Département de géomatique appliquée
Faculté des lettres et sciences humaines
Université de Sherbrooke

Prédiction de la distribution des diamètres des arbres à l'aide de métriques tirées de la donnée lidar aéroporté pour les forêts boréales du Québec et de l'ouest de Terre-Neuve

Xavier Gallagher-Duval

Mémoire présenté pour l'obtention du grade de Maître ès sciences géographiques (M.Sc.),

Cheminement Géomatique profil recherche

Juin 2023 :

Directeur de recherche :
Richard Fournier
Département de géomatique appliquée
Faculté des lettres et sciences humaines
Université de Sherbrooke, Qc.

Co-directeur de recherche :
Olivier van Lier
Service Canadien des Forêts
Centre Canadien sur la fibre de bois
Resource Naturelles Canada,
Corner Brook, Terre-Neuve et Labrador.

Membre du jury interne :
Samuel Foucher
Département de géomatique appliquée
Faculté des lettres et sciences humaines
Université de Sherbrooke, Qc.

Membre du jury externe :
Institut de Recherche sur les forêts
Osvaldo Valeria,
Institut de recherche sur les forêts
Université du Québec en Abitibi-Témiscamingue, Qc

Résumé

Citation : Gallagher-Duval, X. (2022). Prédiction de la distribution des diamètres des arbres à l'aide de métriques tirées de la donnée lidar aéroporté pour les forêts boréales du Québec et de l'ouest de Terre-Neuve. Mémoire de maîtrise, Département de géomatique appliquée, Université de Sherbrooke, 71 pages.

La forêt occupe une proportion importante du territoire Canadien et son exploitation nécessite une connaissance approfondie de la structure des peuplements forestiers. La distribution des diamètres des arbres (DDA) permet d'estimer plusieurs attributs forestiers, notamment le volume de bois ou le taux de croissance. Cette étude vise à prédire la DDA à une résolution fine (20 m²) à partir de données de LiDAR aéroporté pour les forêts boréales de conifères au Québec et à l'ouest de Terre-Neuve. La donnée LiDAR aéroportée permet de produire un modèle de hauteur de canopée. Conséquemment, le premier objectif vise à améliorer les estimés de DDA avec la contribution de métriques texturales dérivées du modèle de hauteur de canopée combinées aux métriques LiDAR standards. Le deuxième objectif consiste à déterminer la meilleure approche pour modéliser les DDA : soit en différentiant a priori la modalité des DDA selon leur modalité (uni/bimodales) ou non. La modélisation de la DDA passe par la prédiction des paramètres de la fonction Weibull ajustée aux DDA unimodales et non-différentiées. Pour les DDA bimodales, un f*inite mixture model,* composé de deux fonctions Weibull, permet d'extraire les paramètres des deux composantes Weibull ajustées à la DDA.

Les paramètres, *échelle* et *forme*, des fonctions Weibull décrivant les DDA unimodales et non différenciées ont été prédits avec des R^2 acceptables (0.40-0.55) comparativement aux paramètres *moyenne*, *proportion* et *écart-type* des DDA bimodales (R^2 moyen<0.30). L'utilisation de métriques de texture a permis d'améliorer la précision globale de la différentiation des modalités de 4%, ce qui a fait augmenter en moyenne de 0.10 le R^2 pour les paramètres des DDA unimodales et non-différentiées, et 0.17 pour les DDA différentiés bimodales. De plus, les DDA unimodales ont prédit en moyenne 79% des diamètres mesurés et 75% pour les bimodales. Les DDA non-différentiées ont prédit 76% des diamètres mesurés. Cependant, les DDA bimodales présentaient des R^2 faibles, causé par

l’absence de deux modes clairement distincts ainsi eu par la difficulté de prédire les faibles diamètres.

Mots clés : lidar, attributs forestiers, distribution des diamètres des arbres, modélisation.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

AAR Analyse axée sur la région

BC Coefficient de bimodalité (*bimodal coefficient*))

BLA Balayage laser aéroporté

DDA Distribution des diamètres des arbres

DHP Diamètre à hauteur poitrine

GMM *Gaussian mixture models*

MHC Modèle de hauteur de canopée

PDA Photogrammétrie digitale aéroportée

PDF *Probability density function*

Remerciements

Le chemin parcouru pour la réalisation de ce projet a été des plus enrichissant. Je tiens tout d'abord à remercier mon directeur, Pr Richard Fournier m'a proposé ce projet de maitrise des plus stimulant. Ses multiples conseils éclairés et pertinents m'apportaient une vision plus grande du monde de la télédétection forestière, étant donné que j'ai la mauvaise habitude de creuser et de foncer dans les détails. Plus d'une fois, ses conseils me remettaient sur la bonne voie. Je suis des plus reconnaissant pour tous ses encouragements et inestimables conseils qui ont contribué à la réalisation de ce mémoire. Je tiens également à remercier Mr. Olivier van Lier qui m'a accueilli plus d'une fois au sein de son équipe pour des stages, mais également pour la réalisation de ce projet. C'est lors de ces stages à Terre-Neuve que ma passion pour la télédétection forestière est née. J'avais la chance de collaborer au quotidien avec Olivier et j'ai pu acquérir un immense bagage de connaissances lors de nos discussions et réflexions. Par ailleurs, ses commentaires étaient toujours les bienvenus et ils permettaient de m'améliorer dans mon domaine de recherche. Je tiens également à remercie Joan Luther, qui m'a également encadré et guidé avec efficacité dans le monde de la modélisation. Elle m'a donné la piqure de R et m'encourageait à comprendre les différentes composantes de ma méthode pour ne rien laisser au hasard. Je remercie également ma conjointe, Stéphanie, qui savait me faire rire et m'encourager dans les moments les plus durs. Merci d'avoir été à mes côtés. À mes amis, Sam, Kevin, Alex, merci de m'avoir aidé à me divertir et d'avoir été là. Encore une fois, Richard, Olivier, merci du fond du cœur d'avoir cru en moi et de m'avoir épaulé tout au long de ce projet. Merci.

Avant-Propos

Ce mémoire est divisé en 4 parties. La première partie correspond à l'introduction générale qui contient les définitions du contexte, de la problématique, de l'État des connaissances ainsi que les objectifs et hypothèses. La deuxième partie présente le manuscrit scientifique publié dans *Forest.* L'article définit le site d'étude, les données et les méthodes employées dans la démarche méthodologique développée pour prédire la distribution des diamètres des arbres, ainsi que les résultats obtenus pour le site d'étude. La troisième partie comporte des analyses complémentaires utilisées pour évaluer la modalité des distributions des diamètres des arbres. Par ailleurs, cette section comporte également une approche alternative pour la prédiction de la distribution des diamètres des arbres. Finalement, une discussion permet de reprendre les grands points abordés dans l'article et dans les compléments de l'article.

1 Introduction

1.1 Problématique

Au Canada, la forêt recouvre plus de 347 millions d'hectares, ce qui représente 35% de son territoire (NRCan, 2018). Cette étendue contribue à 24.6 milliards de dollars de son produit intérieur brut, soit 1.6% de celui-ci (NRCan, 2018). De la planification à la transformation, 209 940 employés participent à ce secteur économique, soit 1.1% de tous les employés canadiens. Les arbres approvisionnent les scieries, les pulperies et les bioraffineries sur tout le territoire du Canada. Sélectionner les bons arbres au bon moment en vue d'une utilisation précise nécessite une connaissance approfondie de l'état de la forêt. La gestion des forêts évolue avec l'amélioration et le développement constant des technologies. Cependant, l'exploitation de la forêt doit se réaliser de façons durables. Avec la demande grandissante des ressources forestières, il y une augmentation du désir de protéger cette ressource pour ces services écosystémiques comme la captation de carbone (Diao *et al*., 2022). Par ailleurs, les marchés liés aux milieux forestiers sont soumis à une forte compétition à l'international (White *et al.*, 2016). L'optimisation de toutes les étapes de la chaine d'approvisionnement représente maintenant un facteur clé dans l'industrie forestière actuelle (Kvalvik *et al*., 2020; Shabani *et al*., 2013; Shahi and Pulkki, 2015). Les décisions stratégiques tant sur la planification et l'approvisionnement doivent se baser sur des informations précises, spatialement accessibles et à jour (Groot *et al*., 2015).

Les inventaires forestiers reposent sur des mesures dans des placettes terrain. Elles contiennent des informations descriptives du peuplements comme la hauteur des arbres, leur diamètre hauteur poitrine (DHP), l'espèces, la surface terrière pour n'en nommer que quelques-unes. Ces informations sont par la suite utilisées à divers degrés dans la planification des opérations forestières. Il est possible de définir trois niveaux de planification (Vandendaele *et al*., 2021). Le niveau à long terme ou stratégique qui comprend les planifications opérationnelles à l'échelle du territoire et sur un horizon d'environ 20 ans. En complément, la planification au niveau tactique couvre un horizon de 5 ans et comprend la planification des cibles à respecter selon les différents scénarios sylvicoles possible au niveau du paysage. Finalement, le niveau opérationnel s'étend sur

une période d'un an et contient principalement les opérations de récoltes. Il s'agit du niveau de planification le plus fin. Ces trois niveaux nécessitent des informations dérivées des inventaires forestiers à différentes échelles spatiales. Il est bénéfique d'extraire le plus d'information possible des inventaires terrain pour appuyer la planification forestière aux niveaux stratégique, tactique et opérationnel.

La modélisation spatiale à l'aide de données issues de la télédétection permet d'utiliser les mesures terrains pour calibrer des modèles prédictifs des attributs forestiers et ensuite les appliquer pour produire des cartes sur de grands territoires (White *et al*., 2016). Les gestionnaires forestiers peuvent utiliser ces informations pour mieux comprendre le territoire à gérer, mais surtout pour planifier les opérations futures. Ces nouvelles informations peuvent complémenter celle déjà existantes, améliorer la qualité et la couverture spatiale et réduire l'incertitude (Ginzler, 2019; Lechner *et al*., 2020; Tomppo *et al.*, 2008). La prédiction des attributs forestiers d'intérêts comme la surface terrière ou la hauteur moyennes sont déjà utilisé au niveau opérationnel à l'aide des données lidar ou diverses sources d'imageries (Valbuena *et al.*, 2014). Cependant, certains attributs reposent uniquement sur les inventaires forestiers et ce, même de façon opérationnelle. Un tel attribut est la distribution des diamètres des arbres (DDA).

La DDA constitue un attribut forestier prisé et riche en informations. Cette distribution correspond à un histogramme de fréquences des DHP retrouvés dans une superficie définie, généralement de l'ordre du peuplement ou de la placette (Leclère *et al.*, 2022). Au Québec, les placettes forestières se présentent sous une forme circulaire avec 400 m^2 de surface (11.28 m de rayon). La DDA renseigne aussi sur les perturbations passées, les patrons de régénérations, le volume de bois, la dynamique et les changements de composition ainsi que le taux de croissance du peuplement (Penner *et al.*, 2015). De plus, la DDA peut procurer un aperçu des classes de diamètres présentes dans le peuplement, une caractéristique corrélée avec la diversité d'espèces (Fries *et al.*, 1997). Par ailleurs, la DDA est une composante essentielle pour le calcul du rendement, mais également pour le potentiel de stockage de carbone (Zhang *et al*., 2019). Certains, comme Rubin *et al.* (2006), utilisent la DDA pour évaluer la durabilité de la forêt, en se basant sur la quantité et la taille de la cohorte d'arbres croissant. Knoebel and Burkhart (1991) l'utilisaient comme un

indicateur de valeur économique du peuplement, en plus d'estimer le type et le moment opportun pour chacune des étapes de la gestion d'un peuplement. Finalement, la DDA renseigne sur la diversité végétale et la diversité d'habitats (Strunk *et al.*, 2017) et à la réhabilitation à la suite des perturbations, car elle peut être couplée à des modèles de croissances et des simulations climatiques. Par exemple, Guo *et al*. (2022) ont utilisé une méthode de recouvrement des paramètres de DDA (simulée par une fonction Weibull) couplé à des variables climatiques. Le recouvrement des paramètres consiste à estimer des attributs reliés à la distribution (moments ou percentiles) et puis prédire les paramètres de la fonction (ex. fonction Weibull). Cette approche peur être paramétrisée à l'aide d'attributs forestier comme l'âge ou la hauteur (Zhang *et al*., 2019). Actuellement, l'évaluation de la DDA dépend principalement d'inventaires terrains couteux, longs et spatialement limités (Packalén and Maltamo, 2008). Ces limites sont très présentes dans les provinces comme Terre-Neuve et Labrador ou le Québec, due à leurs grandes superficies d'aménagements forestiers. La diversité en essences et en structures verticales des peuplements augmente la complexité de l'évaluation de la DDA. Les données issues de la télédétection ont le pouvoir d'améliorer les inventaires forestiers et spécialiser des attributs forestiers, notamment la DDA.

1.2 État de l'art

En pratique, il n'est pas réaliste de mesurer tous les arbres d'un peuplement pour obtenir la DDA (Maltamo *et al.*, 2007). En revanche, il est possible de faire des suppositions implicites que la DDA d'une placette est représentative du peuplement forestier (Garcıa, 1992). Prédire la DDA pour une aire d'étude, à partir de placettes terrain, représente une alternative efficace souvent contrainte par la taille du territoire. Généralement, la représentation de la DDA s'effectue avec une fonction dont les paramètres peuvent être estimés et prédits telles que les fonctions gamma ou log-normale (Figure 1-1). Étant une fonction de vraisemblance, l'axe des y d'une fonction Weibull correspond à la densité de probabilité, soit la probabilité qu'une variable aléatoire prenne une valeur donnée. Cependant, la fonction de Weibull s'avère beaucoup plus flexible et comprend un nombre réduit de paramètres. Bailey and Dell (1973) ont utilisé la fonction Weibull pour

quantifier et représenter la DDA de placettes équiennes. La fonction Weibull s'exprime mathématiquement à l'aide de deux paramètres *shape* et *scale*:

$$f(x) = \frac{c}{b}\left(\frac{x}{b}\right)^{c-1} exp\left[-\left(\frac{x}{b}\right)^{c}\right], x \geq 0\,; b, c \geq 0 \qquad (1)$$

où c représente le paramètre *shape*, b le *scale* et x, la variable aléatoire qui suit la Weibull, soit les fréquences de la DDA.

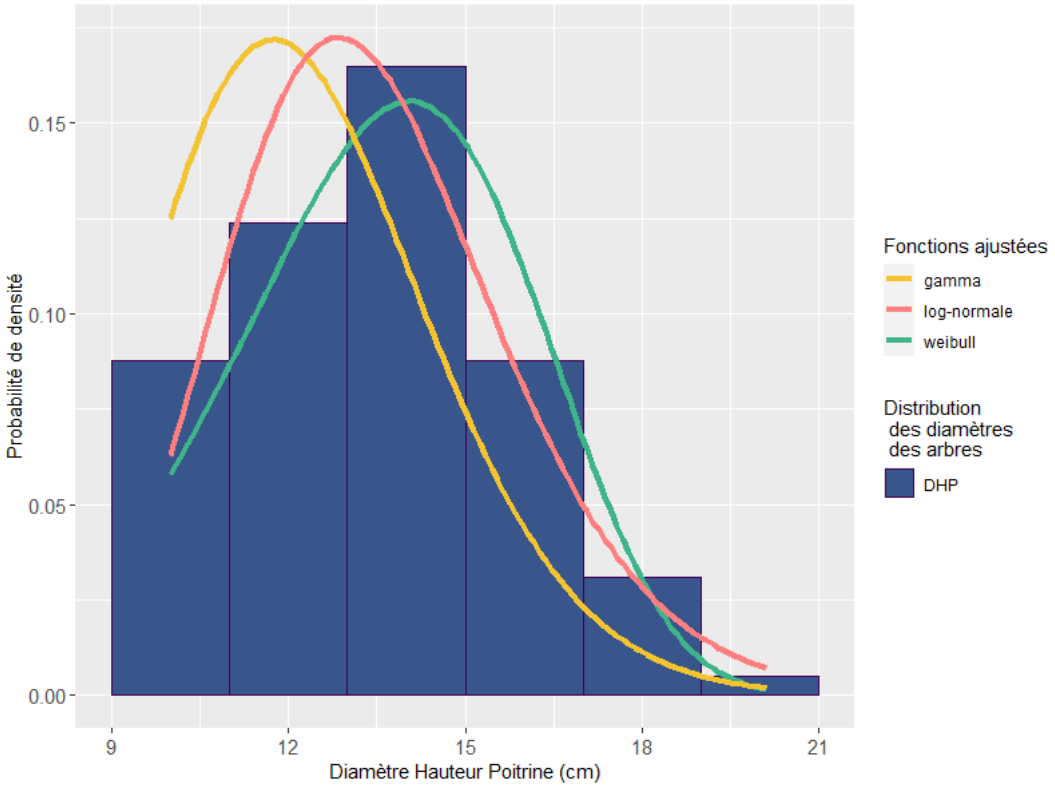

Figure 1-1: Représentation d'une distribution des diamètres des arbres avec une fonction de gamma, log-normale et Weibull ajustées sur les mesures de diamètres hauteur poitrine.

