonctionnel et les descripteurs produits sont physiquement cohérents (l'architecture logicielle est prête à monter en charge). D'autre part, deux verrous indépendants mais cumulatifs doivent être levés avant tout entraînement. Le premier est statistique : avec seulement 4 placettes disponibles répondant aux critères de qualité requis, le modèle produirait inévitablement un surapprentissage incapable de généraliser ses prédictions à d'autres placettes, alors que la littérature indique un minimum de 40 à 50 placettes pour stabiliser un modèle prédictif de biomasse à l'échelle du peuplement ([Zolkos et al., 2013](#)), voire plus de 60 en milieu tropical complexe ([Ploton et al., 2020](#)). Le second verrou est structural : les variables explicatives (X) alimentant le modèle sont issues de la segmentation du nuage de points, laquelle ne capture actuellement qu'environ 97 tiges/ha sur les 400+ que compte une forêt guyanaise fermée. En conséquence, un modèle entraîné sur ces données ne prédirait que la biomasse de la strate dominante visible, et non la biomasse totale de la forêt. La fiabilité du Random Forest est donc directement



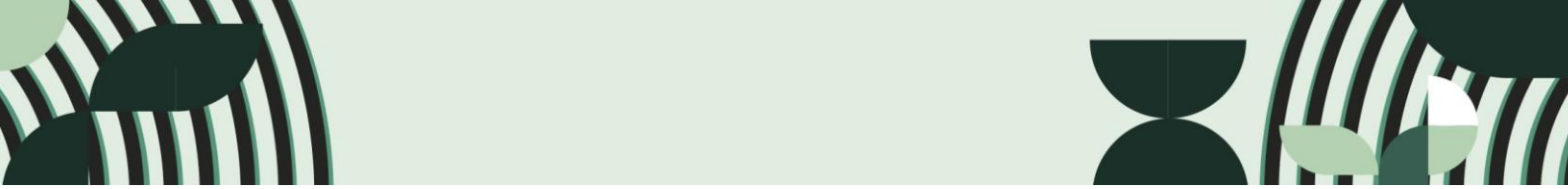
conditionnée à la résolution préalable des limites de la segmentation.

F.3 CONCLUSION DE SYNTHÈSE

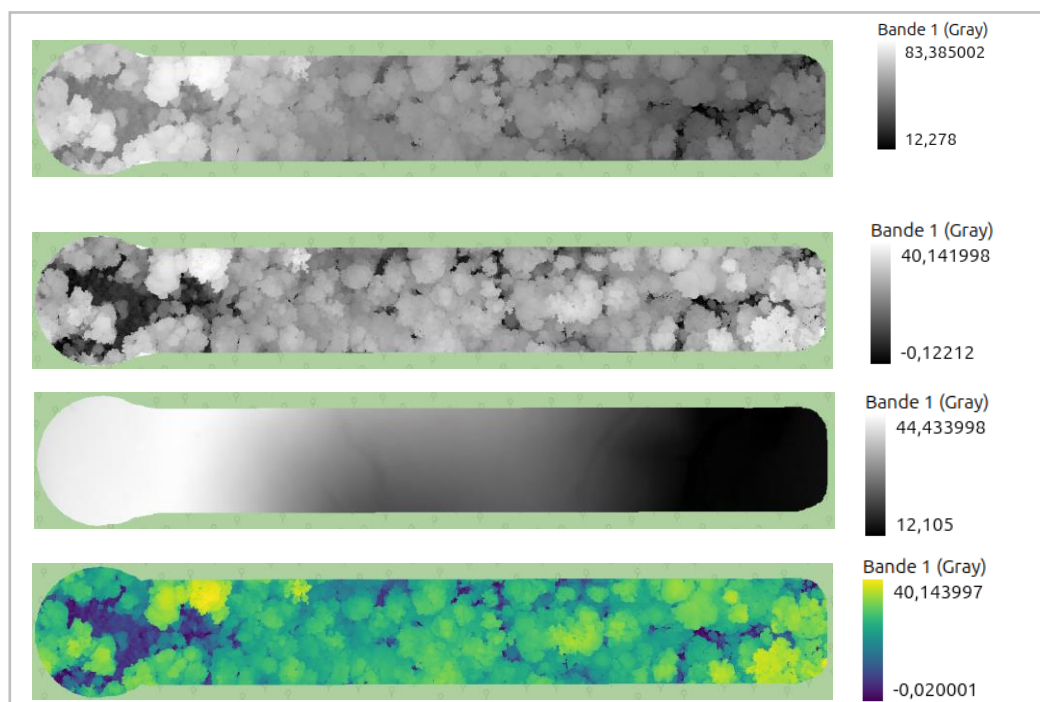
Cette phase exploratoire a permis de valider un pipeline de traitement complet et de sécuriser les choix méthodologiques. Les leviers d'amélioration identifiés — augmentation du nombre de placettes, inventaires exhaustifs et évaluation d'outils de segmentation plus avancés (dont ComputeTree) — constituent la feuille de route prioritaire pour les prochaines phases.