Podlaski (2016) a utilisé une fonction Weibull ajustée sur les mesures de DHP, puis les deux paramètres de la fonction estimés. Par la suite, les mesures terrain (ex: surface terrière et densité des tiges) servaient à créer des modèles de régressions paramétriques pour prédire les valeurs des deux paramètres de la Weibull. Cependant, cette fonction ne permet pas de bien représenter toutes les formes de DDA possible comme celles retrouvés pour les peuplements inéquiennes. Dans certaines circonstances, une distribution unimodale ne décrit pas adéquatement la forme générale de la DDA. C'est particulièrement le cas des distributions qui contiennent des discontinuités dans les diamètres, lesquelles peuvent être parfois observées dans les placettes de faible densité ou suite à des perturbations. Ces

placettes comprennent généralement quelques arbres avec des DHP élevés, mais également un plus grand nombre de petits arbres (Figure 1-2). Ajuster une fonction Weibull sur ce type de DDA aurait pour conséquence une surestimation des DHP moyens, car elle ne pourrait compenser leur absence.

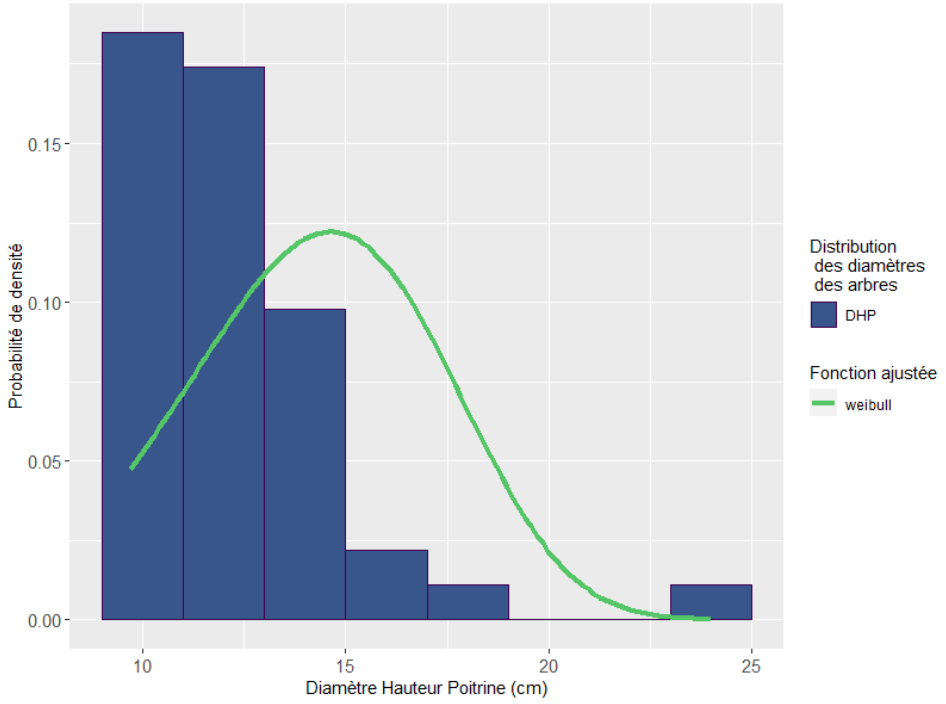

Figure 1-2 : Représentation d'une distribution des diamètres des arbres (DDA) ayant peut de valeurs de diamètres hauteur poitrine (DHP) au milieu de la distribution ainsi qu'une fonction Weibull à deux paramètres (*shape* et *scale*), ajustée aux observations.

Pour affiner la représentation graphique des DDA, il est utile de déterminer si la forme générale d'une DDA adopte l'une des deux formes possibles: unimodale ou multimodale (Mulverhill *et al.*, 2018). La classification de la modalité d'une distribution s'avère complexe et les méthodes existantes se basent sur différentes approches. Une approche simple est le coefficient de bimodalité (Ellison, 1987). Cette méthode nécessite le recours à trois paramètres, soit (1) la taille de l'échantillon, (2) le *skewness* et (3) le kurtosis. Le coefficient prend une valeur entre 0 et 1 et s'il est supérieur à $0.5\overline{5}$, la distribution est considérée comme bimodale. Une autre approche de différentiation de modalité est le Hartigan-Hartigan Dip test (Hartigan and Hartigan, 1985). Ce test permet le calcul d'une

statistique, le *dip*, soit la distance entre la distribution observée et une distribution théorique, et cherche la distribution unimodale la plus près des données observées. La distance maximale entre les distributions cumulatives permet de comparer la distribution observée et la distribution unimodale théorique (Johnsson *et al*., 2017). La valeur du d*ip* augmente plus la distribution observée s'éloigne de la distribution unimodale théorique.

Certaines caractéristiques des peuplements forestiers peuvent renseigner sur la modalité de la DDA. Dans un peuplement forestier, la courbe de Lorenz correspond à la proportion cumulée de la surface terrière en fonction de la proportion cumulée des individus (Figure 1-3). À noter que les arbres sont classés en ordre croissant de surface terrière lors de la création de la courbe.

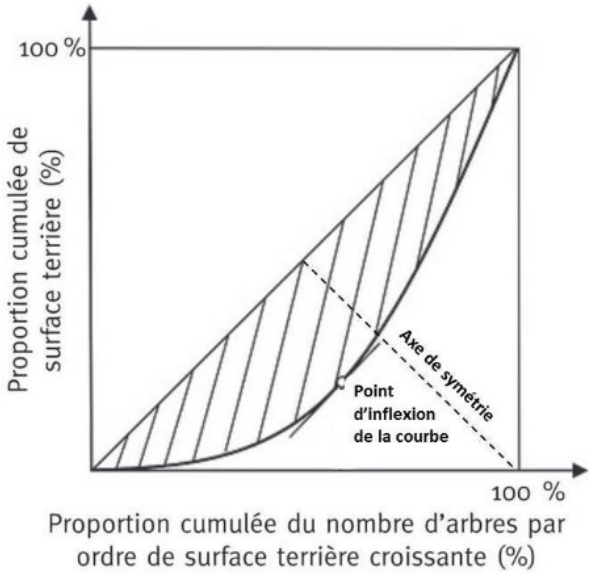

Figure 1-3: Représentation d'une courbe de Lorenz théorique pour une placette terrain (Cordonnier *et al*., 2012).

La diagonale représente le cas parfait où tous les arbres ont la même valeur de surface terrière. Plus un peuplement est considéré hétérogène, plus la courbe de Lorenz s'éloigne de la diagonale (Valbuena *et al*., 2012). Une caractéristique de la courbe de Lorenz est son asymétrie, soit sa déviation de la diagonale. Elle est définie par la position relative du point d'inflexion par rapport à l'axe de symétrie. Un point d'inflexion situé au-dessus de l'axe de symétrie peut représenter la présence de gros arbres. Au contraire, un point d'inflexion sous l'axe de symétrie correspond à une présence plus élevée d'arbre plus

petits. Une autre caractéristique de la courbe est l'indice de Gini qui correspond à deux fois la surface entre la diagonale et la courbe.

Par ailleurs, certaines caractéristiques de la courbe de Lorenz, comme l'asymétrie ainsi que l'indice de Gini, représentent des indicateurs de la structure de la forêt. Ces deux indicateurs ont déjà été utilisés pour caractériser les inégalités de la taille des arbres pour stratifier la forêt en peuplement homogène et hétérogène (Valbuena *et al.*, 2013). Zhang *et al*. (2019) ont calculé la courbe de Lorenz pour chaque placette et ensuite ont dérivé l'asymétrie et l'indice de Gini. Puis, ils ont établi que la DDA était considérée bimodale lorsque la valeur de l'asymétrie de la courbe de Lorenz < 0.5 et que l'indice de Gini > 0.5.

Une fois classifiées, les DDA unimodales sont représentées par une fonction Weibull tronquée et ses paramètres *shape* et *scale* sont modélisés. Une fonction Weibull ne permet pas de bien représenter les DDA plus complexes (McRoberts, 2012). Pour les DDA plus complexes comme pour les multimodales, l'utilisation de l'approche des *finite mixture model* permet d'ajuster deux fonctions Weibull, une sur chaque composante de la DDA. Zhang *et al.* (2019) appliquaient cette approche et la classification des DDA, selon leur modalité, pour améliorer la prédiction de la DDA. Le principal avantage d'utiliser une distribution paramétrique comme la Weibull correspond au faible nombre de paramètres (en l'occurrence deux). Certains auteurs, comme Thomas *et al.* (2008), suggèrent de créer des modèles spécifiques à l'espèces dominante. Ces derniers ont traité les placettes de feuillus séparément des placettes dominées par les conifères et les placettes mixtes pour obtenir de meilleurs résultats.

Il existe deux approches qui permettent the prédire la DDA. La première approche vise à utiliser des prédicteurs, issus de différentes données de télédétection, pour prédire les paramètres qui composent la distribution de Weibull (Gobakken and Næsset, 2004). La 1re approche implique une régression paramétrique indirecte pour inférer la DDA par recouvrement des paramètres. Premièrement, des attributs forestiers sont utilisés pour estimer les valeurs des paramètres de la DDA, par exemple, le paramètres *shape* et *scale* d'une fonction Weibull. Ainsi, les attributs forestiers tels que l'âge de la placette, le DHP ou la hauteur moyenne peuvent servir à estimer les paramètres de la fonction Weibull

(Siipilehto and Mehtätalo, 2013). Le recouvrement des paramètres nécessite autant d'attributs forestiers liés à la distribution qu'il y a de paramètres dans la DDA (ex. : les deux paramètres de la fonction Weibull *shape* et *scale*). Ces attributs peuvent être des moments (ex. : moyenne quadratique et moyenne), des quantiles de la distribution, ou encore des caractéristiques dérivées de la distribution (ex. : médiane). En général, les moments et les quantiles des attributs forestiers sont plus utilisés et mieux compris en raison de leur relation directe avec les caractéristiques de la placette (Gobakken and Næsset, 2004). La 2^{e} approche par régression paramétrique permet d'établir un système d'équation entre des attributs de la placette et leur expression basée sur une distribution de Weibull (Mauro *et al*., 2021). Cependant, l'approche de prédiction des paramètres de la fonction Weibull (1ère approche) produit de meilleures performances pour les forêts boréales du centre de l'Ontario (Mulverhill *et al*., 2018).

Différentes combinaisons de mesures terrain et de diverses sources de données issues de la télédétection offrent la possibilité de caractériser la DDA. Tarp-Johanssen (2002) utilisait un modèle tri-dimensionnel et des photographies aériennes digitalisées pour estimer la distribution des diamètres pour des peuplements mono-espèces (*Quercus robur* L.) au Danemark. Cependant, les images ont été acquises en période sans feuilles pour mieux voir les troncs, ce qui limite la réutilisation des images aériennes. Avec le développement de nouvelles technologies de télédétection, le Balayage Laser Aéroporté (BLA ou en anglais : *Airborne Laser Scanning* - *ALS*) permet de diversifier les méthodes de prédictions de la DDA. Gobakken and Naesset (2004) ont utilisé différentes métriques BLA décrivant la hauteur pour prédire avec précision (R^2 entre 0.20-0.90 et RMSE < 0.15) les deux paramètres des distribution Weibull pour estimer la DDA de forêts boréales du sud-est de la Norvège dominées par l'épinette de Norvège (*Picea abies* (L.) Karst) et le pin sylvester (*Pinus sylvestris* L.). De plus, les technologies de télédétection multi-sources peuvent également être combinées dans le but d'améliorer l'exactitude des prédictions. Peuhkurien *et al.* (2018) utilisaient une combinaison de données BLA et l'image satellitaire de SPOT5 pour estimer avec exactitude (indice de Reynolds situé entre 17.99 et 122.94) la DDA pour la région de Perm en Russie, composée essentiellement d'épinettes et de pins. L'indice de Reynolds (Reynolds *et al.*, 1988) a été utilisé pour évaluer si une distribution est bien représentée par une distribution de densité

de probabilité lors de la comparaison entre la DDA observée et prédite. Les valeurs de l'indice se situent entre 0 et 200, où une valeur de 0 indique deux distributions identiques, et 200, deux distributions complètement différentes. Dans le cas de Peuhkurinen *et al*. (2018), les DDA étaient stratifiées par espèce dominantes, puis combinées. Lorsque l'ensemble des placettes étaient combinées, les valeurs de EI pour les placettes variaient entre 17.99 et 122.94 avec une moyenne de 66.57. La modélisation des paramètres de la DDA s'effectuait avec la méthode *K-Nearest Neighbours*. En plus des métriques de hauteur, des métriques d'intensité à partir des données BLA peuvent aussi être utilisées. Les métriques d'intensité apportent une information sur l'énergie des impulsions laser rétrodiffusées. Shang *et al.* (2017) ont utilisé une combinaison de métriques de hauteur et d'intensité pour prédire la DDA de forêts de feuillus tolérants de l'Ontario. Leurs résultats montrent une amélioration des modèles prédictifs lorsqu'ils ont ajouté des métriques d'intensité aux métriques de hauteur.

De nouvelles sources de données peuvent servir à générer différents prédicteurs pour améliorer les prédictions de la DDA. Le BLA apporte une information en 3D des attributs structuraux de la canopée à l'intérieur du nuage de points. Différentes métriques dérivées du nuage de points BLA contiennent des informations structurelles de la canopée balayée. Au courant de la dernière décennie, plusieurs recherches portaient sur l'intégration des données BLA dans le développement des inventaires forestiers (Thomas *et al.*, 2008; White *et al.*, 2016; Ullah *et al.*, 2017; Noordermeer *et al.*, 2019). Les données BLA occupent une place prépondérante dans le développement de nouveaux métriques pour améliorer les inventaires forestiers (White *et al.*, 2016). Il a été démontré que les données BLA produisent des mesures et des estimations de caractéristiques d'inventaires précises et comparables aux mesures terrain et parfois même les surpassent (Maltamo and Gobakken, 2014). Les données BLA offrent la possibilité de cartographier différents attributs forestiers, pour de grandes superficies de l'ordre du peuplement ou du paysage. Des métriques calculées à partir du nuage de points décrivent trois principales structures: la densité de points, la hauteur moyenne ainsi que l'occlusion de la canopée (Hudak *et al.*, 2008). Les métriques BLA peuvent servir à prédire et cartographier une multitude d'attributs forestiers comme le volume des tiges, la densité des tiges ou la surface terrière (Bouvier *et al.*, 2015; Packalén and Maltamo, 2007; K. van Ewijk *et al.*, 2019). Une

approche couramment utilisée pour la cartographie des attributs forestiers, l'Analyse Axée sur la Région (AAR ou en anglais *Area-Based Approach - ABA*), procure des estimations précises et exactes à l'échelle de la placette (White *et al*. 2017). L'AAR prédit ces attributs forestiers à l'aide de métriques calculées au niveau de la placette. Les méthodes AAR emploient les métriques BLA comme variables indépendantes (explicatives) dans des modèles paramétriques ou non-paramétriques (Næsset, 2002). Cette approche prédictive est adaptée à la cartographie d'attributs forestiers pour de grands territoires.

La prédiction d'attributs forestiers à l'aide de l'approche AAR inclut généralement des métriques BLA. En plus des métriques BLA calculée directement à partir du nuage de points, il est possible de créer des données matricielles qui servent à calculer d'autres métriques. Le modèle de hauteur de canopée (MHC) est calculé à partir du nuage de point et représente la hauteur de la végétation normalisée au niveau du sol.

Plus récemment, des métriques de textures, comme les matrices de co-occurences des niveaux de gris (traduction libre de *Grey-Level Co-occurence Matrix* (GLCM))(Haralick *et al.*, 1973), calculées sur le MHC, sont incorporées à l'AAR pour améliorer la précision des modèles (Pippuri *et al.*, 2012; Packalen *et al.*, 2013; Niemi and Vauhkonen, 2016). Les GLCM sont une des principales méthodes statistiques pour examiner la texture. Ces matrices impliquent des statistiques de second ordre, définies comme la probabilité d'observer une certaine paire de valeurs de pixels dans une direction, un angle et une fenêtre d'observation prédéfinie (Tuceryan and Jain, 1993). Les métriques de textures dérivées de données optiques sont déjà utilisées pour prédire une variété d'attributs forestiers. Par exemple, l'ajout de la texture dérivée d'images Landsat aux ratios de bandes optiques a permis d'améliorer la prédiction de la biomasse aérienne avec un R^2 moyen de 0.76 comparative aux ratios des bandes Landsat-8 ($R^2 = 0.36$) (Dube *et al.*, 2015). Des indices de textures, dérivées d'imagerie aérienne, procuraient à Meng *et al.* (2016) une estimation de la biomasse aérienne à une résolution beaucoup plus fine qu'en utilisant uniquement leurs données BLA qui ont en moyenne 2 pts/m². La prédiction de la biomasse aérienne avait un R^2 de 0.88 lorsque la texture était utilisée comparativement à 0.79 avec seulement les données BLA. Par ailleurs, les métriques de textures peuvent

être dérivées de différentes sources de données (ex : BLA, données d'imagerie infrarouges à haute résolution aéroportée ou des satellites tel que le satellite ALOS AVNIR-2 avec une résolution spatiale de 10m) et elles ont l'avantage d'être indépendantes des caractéristiques spectrales étant donné qu'elles se basent sur la relation spatiale entre les pixels (Hou *et al.*, 2011). Les métriques de textures représentent une mesure de l'information spatiale continue et n'est pas nécessairement corrélée avec les informations spectrales (Hall-Beyer, 2017). Cela permet l'utilisation de plusieurs sources d'information différentes comme les images panchromatiques, multispectrales, aéroportées ou satellitaires. De plus, l'ajout des métriques de textures aux valeurs spectrales, toutes deux utilisées comme variables indépendantes (explicatives), améliore les résultats de modèles prédictifs la prédiction de modèles prédictifs d'attributs forestiers pour la forêt boréale (Persson, 2016).

Ozdemir and Donoghue (2013) ont utilisé une combinaison de ratios de percentiles de hauteurs, à partir du nuage de points ainsi que des métriques de textures, calculées sur le MHC, pour prédire la diversité des tailles d'arbres. Similairement, van Ewijk *et al.* (2019) ont observé que l'ajout de métriques de texture dérivées du MHC ainsi que des métriques BLA, ont amélioré les R^2 des prédictions de la surface terrière. Les améliorations ont augmenté les R^2 de 0.48 en employant uniquement des métriques BLA à 0.67 en ajoutant les métriques de texture. Ces métriques de texture décrivent plus précisément la structure horizontale de la forêt ou la variabilité de la canopée. Indirectement, la texture de la canopée semble informer sur le stade de développement du peuplement (Niemi and Vauhkonen, 2016).

Il existe différentes mesures permettant de comparer l'ajustement des données (traduction libre de *goodness-of-fit*) de la DDA prédite *versus* observée. Poudel and Cao (2013) ont utilisé quatre statistiques pour évaluer cet ajustement des données, soit la statistique de Anderson-Darling, la statistique de Kolmogorov-Smirnov, le Log-vraisemblance négative (*Negative log-likelihood*) et l'indice d'erreur de Reynolds. De plus, l'indice d'erreur de Reynolds a été utilisé dans différentes études sur le DDA qui utilisent des données lidar (Mulverhill *et al*., 2018; Palahi *et al.*, 2007; Zhang *et al.*, 2019).

1.3 Objectifs et hypothèses de recherche

Les études à ce jour permettent de supposer que l'amélioration des prédictions d'attributs forestiers est possible par des techniques de modélisation et l'usage de combinaisons de variables dérivées de différentes sources de données issues de la télédétection. À ce jour, aucune étude n'a spécifiquement évalué si l'inclusion de métriques de texture dérivées de la surface de la canopée améliore l'estimation de la DDA en les comparant avec les modèles issus des métriques directement disponibles des données BLA. Dans cette étude, notre objectif principal vise à établir si l'ajout de métriques de texture permet d'améliorer l'exactitude de la DDA prédite en obtenant de meilleures valeurs d'ajustement des données (*goodness-of-fit*). Il s'agira donc de développer des modèles prédictifs de la DDA avec différents groupes de métriques, soient les métriques BLA couramment utilisés, les métriques de textures issues du MHC et une combinaison des deux. Cette modélisation devra s'effectuer à partir de différentes méthodes statistiques (paramétriques ou non).

L'atteinte de notre objectif principal peut s'exprimer par deux hypothèses. Tout d'abord, notre première hypothèse énonce que les modèles qui incluent une combinaison de métriques BLA et de texture produiront les modèles les plus performants pour prédire les paramètres (*shape* et *scale*) de la DDA comparativement à l'utilisation de métriques BLA ou texturales seules. Notre deuxième hypothèse stipule que la différentiation des peuplements homogènes et hétérogènes en DDA unimodal et bimodal permettra d'améliorer la prédiction des DDA. Nous avons testé ces hypothèses en développant deux approches de modélisation. La première considère une connaissance *a priori* de la modalité des DDA (unimodale ou bimodale), et la deuxième considère toutes les DDA comme étant unimodales. Par la suite, nous avons développé des modèles spécifiques aux deux approches en utilisant les trois groupes de métriques, tout en évaluant la contribution des métriques de textures. Les trois groupes de métriques sont : BLA uniquement, texture et combinaison de BLA et texture. Finalement, nous déterminons quelle approche est la mieux adaptée pour estimer la DDA des forêts boréales du Québec et de l'ouest de Terre-Neuve

Research Article

Estimating stem diameter distributions with airborne laser scanning metrics and derived canopy surface texture metrics.

Xavier Gallagher-Duval [1], Olivier R. van Lier [2,*], Richard A. Fournier [1]

[1] Department of Applied Geomatics, Centre d'Applications et de Recherche en Télédétection (CARTEL), Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1, Canada. Xavier.Gallagher-Duval@USherbrooke.ca; Richard.Fournier@USherbrooke.ca
2 Canadian Forest Service – Canadian Wood Fibre Centre, Natural Resources Canada, Corner Brook, Newfoundland and Labrador A2H 5G4, Canada
* Correspondence: Olivier.vanLier@NRCan-RNCan.gc.ca

Abstract: Stem diameter distribution (SDD) describe tree diameter frequencies within an area and are of fundamental importance in various forestry applications. This study aimed to determine the optimal approach for estimating SDD from airborne laser scanning (ALS) data for the eastern boreal forests of Canada. We used ground plot data from Quebec and Newfoundland to develop area-based models (i) to classify SDD modality and (ii) to predict SDD function parameters. We developed three sets of predictor variables derived directly and indirectly from the ALS data at plot locations: point cloud metrics (M_{als}); canopy height model (CHM) texture metrics (M_{tex}); and a combination thereof (M_{comb}). We created three response datasets from the ground plots; the first two (unimodal, bimodal) were differentiated based on the modality of the SDD, while the third group (undifferentiated) assumed all plots to be unimodal. We used 70% of the plots for model development and the remaining 30% as an independent test case. SDD modality and associated function parameters were response variables for the SDD modality classification and SDD parameter prediction models which we tested for 5 modelling techniques, including Random Forest, support vector machine and generalized linear models with the elastic net penalty. Our results demonstrated little variability in the performance of SDD modality classification models (mean overall accuracy: 72%; SD: 2%). Unlike studies that fitted SDD function parameter models solely with M_{als}, our best models were generally fitted with M_{comb} with R^2 improvements up to 0.25. We found the variable Correlation, originating from M_{tex}, to be the most important predictor within M_{comb}. Trends in the performance of the predictor groups were mostly consistent across the modelling techniques within each parameter. With the best performing models, we evaluated the performance of predicting SDD using an Error Index (EI). The results of our test case indicated that differentiating modality prior to estimating SDD improved the accuracy of estimates for bimodal plots (~12% decrease in EI), which was trivially not the case for unimodal plots (< 1% increase in EI). We therefore concluded that (i) CHM texture metrics can be used to improve the estimate of SDD parameters and that (ii) differentiating for modality prior to estimating SSD is especially beneficial in stands with bimodal SDD. These results may provide for operational efficiencies in modelling and mapping SDD in boreal forest environment dominated either by black spruce or balsam fir.

Keywords: airborne laser scanning, texture, stem diameter distributions, forest inventory, boreal forest

Citation: Gallagher-Duval, X.; van Lier, O. R.; Fournier, R. A. Estimating stem diameter distributions with airborne laser scanning metrics and derived canopy surface texture metrics. *Forests* **2023**, *14*, 287. https://doi.org/10.3390/f14020287

- Academic Editor: Jianping Wu, Zhongbing Chang and Xin Xiong

Received: 10 January 2023
Accepted: 31 January 2023
Published: 2 February 2023

2.1 Introduction

Tree stem diameter distribution (SDD) is a central forest structural attribute that is used by foresters to estimate forest productivity [1]. SDD represents the frequencies of diameters at breast height (DBH, 1.3 m) within an area of predefined dimensions, i.e., typically a stand or a plot [2]. The distributions can be used to describe past disturbances, regeneration patterns, stand growth rates and wood volumes [3], and are known to be correlated with species diversity [4]. This information is an important aid for forest managers, who are planning silvicultural strategies [5], and assessing the economic value of given stands [6].

Airborne laser scanning (ALS) provides information on the 3D structure of the sampled forest and canopy. In the last decade, much effort has been devoted to modelling and mapping forest attributes from ALS data [7–10] to the point where these data are being used operationally over large continuous areas spanning many jurisdictions internationally (e.g., [11–13]). These studies demonstrated that ALS data can provide precise and reliable predictions of many forest attributes (e.g., gross merchantable volume, mean height, basal area) [12,14] using an area-based approach (ABA) [15]. It is possible to predict SDD using an ABA by adjusting a parametric function (e.g., Weibull) to a known SDD and predicting its parameters [16].

The general shape of the SDD is influenced by the stand's structural homogeneity. Homogenous stands tend to have unimodal distributions, while heterogeneous stands with more complex structures are commonly associated with multimodal distributions [17]. It has been demonstrated that a priori knowledge of the stand's complexity (i.e., differentiating between homogeneous and heterogeneous stands) can improve SDD predictions in structurally diverse forests [10], although this may be dependent upon the forest structure that is being assessed. There are numerous statistical measures that can characterize the modality of a distribution [18], such as the bimodality coefficient [19] and Hartigan's dip test [20], together with calculating the Gini index and the asymmetry of a Lorenz curve [16]. However, some techniques can be sensitive to heavily skewed distributions and can misclassify unimodal distributions where a long tail in either direction is observed [21].

Both nonparametric and parametric approaches have been used to predict SDD regardless of the distribution's modality [1,22–24]. Nonparametric methods, such as random forest and k-nearest neighbour, are better suited for heterogeneous, forest stand structures, given that they do not rely upon a probability distribution function (e.g., log-normal, gamma, Weibull) and can represent the local variability in the SDD [16]. Yet, the application of nonparametric approaches can be limited, as they can require very large quantities of training data [25]. In contrast, parametric methods such as generalized linear model and support vector machines do not require large training datasets and can be used to predict the parameters of a probability density function that is fitted over a measured SDD [26]. The Weibull distribution has been considered as one of the better-performing distributions for modelling SDD [27]. It is used widely due to its capacity for fitting a variety of shapes and its relative simplicity of implementation with only two parameters to predict or impute [28]. The Weibull distribution is better suited to representing unimodal SDD that are associated with homogenous stands, given that it contains only one mode [5]. Multimodal SDD, which are observed in heterogeneous or uneven-aged stands, cannot be well represented by a single Weibull distribution, as this would result in biased distribution characterizations [29,30]. In such instances, complex distributions can be represented within a Finite Mixture Model (FMM) by combining two or more Weibull distributions [7]. FMMs have been demonstrated to predict irregular or bi-modal SDD more accurately from ALS data compared to unimodal Weibull modelling in boreal forests [7,17].

Acquiring SDD for large areas relies heavily upon sampled ground plot data that are distributed throughout the entire study site in an effort to capture the full range of variability in SDD [5]. The use of remote sensing data, when linked with ground-sampled data, offers the possibility of characterizing SDD efficiently over broader areas. Tarp-Johansen (2002) [31] used a 3D model and digital aerial photographs to estimate stem diameters for monospecific

English oak (Quercus robur L.) stands in Denmark. With the development of ALS, Gobakken and Næsset (2004) [32] used various ALS height metrics to estimate Weibull parameters accurately (R2 ranging between 0.6-0.9 with RMSE: 0.15) to predict SDD for boreal forest in southeast Norway. Multi-source remote sensing data also can be combined to improve prediction accuracies. Peuhkurinen et al. (2018) [33] combined ALS data and SPOT5 imagery to make accurate predictions (Reynold's Error Index for all plots ranged from 17.99 to 122.94) of SDD for coniferous boreal forests of Russia's Perm Region with the non-parametric k-MSN (k-Most Similar Neighbour) method. In addition to height metrics, intensity metrics can be derived from ALS data, thereby providing indications of the strength of backscattered energy. Shang et al. (2017) [34] used ALS height and intensity metrics to predict SDD for a hardwood forest in Ontario, Canada (Haliburton Forest). They found that combining intensity and height metrics improved the model's performance beyond employing either height-only or intensity-only metrics.

Texture metrics that are derived from remote sensing can provide additional information regarding canopy structure that is independent of spectral features regarding spatial variations [35]. Haralick's Grey-Level Co-occurrence Matrix (GLCM) [36] is one common approach to calculate texture features from a given raster surface. GLCM uses sec-ond-order statistics, which are defined as the probability of observing a certain pair of pixel values within a predefined angle and observation window size [37]. Studies have demonstrated that texture metrics derived from optical data can be used successfully to predict a range of forest attributes [38]. Dube and Mutanga (2015) [39] compared aboveground biomass models that were derived from Landsat-8 spectral bands, spectral band ratios, vegetation indices, texture bands, and texture band ratios and found that models developed from multiple texture band ratios yielded the highest R2 values. Several studies have incorporated canopy height model (CHM)-derived texture metrics in predicting forest attributes. Ozdemir and Donoghue (2013) [40] used CHM-derived texture metrics to explain the tree diversity and found that the combination of ALS metrics with texture metrics explained up to 85% of the measured tree height diversity. Niemi et al. (2016) [41] demonstrated that using texture metrics improved prediction of total stem volume and basal area over models that were developed solely from ALS metrics. Similarly, van Ewijk et al. (2019) [38] found that combining ALS, CHM texture, and intensity metrics improved R2 by 0.19 for the prediction of stem density when compared to models that were developed solely with ALS metrics.

The aforementioned studies provide meaningful insight into potential improvements for predicting forest attributes using a variety of modelling approaches and predictor variables that are derived from remote sensing data. To date, no studies have specifically examined whether the inclusion of canopy surface texture metrics can improve the characterization of SDD from ALS data. In this study, we compared the accuracy of SDD predictions that were modelled independently from commonly used ALS metrics, CHM-derived texture metrics and a combination of the two, using multiple statistical modelling techniques. We first hypothesized that models using texture-derived metrics would more accurately predict SDD parameters than ones using ALS metrics alone. Second, based upon past research, we hypothesized that developing differentiated modality-specific models (unimodal/bimodal) would improve SDD predictions. We tested these hypotheses by developing two modelling approaches; the first considers a priori knowledge regarding the modality of the SDD, while the second considers all SDD to be unimodal. We then evaluated the contribution of texture metrics in both approaches and determined which approach is best suited for estimating SDD in the eastern boreal forests of Quebec and western Newfoundland.

2.2 Materials and Methods

2.2.1 Study Area

Two study areas were selected based on their similarity in forest composition: both are conifer dominated and lie within the eastern extent of the North American boreal forest [42] (**Figure 2-1**). The forests are comprised of balsam fir (*Abies*

balsamea (L.) Miller), black spruce (*Picea mariana* [Miller] Britton, white spruce (*Picea glauca* [Moench] Voss), paper or white birch (*Betula papyrifera* Marshall), yellow birch (*Betula alleghaniensis* Britton) and, to a lesser extent, tamarack or eastern larch (*Larix laricina* [Du Roi] K. Koch). We find balsam fir and white spruce dominated mixed stands south of the 50th parallel in our first study area (123,140 km²), located in the province of Quebec. As we move north, the presence of black spruce increases until it completely dominates the landscape above the 52nd parallel. The second study area (977 km²) is located in the most eastern extent of the Boreal Shield Ecozone in the province of Newfoundland and Labrador and is dominated by balsam fir. The climate in both sites is favourable for forest growth due abundant precipitation and warm summers. The primary silvicultural treatments practiced in these areas are pre-commercial thinning and clear-cut harvesting, which generally yields even-aged, homogeneous, forest stands.