RÉFÉRENCES BIBLIOGRAPHIQUES :

- **Chave, J.**, Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M. S., Delitti, W. B., ... & Vieilledent, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology*, 20(10), 3177–3190. <https://doi.org/10.1111/gcb.12629>
- **Feldpausch, T. R.**, Banin, L., Phillips, O. L., Baker, T. R., Lewis, S. L., Quesada, C. A., ... & Lloyd, J. (2011). Height-diameter allometry of tropical forest trees. *Biogeosciences*, 8(5), 1081–1106. <https://doi.org/10.5194/bg-8-1081-2011>
- **Feldpausch, T. R.**, Lloyd, J., Lewis, S. L., Brien, R. J. W., Gloor, M., Monteagudo Mendoza, A., ... & Phillips, O. L. (2012). Tree height integrated into pan-tropical forest biomass estimates. *Biogeosciences*, 9(8), 3381–3403. <https://doi.org/10.5194/bg-9-3381-2012>
- **Goodman, R. C.**, Phillips, O. L., & Baker, T. R. (2013). Amazon palm biomass and allometry. *Forest Ecology and Management*, 310, 994–1004. <https://doi.org/10.1016/j.foreco.2013.09.045>
- **Huertas, C.**, et al. (2025). The value of local allometries from airborne laser scanning for tropical forest biomass estimates. *Peer Community Journal*, 5, e103. <https://peercommunityjournal.org/articles/10.24072/pcjournal.625/>
- **Jiang, L.**, Zhao, H., Shi, S., Liu, S., Fu, C. W., & Jia, J. (2020). PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4867–4876. <https://doi.org/10.1109/CVPR42600.2020.00492>
- **Li, W.**, Guo, Q., Jakubowski, M. K., & Kelly, M. (2012). A new method for segmenting individual trees from the lidar point cloud. *Photogrammetric Engineering & Remote Sensing*, 78(1), 75–84. *Geoscience and Remote Sensing Letters*
- **Li, B.**, et al. (2026). ITS-Net: A platform and sensor agnostic 3D deep learning model for individual tree segmentation using aerial LiDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231. <https://doi.org/10.1016/j.isprsjprs.2025.12.002>
- **Ploton, P.**, Mortier, F., Réjou-Méchain, M., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1), 4540. <https://doi.org/10.1038/s41467-020-18321-y>
- **Silva, C. A.**, Crookston, N. L., Hudak, A. T., & Vierling, L. A. (2016). Imputation of individual longleaf pine (*Pinus palustris*) attributes from field and LiDAR data. *Canadian Journal of Remote Sensing*, 42(5), 554–572. <https://doi.org/10.1080/07038992.2016.1196582>

- 
- **Wielgosz, M.**, et al. (2024). SegmentAnyTree: A sensor and platform agnostic deep learning model for tree segmentation using laser scanning data. *Remote Sensing of Environment*, 313, 114367. <https://doi.org/10.1016/j.rse.2024.114367>
 - **Xiang, B.**, et al. (2023). Deep learning for individual tree segmentation in complex forest scenes. *International Journal of Applied Earth Observation and Geoinformation*, 122, 103431.
 - **Zolkos, S. G.**, Goetz, S. J., & Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, 128, 289–298. <https://doi.org/10.1016/j.rse.2012.10.017>