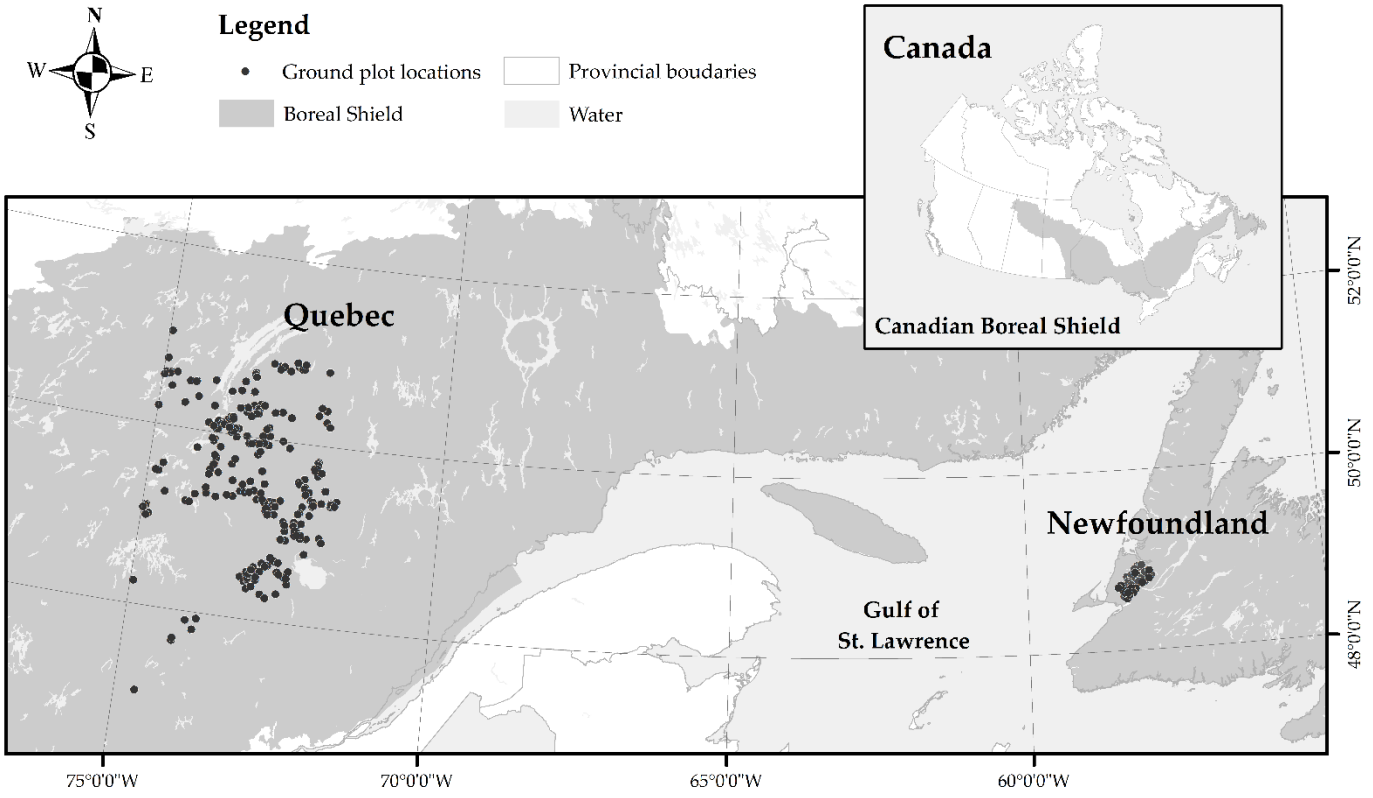

Figure 2-1: Plot Distribution across two sites within the eastern Boreal Shield, Canada

2.2.2 Ground Plots

Fixed-area circular plots were established with radii of 11.28 m where species, diameter at breast height (DBH), height, and status (live or dead) were recorded for all merchantable trees (trees ≥ 9 cm DBH). We retained plots having a total basal area ≥ 75% associated to balsam fir and/or balck spruce with a presence of ≤ 10% hardwoods. We then identi-fied and removed outlier plots by performing a multivariate local outlier factor analysis with the R package DMwR [43]. The analysis was based upon mean DBH and gross merchantable volume, together with the shape and scale parameters of a fitted Weibull function. We differentiated the SDD of each retained plot as unimodal or bimodal using the Bimodality Coefficient (BC) [19], given that its validity has been demonstrated in boreal forest environments [17] (**Figure 2-2**). The BC is proportional to the ratio between squared skewness and uncorrected kurtosis [18]. We associated plots having BC values ≤ 5/9 to unimodal distributions, while bimodal distributions were associated with BC values > 5/9 [21]. In total, we retained 307 plots differentiated as unimodal, and 120 as bimodal, for the anaslysis of our hypotheses.

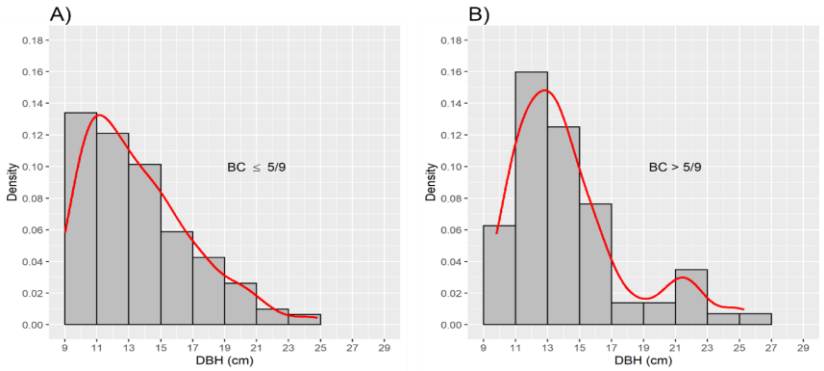

Figure 2-2: Example of Stem Diameter Distribution (SDD) from measured diameter at breast height (DBH) that was differentiated according to the Bimodality Coefficient (BC) as A) unimodal and B) bimodal, and fitted with a Weibull distribution and a Finite Mixture Model, respectively (red lines).

2.2.3 ALS Data and Metrics

All ALS data were acquired within 2 years of ground plot measurements between 2012 and 2016. We calculated the mean point densities from plot locations to be 5.8 points m−2 and 4.9 points m−2 for the Quebec and Newfoundland sites, respectively. We created a CHM at a 1 m × 1 m resolution from first returns that were classified as vegetation using a natural neighbour interpolation. Binning cell assignment was set to maximum value and zeros replaced negative values. We calculated ALS metrics that are commonly used to describe the height, structure and density of the canopy using the lidR package [44] in the R programming environment [45] using only returns ≥ 2 m that were classified as vegetation. We calculated GLCM edge (contrast and dissimilarity) and patch interior texture metrics from the CHM, i.e., correlation, homogeneity, mean and angular second-moment [46]. We considered three window sizes, 3 × 3, 5 × 5 and 7 × 7, for the GLCM texture feature calculations and determined that the 3 x 3 window produced metrics, which explained the most variation in our response variables (i.e., SDD Weibull parameters). We averaged the GLCM feature texture metrics in all directions and limited the number of grey-levels to 32. We then averaged the 1 m × 1 m resolution texture feature values for each ground plot location to produce associated metrics of texture. To evaluate our hypotheses, we grouped the pre-dictor variables into three sets of ALS metrics, based upon: (i) point cloud metrics (Mals); (ii) CHM texture metrics (Mtex); and (iii) a combination thereof (Mcomb) (**Table 2-1**).

Table 2-1: Description of metrics and associated groupings used as predictor variables: ALS metrics (M_{als}), texture metrics (M_{tex}), and combined ALS and texture metrics (M_{comb}).

Group	Metric	Units	Description
M_{als}	MAX	m	Maximum height
	MEAN	m	Mean height [47]
	P25, P75, P90	m	Height percentiles. E.g., P25 is the height of the 25th percentile. [48]
	SKEW		Skewness
	VAR		Variance [47]
	COVAR	%	Coefficient of variation: standard deviation / mean [49]
	VDR		Vertical Distribution Ration: (MAX - MEAN)/MAX [50]
	VCI		Vertical Complexity Index [51]
	ENT		Entropy: normalized Shannon diversity index [52]
	RI		Rumple Index of roughness [53]
	D2, D5, D8	%	Proportion of all vegetation returns found in sections divided within the range of heights of all returns for each plot. [54]
	COVER		Ratio of the number of vegetated returns above 2m to the total number of ground and vegetated returns [55]
	LPI		Light Penetration Index, Ground returns/(Ground returns + Canopy returns). [48]
	LPI1st		Light Penetration Index (first returns): Ground first returns/(Ground returns + Canopy returns) [56]
	FR		First return ratio: number of first return heights below a specified height threshold / total number of first return heights [47]
	RR		All return ratio: all returns < 2 m/all returns [57]
	LAI		Sum of Leaf Area Density [47]
	cvLAI		Coefficient of variation of Leaf Area Density [47]
M_{tex}	CON		Contrast (edge texture) [46]
	COR		Correlation (interior textures) [46]
	DIS		Dissimilarity (edge textures) [46]
	HOM		Homogeneity (interior textures) [46]
	MEAN		Mean (interior textures) [46]
M_{comb}		-	Combination of all metrics (M_{als} and M_{tex})

2.2.4 Overview of the Methods

Figure 2-3 provides an overview of the methodological approach of the study. We used the ground plot data to develop area-based models (i) to classify SDD modality and (ii) to predict SDD function parameters. We first defined three sets of ALS metrics from the ground plot locations (M_{als}, M_{tex}, M_{comb}). We then created three ground plot datasets: the first two, unimodal and bimodal, were differentiated based upon the modality of the SDD, while the third group was undifferentiated and assumed all plots were unimodal. Within each of the differentiated modality groups, we randomly selected 70% of plots for model development and used the remaining 30% for a test case. We developed models using 70% of the model development data for training and the remaining 30% for evaluating model performances. We generated three sets of models for each of the ground plot groups using the ALS metrics sets. We used the modality and associated Weibull parameters as response variables for the SDD modality classification models and the SDD parameter prediction models, respectively. We implemented our best performing models on our reserved test case data and analyzed the contribution of the CHM texture metrics to both groups of models (classification and prediction). Finally, we compared the predicted SDD that was obtained from the differentiated and undifferentiated modality models to assess whether modality differentiation improved prediction of SDD in our data. All calculations were performed in R [45].

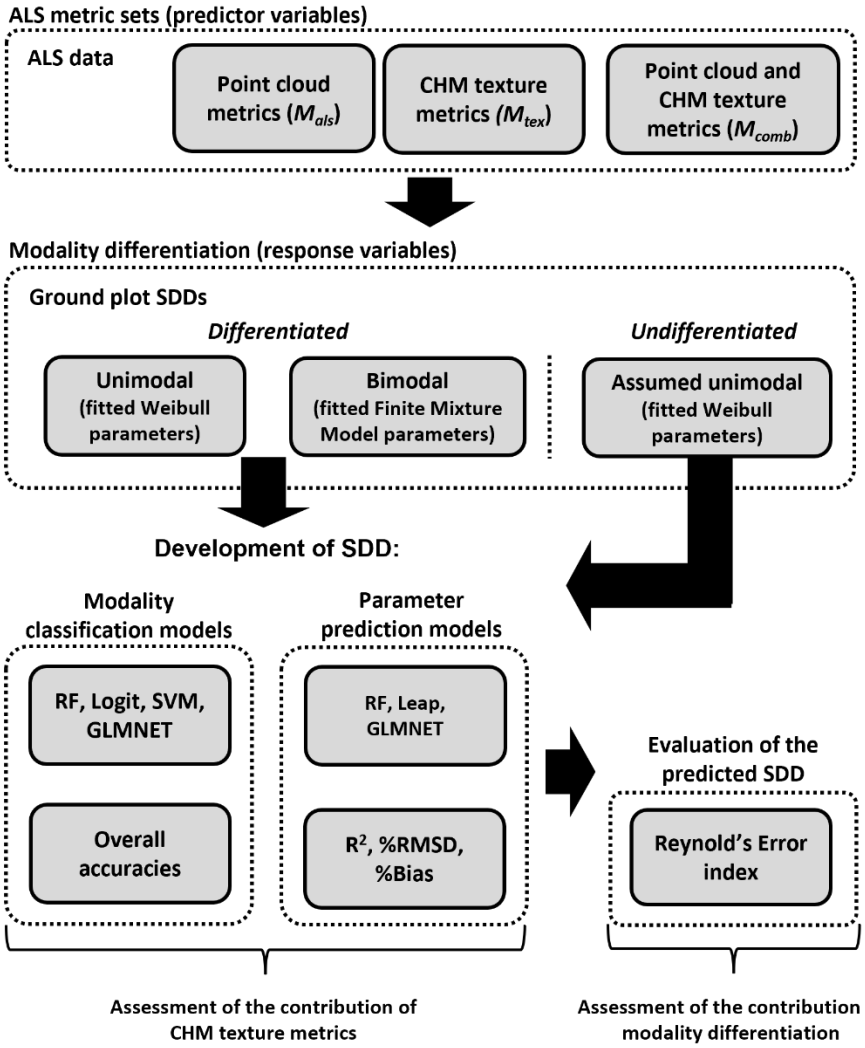

Figure 2-3: Overview of the methodological approach for assessing the contribution of CHM texture metrics and modality differentiation in predicting stem diameter distribution (SDD) parameters. CHM = Canopy Height Model; RF = Random Forest; Logit = generalized linear model with stepwise feature selection; SVM = Support Vector Machine; GLMNET = Generalized linear model through penalized maximum likelihood; Leap = Best subset regression with branch-and-bound algorithm; R^2 = Coefficient of determination; %RMSD = relative root-mean-squared deviation expressed as a percentage of the mean; %Bias = relative Bias expressed as a percentage of the mean.

2.2.5 Development of SDD Modality Classification Models

We developed classification models to classify the modality of SDD using the differentiated SDD modality plot datasets (unimodal and bimodal). We constructed models independently using the three metrics groups (Mals, Mtex and Mcomb) as predictor variables. Herein, we evaluated four statistical techniques: random forest (RF); generalized linear model (Logit); support vector machine (SVM); and generalized linear model through penalized maximum likelihood (GLMNET), which uses the elastic net penalty that mixes the lasso and ridge penalities [58]. These contained internal feature selection mechanisms for selecting the best predictors and model with the caret package [59]. We developed the RF models with the randomForest package [60] and optimized the parameter mtry, which controls the number of predic-tors that were randomly picked at each split, by testing five values, viz., 1, 2, 3, 4, and 5. Logit models were developed with the MASS package [61] and used stepwise model selection based upon the Akaike Information Criterion

(AIC). We defined the family parameter as binomial and conducted no grid search for parameter optimization. SVM models were developed with the kernlab package [62] and used a radial basis function. We tuned two parameters for SVM, sigma, which controls the rigidity of the decision boundaries, and C, which controls the influence of misclassification. Values for sigma were 2-25, 2-20, 2-15, 2-10, 2-5, and 20, while those for C were 20, 21, 22, 23, 24, and 25. Finally, GLMNET models were developed with the glmnet package [63]. GLMNET corresponds to a ratio between model regularization L1 and L2 af-fecting the penalty coefficient and allows the selection of relevant predictors [64]. The two parameters that were tuned were lambda, which controls the overall strength of the penalty, and alpha, which controls the gap between the L1 and L2 regularization. We tested alpha values ranging from 0 to 1 with 0.1 increments and the following lambda values: 0.0001, 0.1112, 0.2223, 0.3334, 0.4445, 0.5556, 0.6667, 0.7778, 0.8889, and 1. We repeated cross-validation 5 times using 70% of the model development data for training and 30% for validation. Finally, we averaged the overall accuracies within each technique and ALS metric group, and applied the best performing models to our test case dataset and as-sessed the contribution of CHM texture metrics.

2.2.6 Development of SDD Prediction Models

We developed three sets of models to predict SDD function parameters using i) differentiated unimodal, ii) differentiated bimodal, and iii) undifferentiated, SDD modality plot datasets. Using the differentiated unimodal plot data, we fitted a truncated Weibull function over the measured SDD and estimated the two function parameters (i.e., shape and scale) using the fitdistrplus package [65]. We implemented the same analysis for the undifferentiated plot data for which all plots were treated as having a unimodal SDD distribution. From the differentiated bimodal plot data, we fitted a FMM composed of two Weibull functions over the SDD. The 1st Weibull related to smaller stem diameters relative to the 2nd Weibull, which described the probability distribution of larger stems. The FMM can be represented by either the scale and shape, or the mean and standard deviation, of each of the two Weibull components and their associated proportions. We estimated the parameters of each function using the mixR package [66]. We assessed three modelling techniques within each model set, which included feature selection that was based on optimizing the root-mean-squared deviation (RMSD) using the caret package. Again, the three metric groups (Mals, Mtex and Mcomb) were used independently as predictor variables. The maximization option of RMSD was set to FALSE to ensure that the best combination of parameters produced the lowest RMSD. The first technique that was used was RF from the randomForest package. Again, the only optimized parameter with grid search was mtry, with values 1, 2, 3, 4, 5. The second technique was GLMNET with two parameters to optimize, i.e., alpha and lambda. The alpha that was tested ranged from 0 to 1 in 0.1 increments; lambda values were 0.0001, 0.1112, 0.2223, 0.3334, 0.4445, 0.5556, 0.6667, 0.7778, 0.8889, or 1. We implemented the third and final technique, i.e., best subset regression with branch-and-bound algorithm (LEAP) [67], with the R package leaps [68]. This best subset regression used the branch-and-bound algorithm (BnB; [69]), which solves and optimizes combinatorial problems to select the best subset of predictors. Herein, we defined the number of predictors allowed in each subset to range between 2 and 6 predictors.

We evaluated the best tuned models from de repeated 5 time cross-validation with the reserved test case dataset not used for model development. We compared the coefficient of determination (R^2), the absolute and relative RMSD (Equation (1) and Equation (2)), the absolute and relative bias (Equation (3) and Equation (4)) for both the model development and test case datasets to assess our two hypotheses:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-1}} \quad (1)$$

$$RMSD\% = \frac{RMSD}{\bar{y}} \times 100 \quad (2)$$

$$Bias = \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})}{n} \quad (3)$$

$$Bias\% = \frac{Bias}{\bar{y}} \times 100 \quad (4)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value for case i, n is the number of observations, and $\bar{y}$ is the mean.

In order to evaluate the composition of metrics used in the best performing models developped with M_{comb}, we calculated the associated variable importance. Since methods to characterize variable importance are dependent on the modelling technique implemented, we first scaled values between 0 and 100 to finally derive an average for each parameter modelled. For random forest models, we calculated variable importance as the percent increase in mean square error (noted %IncMSE) [70]. For GLMNET models, we scaled variable coefficients as a representation of variable importance since they are proportionally indicative of the variables' importance [64] due to the penalization that reduces the coefficients of less-important variables [63] Finally, we calculated variable importance for LEAP models as the absolute value of the t-statistic for each parameter in the final model [71].

2.2.7 Evaluation of the Predicted SDD

The quality of the predicted SDD was estimated with the Reynolds Error Index (EI) [72]. To do so, we predicted the SDD's parameters with the models demonstrating the highest R^2 an lowest RMSD% for the unimodal, bimodal, and undifferentiated plots from both model development and test case datasets. We then grouped the predicted tree DBH into 2 centimetre-wide bins to limit variability at larger intervals [73]. Finally, we evaluated goodness-of-fit between the predicted SDD and observed SDD of each plot with EI as follows:

$$EI = \sum_{i=1}^{m} 100 \left| \frac{f_{refi} - f_{alsi}}{N_{ref}} \right| \quad (5)$$

where m is the total number of bins, f_{refi} is the reference stem count for DBH bin i, f_{alsi} is the predicted stem count in DBH bin i, and N_{ref} is the true stem count of all DBH bins. EI values ranged between 0 and 200, where an EI of 0 indicated a perfect fit between predicted and observed SDD, which an EI of 200 indicated a completely different SDD. To assess the effects of modality differentiation, we averaged the EI from all plots that had been derived independently for both the differentiated (unimodal and bimodal) and undifferentiated modelling approaches.

2.3 Results

2.3.1 SDD Modality Classification Models

Table 2-2 denotes the overall accuracies of the modality classification models using the three ALS metric sets as predictor variables and four modelling techniques for both model development and test case datasets. During model development, we observed M_{als} and M_{comb} to perform best using RF and GLMNET (overall accuracy of 74%). Surprisingly the M_{tex} predictor set was used in both the best (using Logit) and worst (using RF) performing models in our test case. Overall, we observed little variability in the overall model accuracies regardless of the ALS predictor variable set or modelling technique used during model development or in our test case (mean: 72%; SD: 2% in both scenarios).

Table 2-2: Overall accuracies (%) of the SDD modality differentiation models using predictor variables that were derived from the three ALS metrics sets (M_{als}, M_{tex}, M_{comb}) for both model development and test case datasets.

ALS metric set	RF	SVM	Logit	GLMNET
Model development				
M_{als}	74	72	71	74
M_{tex}	73	72	68	68
M_{comb}	74	71	70	74
Test case				
M_{als}	72	73	72	71
M_{tex}	66	72	74	73
M_{comb}	71	71	72	71

2.3.2 SDD Prediction Models

We developed model sets to estimate probability distribution function parameters from the differentiated unimodal, differentiated bimodal and undifferentiated SDD modality plot datasets. We developed models within each model set using the three ALS metrics sets (Mals, Mtex, Mcomb) and three modelling techniques (RF, GLMNET, LEAP). The model performance measures (R2, RMSD%) that were derived from cross-validation are presented as supplementary material (S.1) as we observed for the most part the same trends in results with our case study illustrated in **Figure 2-4**. The results of our test case show that the proportion of the variance in the parameters describing the differentiated unimodal SDD were variable (R2: 0-0.62). We observed associated errors ranging between 9.9% and 13.4%, and 16.4% and 23.8% for models predicting scale and shape, respectively. For both parameters, the results indicate for all but one exception (Shape ~f(Mals) using RF), that models developed with Mcomb consistently outperformed models that were developed with either Mals or Mtex. Both parameters were best predicted with RF, scale was best predicted using Mcomb (R2: 0.62; RMSD%: 9.9%), while shape, using Mals (R2: 0.39; RMSD%: 16.4%).

The performance of models that were developed using the differentiated bimodal SDD modality plot data were again variable (R2: 0-0.53; RMSD%: 8.2%-52.1%). The results indicated that the FMM could not be represented by the parameters scale and shape; the parameter shape of the first Weibull component could not be predicted given the resulting models could never explain any of the variation in the parameter around it's mean (R2: 0), regardless of the ALS metric set or modelling approach. We therefore used the parameters mean and standard deviation to describe each component of the FMM. As expected, variation in the two proportion parameters was very poorly explained, if at all, by the predictor sets (R2: 0-0.15), with associated errors ranging from 17.5% to 36.9%. As expected, the two Weibull component proportions of the FMMs were poorly predicted with best predictions modeled with RF using Mals (R2: 0.15, 0.15; RMSD%: 17.5%, 33.8% for proportions of the first and second components, respectively). The parameter mean was best predicted using Mcomb for both components (R2: 0.27, 0.53; RMSD%: 8.2%, 14.4%; using LEAP and GLMNET for mean 1 and 2, respectively). Of note, GLMNET only marginally outperformed LEAP for mean of the second FMM component (increase in R2 < 0.01, decrease in RMSD% <0.13%), both using Mcomb. Standard deviation was best predicted with LEAP using Mtex for the first Weibull component (R2: 0.34; RMSD%: 45.13%) and using Mcomb for the second (R2: 0.43; RMSD%: 37.6%) with either LEAP or GLMNET.

The development of models using the undifferentiated modality SDD plot data involved applying the unimodal fit-ting analysis to all plots, regardless of modality. Herein, models performed better for the scale parameter (R2: 0.37-0.73; RMSD%: 8.4%-12.9%) than for shape (R2: 0.12-0.52; RMSD%: 17.7-23.9%). We consistently observed improvements in model performance associated with models that have been developed with Mcomb. Scale was best predicted with LEAP (R2: 0.73; RMSD%: 8.4%), while shape was best predicted with GLMNET (R2: 0.52; RMSD%: 17.7%). For these

models, we observed a mean increase in R2 of 0.08 (SD: 0.03) and a mean decrease in RMSD% of 1.3% (SD: 0.6%) with models that were developed using Mcomb over those developed using Mals.

Analysis of the variable importance indicated that the Correlation metric from Mtex is holistically the most important predictor within Mcomb (**Figure 2-5**). The most important predictors thereafter are, for the majority, from Mals. In summary, we generally observed higher R2 and lower RMSD% to be associated with models that were developed with Mcomb compared with those using Mals or Mtex, regardless of the parameter being modelled or modelling technique being used. We found the variable Correlation, originating from Mtex, to be the most important predictor within Mcomb. Relative biases remained very low regardless of the parameter being modelled, the ALS metric set that was used, or the model-ling approach that was employed (min.: -8.8; max.: 9.2; mean: 1.0; SD: 2.7 in absolute values of Bias; data not shown). We observed no trend in the performance of the modelling techniques across all parameters.

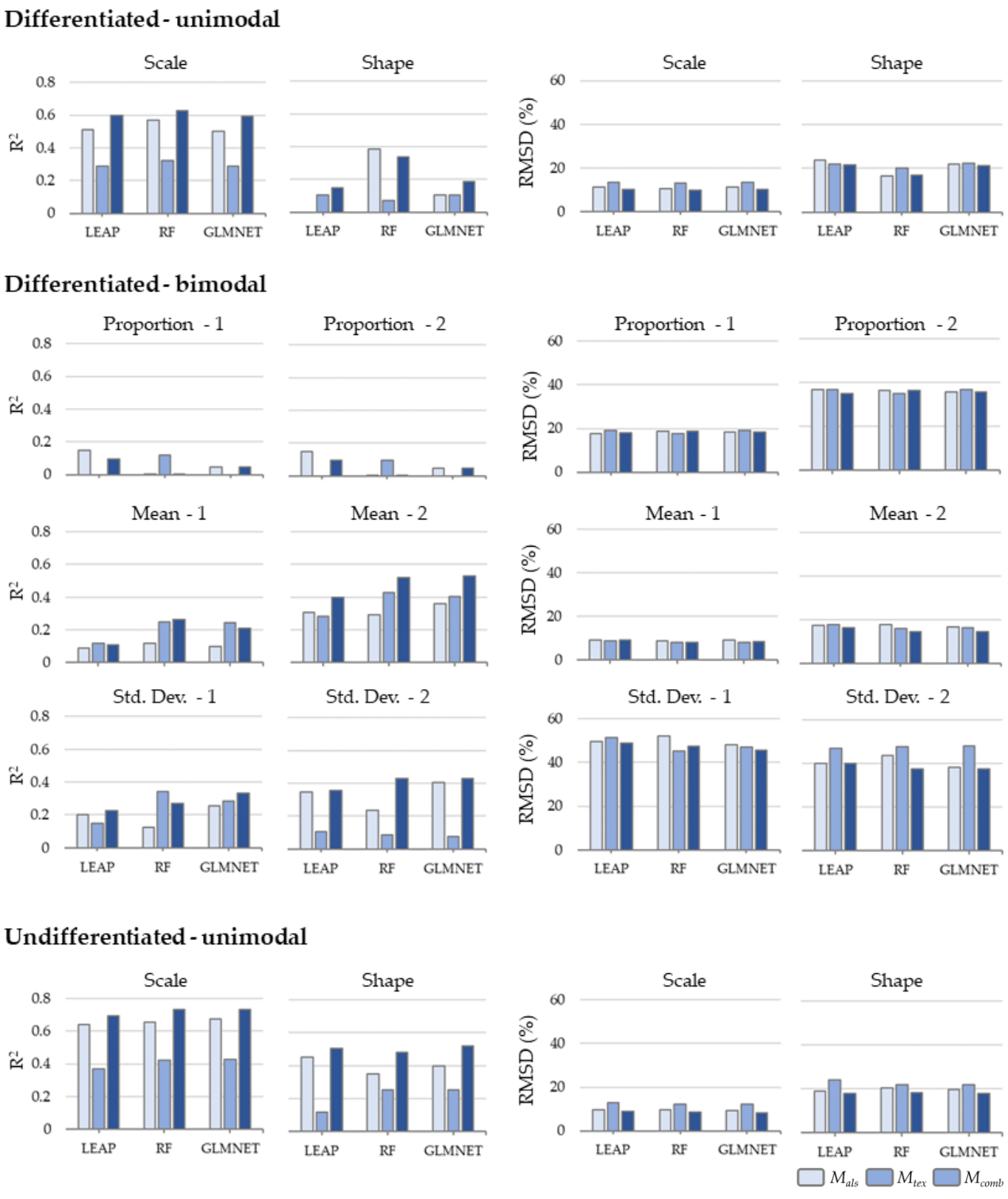

Figure 2-4: Coefficient of determination (R^2) and relative root-mean-squared deviation (RMSD%) that was derived from the application of the SDD prediction models to the test case data using the differentiated unimodal, differentiated bimodal, and undifferentiated SDD modality plot groupings; three ALS metrics sets (M_{als}, M_{tex}, M_{comb}) and three modelling techniques (RF, GLMNET, LEAP) were used.

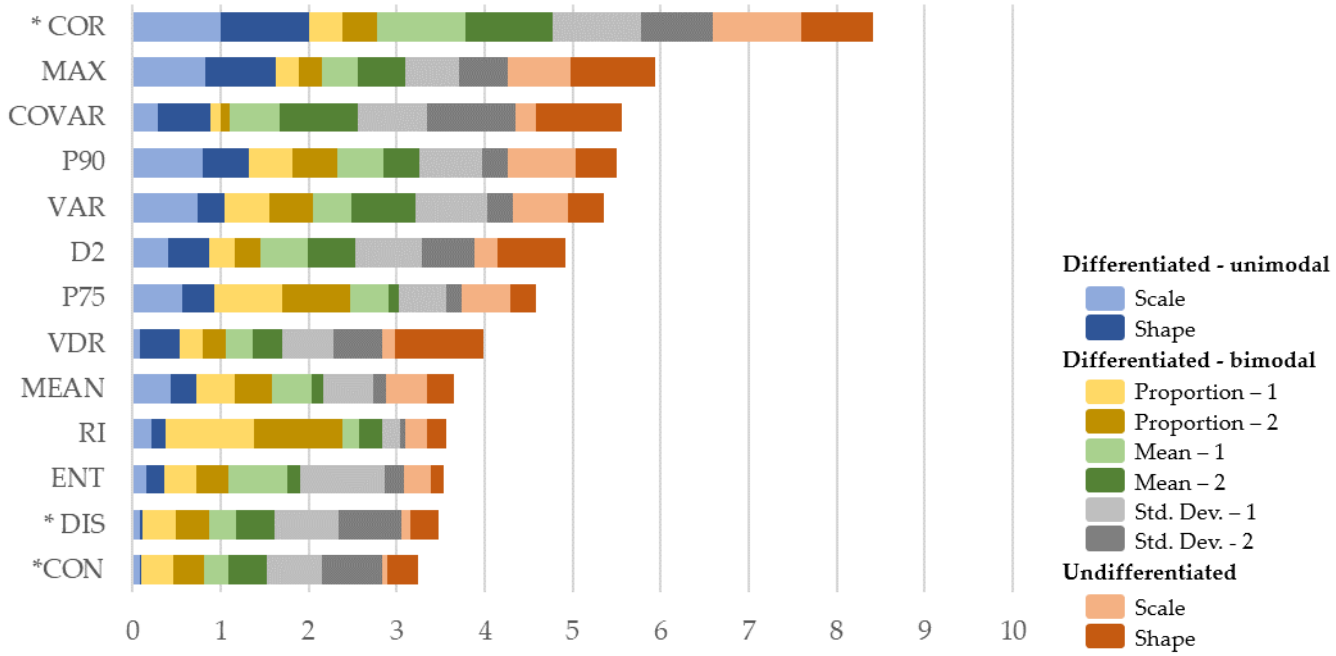

Figure 2-5: Cumulative variable importance values for metrics used in the best SDD parameter models which used M_{comb} during model development. Individual values represent the average variable importance across the three modelling techniques within each parameter and was scaled between 0 and 1. Only metrics with a cumulative value > 3 are shown. Asterix denotes metrics originating from M_{tex}.

2.3.3 Goodness-of-fit of the Predicted SDD

We applied the best model within each model set independently to each plot and calculated mean Error Indices (EIs) from the predicted SDD parameters for both the model development and test case datasets (**Table 2-3**). We observed the same trends from both datasets. Surprisingly, we observed an increase in EI by applying differentiated unimodal models on unimodal plots, albeit the increase is negligible (< 1%). Differentiating modality prior to estimating SDD most improved the accuracy of estimates for bimodal plots (~12% decrease in EI). Of the 120 plots that were used test our models, 50 (41.7%) had a better EI when derived from differentiated modality model predictions (31 and 19 plots within the differentiated unimodal and bimodal plots, respectively). Overall, we observed a marginally better fit (~4% decrease in EI) for SDD that were estimated from the differentiated modality model set in comparison with those estimated from the undifferentiated modality model set. The results therefore indicate improvements in SDD predictions by using differentiated modality-specific models, namely for heterogeneous (bimodal) stands.

Table 2-3: Plot-level Reynold's Error Index means for each ground plot dataset and model set. EI values ranged between 0 and 200, where an EI of 0 indicated a perfect fit between predicted and observed SDD, which an EI of 200 indicated a completely different SDD.

Plot dataset	n	Model set	
		Differentiated	Undifferentiated
Model development			
Differentiated as unimodal	215	50.4	50.3
Differentiated as bimodal	92	65	74
Undifferentiated modality	307	54.8	57.4
Test case			
Differentiated as unimodal	88	50.8	50.5
Differentiated as bimodal	32	59.1	67
Undifferentiated modality	120	53	54.9

2.4 Discussion

From our first hypothesis, we expected models that were developed with CHM texture metrics to outperform SDD prediction models developed solely with ALS metrics. This expectation was based upon previous studies that related CHM texture metrics (M_{tex}) to properties of the growing stock, such as the spatial pattern of trees [74], and furthermore, have demonstrated that their inclusion as predictors in modelling forest attributes improved predictions over using ALS metrics alone [38,40,41]. For example, van Ewijk et al. (2019) [38] tested multiple predictor sets using height metrics with combinations of CHM texture and intensity metrics and found that the addition of texture metrics improved prediction accuracies for basal area, quadratic mean DBH and stem density. To our knowledge, no published studies have directly assessed the contribution of CHM texture metrics in estimating SDD. Hence, the innovative aspects of our study make direct comparisons with past research challenging, especially regarding the attributes that we assessed (i.e., SDD modality and parameters), together with the CHM texture metrics that were included in our analyses. Nevertheless, our study demonstrated comparable results in classifying SDD modality with Zhang et al. (2019) [16] and Mulverhill et al. (2018) [17] using M_{als} (range in overall accuracies: 71%-73% vs. 49%-76% and 47%-78%, respectively). Our results for estimating SDD were generally comparable with those presented in Mulverhill et al. (2018) [17] for the differentiated unimodal distributions' modelled parameters, albeit with consistently lower error. Consistent with Thomas et al. (2008) [7] and Zhang et al. (2019) [16], the second component of the FMM that was associated with differentiated bimodal distributions was better predicted than the first. As highlighted by Thomas et al. (2008) [7], the main drawback of FMM is the increase in parameters that are needed to describe it. With the increase of modelled parameters, it becomes unlikely that each can be predicted accurately with M_{als}. Apart from the proportions associated to the FMM's components, the parameters of the differentiated bimodal distributions were best predicted with M_{comb}. Unlike Zhang et al. (2019) [16] and Mulverhill et al. (2018) [17], who had developed models solely from M_{als}, our best SDD prediction models were generally developed with M_{comb}. Therefore, we could confirm our first hypothesis given that our study demonstrated that SDD prediction models developed with M_{comb} usually outperformed those developed with M_{als} (Figure 4).