ANNEXES



Sources: OpenStreetMap, Geofit

ANNEXE 1 : DONNEES DERIVEES DU NUAGE DE POINT SUR UNE DES PLACETTES EN GUYANE (PARTANT DU HAUT MNS : MODELE NUMERIQUE DE SURFACE, MNE : MODELE NUMERIQUE D'ELEVATION, MNT : MODELE NUMERIQUE DE TERRAIN ET EN DERNIER UN MNHC : MODELE NUMERIQUE DE HAUTEUR DE CANOPEE, TOUS A 10CM DE RESOLUTION)



1



2

ANNEXE 2: NUAGE DE POINT D'UNE DES PLACETTES SUR L'ÎLE DE LA REUNION (1 PLACETTE ENTIÈRE ET 2 UN ZOOM SUR LE NUAGE DE POINT COLORISÉ)

Outil / Méthode	Catégorie	Statut	Points forts	Limites principales	Recommandation
Silva et al. 2016 (segmentation MNHC)	Segmentation LiDAR 2D.5	Fort potentiel	Simple, robuste, sans annotation	Aveugle au sous-étage dominé	Évaluer ComputeTree en complément sur la phase suivante
ComputeTree (ONF)	Segmentation LiDAR 3D	À évaluer (prioritaire)	Reconstruction 3D complète de l'arbre (branches + tronc), conçu pour nuages haute densité	Non encore testé dans ce projet	Priorité absolue pour la phase suivante
lidaRtRee (INRAE)	Segmentation LiDAR	Écarté à ce stade	Validé en forêt tempérée	Nécessite des ajustements importants pour la forêt tropicale guyanaise	À reconsidérer uniquement dans le cadre d'un projet dédié

Detectree2	Segmentation optique 2D	Fort potentiel	Bon taux de détection global des houppiers	Aveugle au sous-bois ; ne détecte que la canopée dominante et visible dans l'orthophoto	Combiner avec SAM3 pour optimiser la détection et la précision des contours
SAM3 / Segment Anything Model 2025	Segmentation optique 2D	Fort potentiel	Précision supérieure dans la délimitation des contours de couronnes	Taux de détection plus faible que Detectree2 ; aveugle au sous-bois	Combiner avec Detectree2 ; optimisation en cours
PointGroup / ITS-Net / SegmentAnyTree / MinkowskiEngine	Deep Learning 3D	Non testé	État de l'art	Requiert des milliers d'arbres annotés manuellement ; coût et délai incompatibles	Projet dédié avec budget et ressources pour la constitution d'une base d'annotation
Protocole hybride Huertas et al. 2025 + Chave et al. 2014	Calcul de biomasse de référence	Opérationnel (Guyane)	Réduit l'erreur relative de 27,7 % à 15,3 % vs approche sans hauteur ; robustesse confirmée par propagation bayésienne ($\pm 0,7$ à $\pm 1,1$ t/ha)	Dépend du Canopy Constructor calibré spécifiquement pour Paracou ; non transposable en l'état à La Réunion	Maintenir pour la Guyane ; engager la calibration du Canopy Constructor pour La Réunion avec CIRAD-IRD/AMAP
Allométries pantropicales (Feldpausch et al. 2011/2012 + Chave et al. 2014)	Calcul de biomasse de référence	Fort potentiel (solution interim Réunion)	Applicable sans Canopy Constructor ; immédiatement déployable	Biais d'estimation potentiel selon la composition spécifique du peuplement	Retenir comme alternative à court terme pour La Réunion en quantifiant et documentant explicitement les biais
Random Forest (régression multivariée)	Modélisation prédictive AGB	Fort potentiel — non déclenché	Pipeline de préparation des données entièrement fonctionnel ;	4 placettes disponibles vs minimum 40-50 requis (Zolkos et al.	Conditionné à : (1) acquisition de 60+ placettes spatialement

		ble à ce stade	descripteurs LiDAR physiqueme nt cohérents	2013) voire 60+ en tropical (Ploton et al. 2020) ; segmentatio n ne capte que la strate dominante	indépendantes ; (2) amélioration de la segmentation pour couvrir l'ensemble du profil vertical
--	--	-------------------	---	---	--

ANNEXE 3 : SYNTHÈSE DU PROTOCOLE ET RECOMMANDATIONS