Our second hypothesis stated that developing differentiated modality-specific models (i.e., unimodal/bimodal) would improve SDD predictions for heterogeneous stands in our study site. The literature demonstrates improvements in estimating SDD with approaches that differentiate stand modality over approaches that do not (e.g., [16,17]). Our results indicated a similar trend. Yet, when interpreted globally, the improvements were marginal (~4 decrease in EI). Surprisingly, within our differentiated plot datasets, we observed that SDD was marginally better predicted by the

undifferentiated modality model set that was intended for unimodal plots. Notably, and in support of our hypothesis, we observed SDD to be better predicted by the differentiated bimodal model set for bimodal plots (mean EI of 59.1 vs. 67.0). Our results therefore support the idea that developing model sets based on the modality of stands can improve SDD predictions for bimodal stands. Given this, we can confirm our hypothesis that differentiating for modality prior to estimating SSD improved the accuracy of estimates for the bimodal SDD conifer stands of our study site.

Accurate differentiation of the SDD modalities was assumed in our analyses and therefore potential errors in differentiation would directly impact model performances. Of the multiple available approaches to differentiate SDD modalities, we implemented BC, as it has been successfully implemented in similar studies (e.g., [17]). Yet, it should be noted that BC is directly influenced by the kurtosis and, more so, by the skewness of a given distribution [21]. A distribution with high skewness and low kurtosis can inflate BC and subsequently differentiate the distribution as bimodal. Left-skewed distributions are observed when larger diameter trees dominate, while right-skewed distributions are associated with stands that are dominated by smaller diameter trees. Both situations will yield, however, a skewness value greater than zero. The closer that observed skewness is to zero, the more homogeneous the distribution will be: the stand can be described as having an even-aged distribution [29]. Freeman and Dale (2013) [18] evaluated the effect of the skewness, the proportion, and the distance between the modes on the BC value. In their study, BC produced 21% false positives where simulated unimodal distributions had BC values greater than the bimodality threshold of 5/9 and were subsequently classified as bimodal. The BC relies upon the basic assumption that bimodality involves an increase in distribution asymmetry; therefore, an increase in skewness within a unimodal context can increase the BC and produce misclassification. Furthermore, the BC is not calibrated to proportion size; a small proportion in either component of a bimodal distribution can also produce false positives, when the former is combined with a small distance between associated means. Of the 124 (92 in model development and 32 in test case) plots that were differentiated as bimodal in our study, 81 had skewness estimates > 1 and, thus, can be considered substantially skewed. Furthermore, the proportions that were associated with the second component of the bimodal distributions of our bimodal plots were low, as were distances between observed means (mean 5.8 cm). Given these results, it is possible that the combination of these factors could have inflated the BC and, therefore, mis-differentiated plots as bimodal. We can advance this as a plausible explanation, given the observed better fit for SDD that was estimated from the differentiated modality model set was minimal (decrease in RI ~4%). These effects upon the BC suggest that relying solely on this differentiation method may not be advisable for all forest types. Zhang et al. (2019) [16] used a combination of the Gini Coefficient and the asymmetry of the Lorenz curve to differentiate SDD modality, given that both measures are related to stand heterogeneity and the skewness of the diameter distribution [7,47]. Additional research is required to determine the optimal approach for differentiating the modality of SDD for a given forest type.

Nevertheless, the research presented here is important for several reasons. First, the methodology is used to differentiate the SDD modality and to develop the modality classification model, which can be used by foresters to improve differentiation of stand structure types and to select the most appropriate models for accurately estimating diameter distributions across large ALS coverages. Second, we demonstrated that models fitted with Mcomb yielded higher R2 and lower RMSD% in comparison with those using solely Mals, thereby indicating that textural metrics contain additional information useful for the estimation of SDD.

2.5 Conclusion

This study compared two approaches for estimating SDD, one in which the SDD probability function parameters were predicted with modality-specific models, and the other without. We compared model performances when fitted independently to three ALS metric sets (Mals, Mtex, Mcomb) to assess the contribution of CHM texture metrics: overall SDD were generally best estimated using a combination of ALS and texture metrics (Mcomb), thereby emphasizing the

additional information contained in CHM texture metrics. As expected, we confirmed that developing modality specific models improved SDD predictions for bimodal distributions, which surprisingly, was not the case for unimodal distributions. The performance of SDD predictions for the latter, as assessed through the EI, varied negligibly between modality-specific and undifferentiated- modality models. Of note, the performance of modality-specific models will depend on the modality differentiation method implemented, which must be assessed thoroughly. These results may provide for operational efficiencies in modelling and mapping SDD in this balsam fir or spruce dominated forest environments.

Author Contributions: Conceptualization, X.G.-D., O.R.L. and R.A.F.; methodology, X.G.-D., O.R.L. and R.A.F.; validation, X.G.-D.; formal analysis, X.G.-D.; investigation, X.G.-D., O.R.L.; resources, O.R.L. and R.A.F.; data curation, X.G.-D. and O.R.L.; writing—original draft preparation, X.G.-D.; writing—review and editing, O.R.L. and R.A.F.; visualization, X.G.-D.; supervision, R.A.F. and O.R.L.; project administration, R.A.F.; funding acquisition, R.A.F.

Funding: This work was supported by Natural Resources Canada's Canadian Forest Service – Canadian Wood Fibre Centre; and the Assessment of Wood Attributes using Remote Sensing Project (National Sciences and Engineering Research Council of Canada Collaborative Research and Development Grant PJ-462973-14, grantee N.C. Coops, UBC); in collaboration with Corner Brook Pulp and Paper Limited; and the Newfoundland and Labrador Department of Fisheries and Land Resources.

Data Availability Statement: The data underlying this article will be shared on reasonable request to the corresponding author.

Acknowledgments: This research was mainly developed in the Centre d'Applications et de Recherche en TÉLédétection of the Université de Sherbrooke, Canada. We thank Faron Knott and Kim Childs of Corner Brook Paper Limited for their input and assistance with the project.

Conflict of interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

2.6 References

1. Martin Bollandsås, O.; Næsset, E. Estimating Percentile-Based Diameter Distributions in Uneven-Sized Norway Spruce Stands Using Airborne Laser Scanner Data. *Scand. J. For. Res.* **2007**, *22*, 33–47, doi:10.1080/02827580601138264.
2. Maltamo, M.; Suvanto, A.; Packalén, P. Comparison of Basal Area and Stem Frequency Diameter Distribution Modelling Using Airborne Laser Scanner Data and Calibration Estimation. *For. Ecol. Manage.* **2007**, *247*, 26–34, doi:10.1016/j.foreco.2007.04.031.
3. Penner, M.; Woods, M.; Pitt, D. A Comparison of Airborne Laser Scanning and Image Point Cloud Derived Tree Size Class Distribution Models in Boreal Ontario. *Forests* **2015**, *6*, 4034–4054, doi:10.3390/f6114034.
4. Fries, C.; Johansson, O.; Pettersson, B.; Simonsson, P. Silvicultural Models to Maintain and Restore Natural Stand Structures in Swedish Boreal Forests. *For. Ecol. Manage.* **1997**, *94*, 89–103, doi:10.1016/S0378-1127(97)00003-0.
5. Packalén, P.; Maltamo, M. Estimation of Species-Specific Diameter Distributions Using Airborne Laser Scanning and Aerial Photographs. *Can. J. For. Res.* **2008**, *38*, 1750–1760, doi:10.1139/x08-037.

6. Knoebel, B.R.; Burkhart, H.E. A Bivariate Distribution Approach to Modeling Forest Diameter Distributions at Two Points in Time. *Biometrics* **1991**, *47*, 241–253, doi:10.2307/2532509.

7. Thomas, V.; Oliver, R.D.; Lim, K.; Woods, M. LiDAR and Weibull Modeling of Diameter and Basal Area. *For. Chron.* **2008**, *84*, 866–875, doi:10.5558/tfc84866-6.

8. White, J.C.; Coops, N.C.; Wulder, M.A.; Vastaranta, M.; Hilker, T.; Tompalski, P. Remote Sensing Technologies for Enhancing Forest Inventories: A Review. *Can. J. Remote Sens.* **2016**, *42*, 619–641, doi:10.1080/07038992.2016.1207484.

9. Ullah, S.; Dees, M.; Datta, P.; Adler, P.; Koch, B. Comparing Airborne Laser Scanning, and Image-Based Point Clouds by Semi-Global Matching and Enhanced Automatic Terrain Extraction to Estimate Forest Timber Volume. *Forests* **2017**, *8*, 215, doi:https://doi.org/10.3390/f8060215.

10. Noordermeer, L.; Bollandsås, O.M.; Ørka, H.O.; Næsset, E.; Gobakken, T. Comparing the Accuracies of Forest Attributes Predicted from Airborne Laser Scanning and Digital Aerial Photogrammetry in Operational Forest Inventories. *Remote Sens. Environ.* **2019**, *226*, 26–37, doi:10.1016/j.rse.2019.03.027.

11. Penner, M.; Pitt, D.G.; Woods, M.E. Parametric vs. Nonparametric LiDAR Models for Operational Forest Inventory in Boreal Ontario. *Can. J. Remote Sens.* **2013**, *39*, 426–443, doi:10.5589/m13-049.

12. Næsset, E. Area-Based Inventory in Norway – From Innovation to an Operational Reality. In *Forestry Applications of Airborne Laser Scanning*; Springer, 2014; Vol. 27, p. 460.

13. Junttila, V.; Kauranne, T.; Finley, A.O.; Bradford, J.B. Linear Models for Airborne-Laser-Scanning-Based Operational Forest Inventory With Small Field Sample Size and Highly Correlated LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 5600–5612, doi:10.1109/TGRS.2015.2425916.

14. Maltamo, M.; Gobakken, T. Predicting Tree Diameter Distributions. In *Forestry Applications of Airborne Laser Scanning*; Springer, 2014; pp. 177–191.

15. White, J.C.; Tompalski, P.; Vastaranta, M.; Wulder, M.A.; Saarinen, N.; Stepper, C.; Coops, N.C. *A Model Development and Application Guide for Generating an Enhanced Forest Inventory Using Airborne Laser Scanning Data and an Area-Based Approach. Information Report FI-X-018*; Natural Resources Canada, Canadian Forest Service, Canadian Wood Fibre Center, Victoria, BC, Canada, 2017;

16. Zhang, Z.; Cao, L.; Mulverhill, C.; Liu, H.; Pang, Y.; Li, Z. Prediction of Diameter Distributions with Multimodal Models Using LiDAR Data in Subtropical Planted Forests. *Forests* **2019**, *10*, 125, doi:10.3390/f10020125.

17. Mulverhill, C.; Coops, N.C.; White, J.C.; Tompalski, P.; Marshall, P.L.; Bailey, T. Enhancing the Estimation of Stem-Size Distributions for Unimodal and Bimodal Stands in a Boreal Mixedwood Forest with Airborne Laser Scanning Data. *Forests* **2018**, *9*, 95, doi:10.3390/f9020095.

18. Freeman, J.B.; Dale, R. Assessing Bimodality to Detect the Presence of a Dual Cognitive Process. *Behav. Res. Methods* **2013**, *45*, 83–97, doi:10.3758/s13428-012-0225-x.

19. Ellison, A.M. Effect of Seed Dimorphism on the Density-Dependent Dynamics of Experimental Populations of Atriplex Triangularis (Chenopodiaceae). *Am. J. Bot.* **1987**, *74*, 1280–1288, doi:10.2307/2444163.

20. Hartigan, J.A.; Hartigan, P.M. The Dip Test of Unimodality. *Ann. Stat.* **1985**, *13*, 14, doi:10.1214/aos/1176346577.

21. Pfister, R.; Schwarz, K.A.; Janczyk, M.; Dale, R.; Freeman, J. Good Things Peak in Pairs: A Note on the Bimodality Coefficient. *Front. Psychol.* **2013**, *4*, 700, doi:https://doi.org/10.3389/fpsyg.2013.00700.

22. Maltamo, M.; Malinen, J.; Kangas, A.; Härkönen, S.; Pasanen, A.M. Most Similar Neighbour-Based Stand Variable Estimation for Use in Inventory by Compartments in Finland. *Forestry* **2003**, *76*, 449–463, doi:10.1093/forestry/76.4.449.

23. Rana, P.; Vauhkonen, J.; Junttila, V.; Hou, Z.; Gautam, B.; Cawkwell, F.; Tokola, T. Large Tree Diameter Distribution Modelling Using Sparse Airborne Laser Scanning Data in a Subtropical Forest in Nepal. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 86–95, doi:10.1016/J.ISPRSJPRS.2017.10.018.

24. Spriggs, R.A.; Coomes, D.A.; Jones, T.A.; Caspersen, J.P.; Vanderwel, M.C. An Alternative Approach to Using LiDAR Remote Sensing Data to Predict Stem Diameter Distributions across a Temperate Forest Landscape. *Remote Sens.* **2017**, *9*, doi:10.3390/rs9090944.

25. Haara, A.; Maltamo, M.; Tokola, T. The K-Nearest-Neighbour Method for Estimating Basal-Area Diameter Distribution. *Scand. J. For. Res.* **1997**, *12*, 200–208, doi:10.1080/02827589709355401.

26. Strunk, J.L.; Gould, P.J.; Packalen, P.; Poudel, K.P.; Andersen, H.-E.E.; Temesgen, H. An Examination of Diameter Density Prediction with K-NN and Airborne Lidar. *Forests* **2017**, *8*, 444, doi:10.3390/f8110444.

27. Poudel, K.P.; Cao, Q. V. Evaluation of Methods to Predict Weibull Parameters for Characterizing Diameter Distributions. *For. Sci.* **2013**, *59*, 243–252, doi:10.5849/FORSCI.12-001.

28. Bailey, R.L.; Dell, T.R. Quantifying Diameter Distributions with the Weibull Function. *For. Sci.* **1973**, *19*, 97–104, doi:10.1093/forestscience/19.2.97.

29. Zhang, L.; Liu, C. Fitting Irregular Diameter Distributions of Forest Stands by Weibull, Modified Weibull, and Mixture Weibull Models. *J. For. Res.* **2006**, *11*, 369–372, doi:10.1007/s10310-006-0218-7.

30. Tompalski, P.; Coops, N.C.; White, J.C.; Wulder, M.A. Enriching ALS-Derived Area-Based Estimates of Volume through Tree-Level Downscaling. *Forests* **2015**, *6*, 2608–2630, doi:10.3390/f6082608.

31. Tarp-Johansen, M.J. Stem Diameter Estimation from Aerial Photographs. *Scand. J. For. Res.* **2002**, *17*, 369–376, doi:10.1080/02827580260138116.

32. Gobakken, T.; Næsset, E. Estimation of Diameter and Basal Area Distributions in Coniferous Forest by Means of Airborne Laser Scanner Data. *Scand. J. For. Res.* **2004**, *19*, 529–542, doi:10.1080/02827580410019454.

33. Peuhkurinen, J.; Tokola, T.; Plevak, K.; Sirparanta, S.; Kedrov, A.; Pyankov, S. Predicting Tree Diameter Distributions from Airborne Laser Scanning, SPOT 5 Satellite, and Field Sample Data in the Perm Region, Russia. *Forests* **2018**, *9*, 639, doi:10.3390/f9100639.

34. Shang, C.; Treitz, P.; Caspersen, J.; Jones, T. Estimating Stem Diameter Distributions in a Management Context for a Tolerant Hardwood Forest Using ALS Height and Intensity Data. *Can. J. Remote Sens.* **2017**, *43*, 79–94, doi:10.1080/07038992.2017.1263152.

35. Hou, Z.; Xu, Q.; Tokola, T. Use of ALS, Airborne CIR and ALOS AVNIR-2 Data for Estimating Tropical Forest Attributes in Lao PDR. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 776–786, doi:10.1016/j.isprsjprs.2011.09.005.

36. Haralick, R.M.; Shanmugam, K.; others Textural Features for Image Classification. *IEEE Trans. Syst. Man. Cybern.* **1973**, 610–621, doi:10.1109/TSMC.1973.4309314.

37. TUCERYAN, M.; JAIN, A.K. TEXTURE ANALYSIS. In *Handbook of Pattern Recognition and Computer Vision*; Chen, C.H., Pau, L.F., Wang, P.S.P., Eds.; WORLD SCIENTIFIC: Singapore, Singapore, 1999; pp. 207–248.

38. van Ewijk, K.; Treitz, P.; Woods, M.; Jones, T.; Caspersen, J. Forest Site and Type Variability in ALS-Based Forest

Resource Inventory Attribute Predictions over Three Ontario Forest Sites. *Forests* **2019**, *10*, 226, doi:10.3390/f10030226.

39. Dube, T.; Mutanga, O.; Abdel-Rahman, E.M.; Ismail, R.; Slotow, R. Predicting Eucalyptus Spp. Stand Volume in Zululand, South Africa: An Analysis Using a Stochastic Gradient Boosting Regression Ensemble with Multi-Source Data Sets. *Int. J. Remote Sens.* **2015**, *36*, 3751–3772, doi:10.1080/01431161.2015.1070316.

40. Ozdemir, I.; Donoghue, D.N.M. Modelling Tree Size Diversity from Airborne Laser Scanning Using Canopy Height Models with Image Texture Measures. *For. Ecol. Manage.* **2013**, *295*, 28–37, doi:10.1016/j.foreco.2012.12.044.

41. Niemi, M.T.; Vauhkonen, J.; Shan, J.; Hyyppä, J.; Baghdadi, N.; Thenkabail, P.S. Extracting Canopy Surface Texture from Airborne Laser Scanning Data for the Supervised and Unsupervised Prediction of Area-Based Forest Characteristics. *Remote Sens.* **2016**, *8*, 582, doi:10.3390/rs8070582.

42. Rowe, J.S. *Forest Regions of Canada. Based on W. E. D. Halliday's "A Forest Classification for Canada"*; PublicationNo 1300; Department of the Environment, Canadian Forestry Service: Ottawa, ON, Canada, 1972;

43. Torgo, L. *Data Mining with R: Learning with Case Studies.*; Second Edi.; Chapman and Hall/CRC, 2017; ISBN 978-1482234893.

44. Roussel, J.-R.; Auty, D. LidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications Available online: https://cran.r-project.org/web/packages/lidR/index.html.

45. R Core Team R: A Language and Environment for Statistical Computing Available online: https://www.r-project.org/.

46. Hall-Beyer, M. Practical Guidelines for Choosing GLCM Textures to Use in Landscape Classification Tasks over a Range of Moderate Spatial Scales. *Int. J. Remote Sens.* **2017**, *38*, 1312–1338, doi:10.1080/01431161.2016.1278314.

47. Bouvier, M.; Durrieu, S.; Fournier, R.A.; Renaud, J.P. Generalizing Predictive Models of Forest Inventory Attributes Using an Area-Based Approach with Airborne LiDAR Data. *Remote Sens. Environ.* **2015**, *156*, 322–334, doi:10.1016/j.rse.2014.10.004.

48. Peduzzi, A.; Wynne, R.H.; Fox, T.R.; Nelson, R.F.; Thomas, V.A. Estimating Leaf Area Index in Intensively Managed Pine Plantations Using Airborne Laser Scanner Data. *For. Ecol. Manage.* **2012**, *270*, 54–65, doi:10.1016/j.foreco.2011.12.048.

49. Pope, G.; Treitz, P. Leaf Area Index (LAI) Estimation in Boreal Mixedwood Forest of Ontario, Canada Using Light Detection and Ranging (LiDAR) and Worldview-2 Imagery. *Remote Sens.* **2013**, *5*, 5040–5063, doi:10.3390/rs5105040.

50. Goetz, S.; Steinberg, D.; Dubayah, R.; Blair, B. Laser Remote Sensing of Canopy Habitat Heterogeneity as a Predictor of Bird Species Richness in an Eastern Temperate Forest, USA. *Remote Sens. Environ.* **2007**, *108*, 254–263, doi:10.1016/j.rse.2006.11.016.

51. van Ewijk, K.Y.; Treitz, P.M.; Scott, N.A. Characterizing Forest Succession in Central Ontario Using Lidar-Derived Indices. *Photogramm. Eng. Remote Sensing* **2011**, *77*, 261–269, doi:10.14358/PERS.77.3.261.

52. Pretzsch, H. Description and Analysis of Stand Structures. In *Forest Dynamics, Growth and Yield.*; Springer, Berlin, Heidelberg, 2010 ISBN 9783540883067.

53. Jenness, J.S. Calculating Landscape Surface Area from Digital Elevation Models. *Wildl. Soc. Bull.* **2004**, *32*, 829–

839, doi:10.2193/0091-7648(2004)032[0829:clsafd]2.0.co;2.

54. Woods, M.; Pitt, D.; Penner, M.; Lim, K.; Nesbitt, D.; Etheridge, D.; Treitz, P. Operational Implementation of a LiDAR Inventory in Boreal Ontario. *For. Chron.* **2011**, *87*, 512–528, doi:10.5558/tfc2011-050.

55. Beets, P.N.; Reutebuch, S.; Kimberley, M.O.; Oliver, G.R.; Pearce, S.H.; McGaughey, R.J. Leaf Area Index, Biomass Carbon and Growth Rate of Radiata Pine Genetic Types and Relationships with LiDAR. *Forests* **2011**, *2*, 637–659, doi:10.3390/f2030637.

56. Solberg, S.; Brunner, A.; Hanssen, K.H.; Lange, H.; Næsset, E.; Rautiainen, M.; Stenberg, P. Mapping LAI in a Norway Spruce Forest Using Airborne Laser Scanning. *Remote Sens. Environ.* **2009**, *113*, 2317–2327, doi:10.1016/j.rse.2009.06.010.

57. Hopkinson, C.; Chasmer, L. Testing LiDAR Models of Fractional Cover across Multiple Forest Ecozones. *Remote Sens. Environ.* **2009**, *113*, 275–288, doi:10.1016/j.rse.2008.09.012.

58. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320, doi:10.1111/j.1467-9868.2005.00527.x.

59. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Software, Artic.* **2008**, *28*, 1–26, doi:10.18637/jss.v028.i05.

60. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18–22.

61. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*; Fourth.; Springer: New York, 2002;

62. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. Kernlab -- An {S4} Package for Kernel Methods in {R}. *J. Stat. Softw.* **2004**, *11*, 1–20, doi:10.18637/jss.v011.i09.

63. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1, doi:10.18637/jss.v033.i01.

64. Hastie, T.; Tibshirani, R.; Friedman, J. *Springer Series in Statistics The Elements of Statistical Learning*; 2nd ed.; Springer, 2009;

65. Delignette-Muller, M.L.; Dutang, C. {fitdistrplus}: An {R} Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*, 1–34, doi:10.18637/jss.v064.i04.

66. Yu, Y. MixR: An R Package for Finite Mixture Modeling for Both Raw and Binned Data. *J. Open Source Softw.* **2022**, *7*, 4031, doi:10.21105/joss.04031.

67. Furnival, G.M.; Wilson, R.W. Regressions by Leaps and Bounds. *Technometrics* **1974**, *16*, 499–511, doi:10.2307/1271435.

68. Lumley, T. Leaps: Regression Subset Selection, Based on Fortran Code by Alan Miller, r Package Version 3.1 2020.

69. Land, A.H.; Doig, A.G. An Automatic Method of Solving Discrete Programming Problems. *Econometrica* **1960**, *28*, 497, doi:10.2307/2223855.

70. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M.C. Modeling Spatial Patterns of Fire Occurrence in Mediterranean Europe Using Multiple Regression and Random Forest. *For. Ecol. Manage.* **2012**, *275*, 117–129, doi:10.1016/j.foreco.2012.03.003.

71. Kuhn, M. Caret: Classification and Regression Training 2022.

72. Reynolds, M.R.; Burk, T.E.; Huang, W.-C. Goodness-of-Fit Tests and Model Selection Procedures for Diameter Distribution Models. *For. Sci.* **1988**, *34*, 373–399, doi:10.1093/forestscience/34.2.373.

73. Coomes, D.A.; Duncan, R.P.; Allen, R.B.; Truscott, J. Disturbances Prevent Stem Size-Density Distributions in Natural Forests from Following Scaling Relationships. *Ecol. Lett.* **2003**, *6*, 980–989, doi:10.1046/j.1461-0248.2003.00520.x.

74. Pippuri, I.; Kallio, E.; Maltamo, M.; Peltola, H.; Packalén, P. Exploring Horizontal Area-Based Metrics to Discriminate the Spatial Pattern of Trees and Need for First Thinning Using Airborne Laser Scanning. *Forestry* **2012**, *85*, 305–314, doi:10.1093/forestry/cps005.

2.7 Supplementary Material

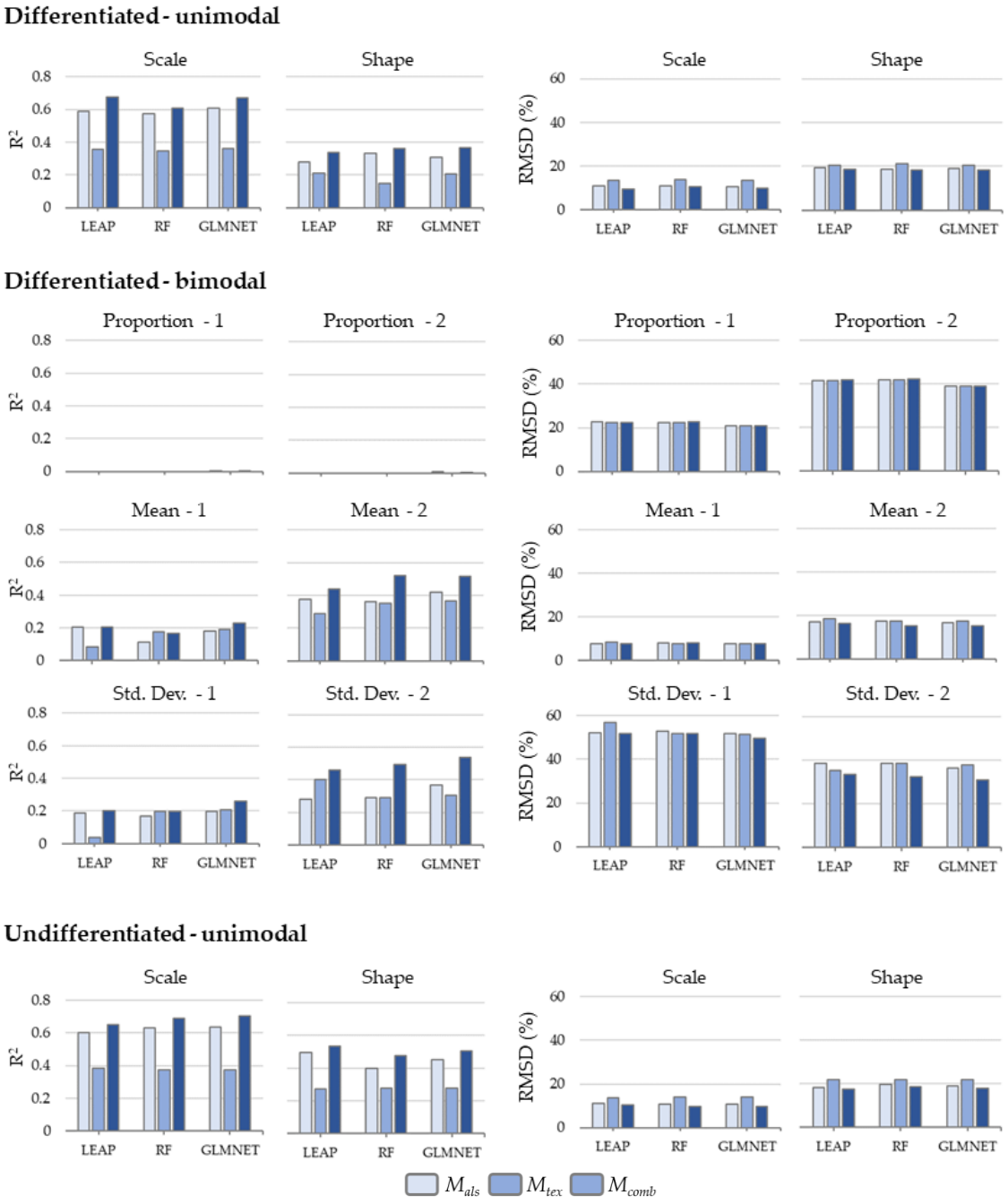

Figure S1: Average of 5 repeated cross-validation performance measures (R^2, RMSD%) derived during model development using the various SDD modality plot groupings, ALS metrics sets and modelling techniques.

3 Analyses complémentaires

3.1 Paramètres des DDA

Un aspect qui n'a pas été abordé dans l'article est la distribution des valeurs des paramètres qui composent les DDA représentées par une seule Weibull. Le Tableau 3-1 comprend les valeurs minimales, moyennes et maximales des paramètres des DDA pour les trois jeux de données utilisés dans l'article (placettes différentiées unimodale et multimodales, et toutes les placettes sans différenciation). Les valeurs des paramètres *shape* et *scale* des placettes unimodales et sans différentiations sont relativement similaires. Il est possible d'observer que les valeurs sont identiques pour le paramètre *scale* avec ou sans différentiation de la modalité (placettes unimodales vs sans différentiation). Donc, les placettes bimodales ont obtenu des valeurs de paramètre *scale* comprises entre 10.770 et 27.700 lorsque leur DDA était représenté par une Weibull unimodale. Au contraire, la valeur minimale du paramètre *shape* est plus basse pour les données sans différentiation que pour les unimodales. Ceci indique qu'une placette classifiée bimodale par le BC a obtenue une valeur de *shape* plus base qu'une placette unimodale. Le paramètre *shape* contrôle la pente de la fonction Weibull. Plus la valeur du *shape* est faible, plus le mode est étroit et près de 0. Ainsi, un paramètre faible de *shape* tend à produire une distribution qui présente un *skweness* à droite de la distribution. Par ailleurs, la valeur moyenne du paramètre *shape* sans différentiation est plus basse que pour les données différentiées unimodale.

Le paramètre *scale* représente l'étendue de la variabilité de la distribution. En conséquent, une valeur de *scale* faible représente un intervalle de valeur possible étroit. Au contraire, une valeur élevée augmente l'étendue des valeurs possibles et réduit du même coup la hauteur du mode. Dans le cas des données sans différentiation, la moyenne est plus faible due à l'ajout des placettes différentiées bimodales ayant une étendue de valeurs inférieure. La combinaison de *shape* et *scale* ayant de faibles valeurs génère des distributions qui ont un mode présent à gauche de la distribution, donc un *skewness* positif.

Tableau 3-1: Valeurs minimales, moyennes et maximales pour chaque paramètre des DDA extraites pour les jeux de données différentiés unimodale et bimodale ainsi que les DDA sans différentiation.

Jeu de données	Paramètre	Minimum	Moyenne	Maximum
Unimodal	*shape*	2.460	4.425	11.742
	scale	10.770	16.590	27.700
Bimodale	*prop1*	0.358	0.654	0.958
	prop2	0.042	0.346	0.642
	scale1	9.991	11.628	16.444
	scale2	12.390	18.200	29.000
	shape1	4.098	26.042	160.001
	shape2	2.909	8.175	160.001
	mean1	9.739	11.065	15.042
	mean2	11.990	16.870	27.150
	sd1	0.079	1.356	3.943
	sd2	0.111	3.463	8.215
Sans Différentiation	*shape*	2.387	4.365	11.742
	scale	10.770	15.940	27.700

3.2 Tests complémentaires pour la différentiation de la modalité

Comme mentionné dans le manuscrit, l'ensemble des résultats repose sur la qualité de la différentiation des modalités. Les articles portant sur le développement de spécifiques pour les DDA unimodales et bimodales utilisaient le BC (Mulverhill *et al*., 2018; Zhang *et al*., 2019), car il s'agit d'une approche simple et rapide. Cependant, cette méthode n'était pas adaptée pour les peuplements dominés par le sapin baumier de l'ouest de Terre-Neuve qui présentent de fortes asymétries. Cependant, il existe d'autres méthodes de différentiation de la modalité. Nous avons testé d'autres méthodes pour déterminer si les erreurs de différentiation de la modalité provenaient de la méthode ou des données. Ces trois autres méthodes sont : le Hartigan-Hartigan Dip test, les caractéristiques de la courbe de Lorenz et les *gaussian mixture models* (GMM). Il est important de préciser que ces méthodes ont été uniquement testées à l'aide des données du site de Terre-Neuve. Les données du Québec n'étaient pas disponibles lors du choix final de la méthodologie.

3.2.1 Hartigan-Hartigan *Dip test*

La première technique testée était le test de Hartigan-Hartigan Dip (Hartigan and Hartigan, 1985). À la place d'une valeur seuil comme avec le BC, le Dip utilise les valeurs de p (*p-values*) pour déterminer si la distribution se présente sous forme multimodale ou non. Il est possible de tester différentes valeurs de p selon le niveau de confiance voulu (Cheng and Hall, 1999). Ce test calcule le Dip, soit la distance entre la distribution observée et une distribution théorique, et cherche la distribution de référence unimodale la plus près des données observées. Le Dip est calculé en mesurant la distance maximale entre les distributions cumulatives (Johnsson *et al*., 2017). La valeur du Dip est testée à différentes valeurs de p. Dans le cas de ce projet, les seuils testés étaient 0.05 et 0.10. Même si ce test grandement utilisé pour vérifier la bimodalité d'une distribution (Freeman and Dale, 2013; Mulverhill *et al*., 2018; Zhang *et al*., 2019), il a été originalement développé pour explorer la divergence d'une distribution unimodale, qui tend vers la présence de 2 modes (SAS Institute Inc., 1990). Il teste l'éloignement d'une *probability density function* (PDF) unimodale observée d'une PDF unimodale de référence. Cette PDF de référence contient un seul point d'inflexion entre un segment convexe et concave. Cette caractéristique du Dip test fait en sorte qu'il est plus robuste à la présence de distribution avec des valeurs de *skewness* s'éloignant de 0. Cependant, le Dip test a classé seulement 5 placettes du jeu de données de Terre-Neuve comme étant bimodales. Comme mentionné précédemment, certaines DDA ont présenté des discontinuités dans les mesures de DHP, surtout dans les valeurs localisées au milieu. Cette discontinuité a peut-être augmenté la sévérité du Dip test et classé toutes les placettes comme unimodales.

3.2.2 *Gaussian Mixture Models*

La troisième technique utilisait les *Gaussian Mixture Models* (GMM), qui permettaient d'estimer la distribution d'une variable aléatoire en modélisant la somme de plusieurs courbes gaussiennes. Si une variable aléatoire, comme les mesures de DHP, est mieux représentée par la somme de plusieurs courbes gaussiennes, c'est qu'elle comporte alors plus d'un mode. Au contraire, une DDA mieux représentée par

une seule courbe gaussienne est considérée comme unimodale. L'implémentation des GMM s'effectuait dans R à l'aide du package mclust (Scrucca *et al*., 2016) et le nombre de gaussiennes possible était limité à 1 ou 2. L'algorithme, *Expectation Maximization* (EM), teste ces deux possibilités et détermine la meilleure solution pour chacune des placettes (Nasrabadi, 2007). Si une somme de 2 gaussiennes représente mieux une DDA, alors celle-ci est considérée bimodale.

L'approche avec les GMM fournit des résultats totalement différents des autres méthodes, y compris le BC. Les GMM identifiaient 38 placettes comme unimodales et 23 comme bimodales. Le principal désavantage de la méthode avec DMM est lié au fait qu'elle repose sur l'ajustement d'une ou de plusieurs gaussiennes pour simuler la distribution observée. Dans le cas d'une distribution ayant une forte densité d'observation à la même place, l'algorithme ajustera une gaussienne très étroite. Si toutes les observations sont captées par cette gaussienne, alors l'algorithme identifiera uniquement un seul mode. Cependant, l'algorithme ne peut pas capter toutes les valeurs avec une seule gaussienne lorsque la distribution contient à la fois une forte densité (un pic) et une valeur de *skewness* élevée. En pareil cas, il créera une deuxième gaussienne pour capter les éléments présents dans la queue causée par la valeur élevée de *skewness*. La Figure 3-1 comprend un exemple de classification d'une DDA comme bimodale, alors que le BC était unimodale. La première gaussienne ne pouvant couvrir l'ensemble des mesures, l'algorithme ajuste une deuxième gaussienne sur les mesures restantes.

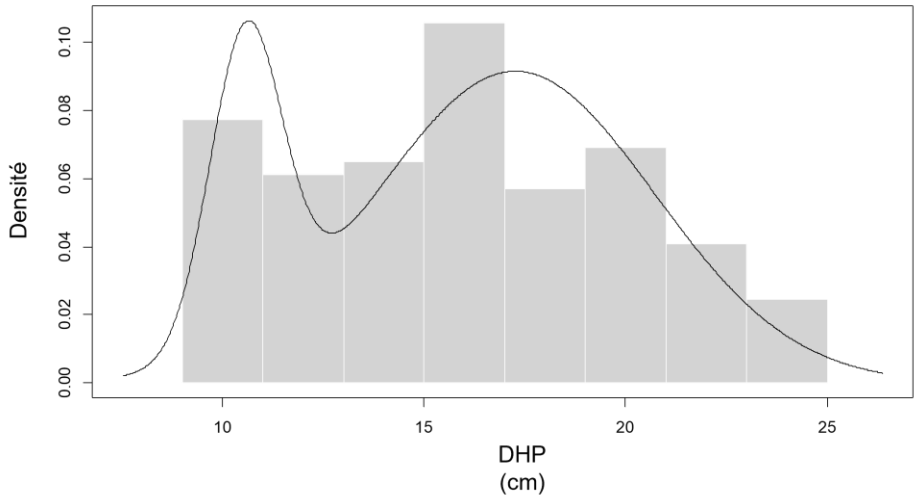

Figure 3-1: Exemple d'un *Gaussian Mixture Model* (GMM) ajusté sur une distribution des mesures de diamètre hauteur poitrine (DHP) d'une distribution des diamètres de arbres (DDA). L'histogramme comprend des intervalles de 2 cm et le GMM a ajusté deux gaussienne, ce qui signifie que la DDA est classifiée comme bimodale.

Comme mentionné dans l'article, plusieurs placettes présentaient une forte concentration d'arbres à petit DHP, ce qui crée une plus grande densité au début de la distribution. L'algorithme n'est donc pas en mesure d'ajuster une gaussienne sur l'entièreté des mesures. Donc, la combinaison d'une forte densité de petites valeurs de DHP à une valeur plus élevée de *skewness* crée la combinaison parfaite pour que l'algorithme de GMM détecte plus d'un mode et classifie la DDA comme bimodale par erreur.

3.2.3 Choix du test de classification de modalité

À la lumière de ces tests, deux constats sont possibles. Le premier est qu'il s'avère primordial de tester plus d'une méthode de différentiation des modalités pour s'assurer qu'elles produisent des résultats similaires, et ce, afin de développer des modèles spécifiques selon la modalité. Le deuxième constat est que deux des quatre techniques testées, le Hartigan-Hartigan dip test et les indicateurs de la courbe de Lorenz, présentaient des résultats identiques. Selon ces résultats, aucune DDA était bimodale. Les deux autres techniques, BC et GMM, permettaient de confirmer que la présence de *skewness* dans les DDA pouvait créer des faux-positifs (Freeman and

Dale, 2013; Pfister *et al.*, 2013) et ainsi sélectionner une approche de modélisation des DDA avec des modèles spécifiques pour les deux modalités possibles (unimodale et bimodale). Comparativement à la courbe de Lorenz, le BC permet une implémentation rapide de la classification des DDA en se basant uniquement sur les mesures de DHP. Au contraire, la courbe de Lorenz nécessite la densité des tiges, mais également la surface terrière. De plus, le BC a fourni des bons résultats pour des forêts boréales mixtes dominées par les feuillus (Mulverhill *et al*., 2018). En revanche, la courbe de Lorenz a servi dans une étude similaire sur une plantation d'eucalyptus (Zhang *et al*., 2019). Selon cette observation, la méthode du BC a permis une implémentation rapide et comparable avec des études utilisant une approche méthodologique similaire de prédiction de la DDA.

3.2.4 Caractéristiques de la courbe de Lorenz

La deuxième technique testée était celle utilisée par Zhang *et al.* (2019), soit calculer deux statistiques descriptives de la courbe de Lorenz. Dans un peuplement forestier, la courbe de Lorenz représente la relation des arbres dominants en comparant les proportions cumulatives relatives de la surface terrière et la densité des tiges pour chacun des arbres (Valbuena *et al.*, 2012). Par ailleurs, certaines caractéristiques de la courbe de Lorenz, comme l'asymétrie ainsi que l'indice de Gini, représentent des indicateurs de la structure de la forêt. Ces deux indicateurs ont déjà été utilisés pour caractériser les inégalités de la taille des arbres pour stratifier la forêt en peuplement homogène et hétérogène. Suivant la méthodologie proposée par Zhang *et al*. (2019), nous avons calculé pour chacune des placettes la courbe de Lorenz, l'asymétrie et l'indice de Gini. Si la valeur de l'asymétrie de la courbe de Lorenz < 0.5 et que l'indice de Gini > 0.5, la DDA était alors considérée bimodale. Ces deux indicateurs étaient calculés à l'aide du package R ineq (Zeileis, 2014).

La méthode de la courbe de Lorenz n'a pas différentié une placette comme étant bimodale. Contrairement à Zhang *et al*. (2019), pour qui cette méthode était la meilleure pour différentier les modalités, nous n'avons pas obtenu de valeurs d'indice de Gini et d'asymétrie qui caractérisent généralement les peuplements hétérogènes et

bimodaux. Des valeurs d'indices de Gini < 0.5 suggèrent des peuplements unimodaux. Par contre, les valeurs seuils de 0.5 pour l'asymétrie et l'indice de Gini ont été déterminées par Valbuena *et al.* (2012) sur des forêts de mono-espèce de pins sylvestres, même si les auteurs mentionnaient que cette propriété de l'indice de Gini est indépendante de l'espèce.

3.3 Modélisations spécifiques aux espèces dominantes

Tel que supporté par différentes études (ex: Thomas *et al*., 2008; Zhang *et al*., 2019), plus précise avec des modèles développés spécifiquement pour l'espèce ou par l'application de d'une stratification des essences dominantes. Bien que les détails de cette analyse ne figurent pas dans l'article, nous avons mis en place différents jeux de données pour tester cette hypothèse. Les espèces dominantes dans les deux sites étaient le sapin baumier et l'épinette noire. Pour ce faire, les placettes dominées par l'épinette noire du Québec et de Terre-Neuve ont été combinées, de même que pour les placettes dominées par le sapin baumier. Cette combinaison a créé trois jeux de données : QcNL_EPN, QcNL_SAB et QcNL_Comb. Le jeu de données QcNL_Comb comprend l'ensemble des placettes (QcNL_EPN + QcNL_SAB) et correspond aux données utilisées dans le manuscrit. Cependant, les données QcNL_EPN et QcNL_SAB ont également été utilisées pour prédire la DDA à l'aide de modèles spécifiques à l'espèce. Les mêmes étapes méthodologiques définies dans le manuscrit ont permis de prédire la DDA à partir de chacune de ces trois bases de données.

3.3.1 Différentiation de la modalité des DDA

Tout d'abord, les différents groupes de métriques ont servi à développer des modèles prédictifs pour différentier la modalité de la DDA en unimodale et bimodale. La mesure d'évaluation des performances est la précision globale (Tableau 3-2). À première vue, les performances entre les espèces spécifiques sont similaires. Les moyennes des précisions globales par espèces se situent à 0.70 pour QcNL_EPN, 0.70 pour QcNL_SAB et 0.71 pour QcNL_Comb. Au niveau des groupes de métriques QcNL_SAB présentte des meilleurs résultats avec les groupes M_{tex} ou M_{comb} uniquement. Cependant, QcNL_EPN et

QcNL_Comb ont obtenu des performances similaires, indépendamment des groupes de métriques.

Tableau 3-2: Précision globale des modèles de classification de la forme générale de la DDA (unimodale ou bimodale) pour les placettes du Québec et de Terre-Neuve dominées par l'épinette noire (QcNL_EPN), le sapin baumier (QcNL_SAB) et la combinaison des essences (QcNL_Comb).

	GLMNET			RF			SVM			Logit		
	M_{als}	M_{tex}	M_{comb}	M_{stand}	M_{tex}	M_{comb}	M_{stand}	M_{tex}	M_{comb}	M_{stand}	M_{tex}	M_{comb}
QcNL_EPN	0,71	0,68	0,67	0,68	0,66	0,71	0,68	0,69	0,66	0,71	0,77	0,74
QcNL_SAB	0,64	0,72	0,72	0,66	0,74	0,70	0,66	0,66	0,72	0,68	0,72	0,70
QcNL_Comb	0,71	0,73	0,71	0,72	0,66	0,71	0,73	0,72	0,71	0,72	0,74	0,72

3.3.2 Prédiction des paramètres des DDA unimodales.

Pour la prédiction des paramètres des DDA unimodales, certains modèles obtenaient des R^2 négatifs. Ce résultat signifie que la moyenne de la variable réponse performe mieux que les modèles. À des fins de simplicité d'affichage, les valeurs négatives sont remplacées par des zéro. À première vue, le paramètre *shape* obtient des R^2 plus faibles que le paramètre *scale* pour les trois groupes d'essences. La moyennes des paramètres *shape* et *scale* pour les trois groupes d'espèces se situaient respectivement à 0,20, 0.36 et 0.18. Les groupes QcNL_EPN et QcNL_Comb ont obtenu les valeurs les plus faibles de R^2. Cependant, QcNL_EPN a obtenu trois valeurs négatives de R^2 alors que QcNL_Comb n'en obtient qu'une seule. En général, les trois groupes d'essences obtiennent de meilleurs R^2 avec les métriques M_{tex} ou M_{comb}. Dans le cas du paramètre *scale*, les R^2 étaient plus élevés (QcNL_EPN = 0.45, QcNL_SAB = 0.48 et QcNL_Comb = 0.48).

En revanche, la présence des métriques de texture n'est pas marquée pour le paramètre *shape* pour QcNL_EPN. Seulement RF a permis de produire des modèles avec les 3 groupes de métriques. Dans le cas de QcNL_SAB, les R^2 les plus élevés sont obtenus avec le groupe M_{tex} pour les modèles de LEAP et GLMNET. Pour QcNL_Comb, le R^2 le plus élevé est retrouvé avec le modèles RF et le groupe de métriques M_{als}. Pour le paramètre *scale*, M_{comb} a fourni des R^2 légèrement plus élevés pour les placettes QcNL_EPN. Cependant, cette augmentation de performance est aussi observable pour les placettes

QcNL_SAB où la différence entre M_{als} et M_{comb} est plus marquée, avec des augmentations de R^2 situées entre 0 et 0.16 selon le modèle. Finalement, QcNL_Comb obtient également des améliorations entre M_{als} et M_{comb} de l'ordre de 0.05 à 0.10. Pour le paramètre *shape*, il n'y a pas vraiment de groupe de métrique qui produit l'ensemble des R^2 maximum. Au contraire, le paramètre *scale* obtient les valeurs les plus élevées avec M_{tex} ou M_{comb}, pour les trois groupes d'essences et les trois techniques de modélisation.

Tableau 3-3: Coefficient de détermination (R^2) des modèles de prédiction des paramètres de formes des distributions des diamètres des arbres (DDA) différentiées unimodales, pour les placettes du Québec et de Terre-Neuve dominées par l'épinette noire (QcNL_EPN), le sapin baumier (QcNL_SAB) et la combinaison des essences (QcNL_Comb).

		RF				**LEAP**			**GLMNET**	
		M_{als}	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**
Shape	QcNL_EPN	0,39	0,16	0,39	0,14	0,00	0,11	0,00	0,00	0,00
	QcNL_SAB	0,39	0,15	0,36	0,35	0,45	0,42	0,30	0,45	0,41
	QcNL_Comb	0,39	0,07	0,34	0,00	0,11	0,15	0,11	0,10	0,19
scale	QcNL_EPN	0,64	0,28	0,65	0,57	0,09	0,57	0,59	0,08	0,60
	QcNL_SAB	0,72	0,43	0,72	0,40	0,30	0,56	0,41	0,31	0,50
	QcNL_Comb	0,57	0,32	0,62	0,51	0,29	0,60	0,50	0,29	0,60

3.3.3 Paramètres des placettes bimodales.

Les résultats obtenus pour la prédiction des paramètres des DDA bimodales comportent plus de R^2 avec des valeurs négatives (Tableau 3-4.) Deux options de paramètres sont disponibles pour reconstruire une DDA bimodale composée d'un *finite mixture model* de deux fonctions de Weibull. Le premier groupe de paramètres est composé de la proportion, du *shape* et du *scale* de chaque Weibull. La deuxième option est composée de la proportion, de la moyenne et de l'écart-type de chaque composante. Cependant, les résultats obtenus pour le paramètre *shape1* (*shape* de la première composante) étaient tous négatifs et *shape2* n'avait que trois modèles avec des $R^2 < 0.04$. Aucun modèle n'a été en mesure de générer une prédiction adéquate, indépendamment du groupe de métriques. Au contraire, le groupe de paramètres formé de la proportion, de la moyenne et de l'écart-type ont obtenu des R^2 plus élevés (notés mean1, mean2, sd1, sd2). Tout d'abord, les valeurs de R^2 obtenues pour les deux paramètres de proportion des composantes des DDA bimodales (prop1 et prop2)

étaient les mêmes pour chaque combinaison de techniques statistiques et de groupe de métriques. C'est pour cette raison que prop1 et prop2 sont représentées ensemble dans le Tableau 3-4. Parmi les trois essences, QcNL_EPN et QcNL_Comb ont obtenu des R^2 moyen de 0 et 0.05 respectivement. Pourtant, la moyenne des R^2 pour QcNL_SAB est de 0.16 et le maximum est de 0.28. Au contraire, la valeur maximale de R^2 pour QcNL_EPN est de seulement 0.09, alors que pour QcNL_Comb à 0.15.

Pour le paramètre mean1, QcNL_EPN a obtenu un R^2 moyen de 0 et un maximum de 0.20. Au contraire, QcNL_Comb a obtenu une moyenne de R^2 de 0.17 et un maximum de 0.26. Par ailleurs, QcNL_Comb ne présentait pas de valeur négative. Au contraire, le paramètre mean2 a été mieux prédit pour les 3 groupes d'essences avec une moyenne de 0.13 pour QcNL_EPN, 0.30 pour QcNL_SAB et 0.39 pour QcNL_Comb. Or, la valeur maximale de R^2 appartient à QcNL_EPN à 0.55, suivi de QcNL_Comb à 0.53.

Tout comme les paramètres précédents, QcNL_EPN a rapporté le plus faible R^2 moyen pour sd1, avec une valeur de 0 causé par 4 R^2 négatifs sur 9. Les moyennes de QcNL_SAB et QcNL_Comb étaient similaires à 0.20 et 0.24 respectivement. Finalement, le paramètre sd2 a été le mieux prédit avec QcNL_EPN avec un R^2 moyen de 0.40 et un maximum de 0.64. Au contraire, QcNL_SAB a le moins bien performé avec une moyenne de 0.04 et un maximum de 0.12.

Dans le cas des groupes de métriques, les paramètres prop1 et prop2 n'ont pas semblé obtenir d'amélioration de performances. Pour les trois groupes d'essences, aucun groupe de métrique ne semblait performer plus que les autres pour les 3 trois techniques de modélisation. Pour Mean1, QcNL_EPN a obtenu des résultats médiocres avec les trois groupes de métriques. Au contraire, QcNL_SAB avait des performances plus élevées avec le groupe M_{tex} pour les trois techniques. Pour QcNL_Comb, M_{tex} et M_{comb} ont permis d'atteindre les R^2 les plus élevés, et M_{als} a obtenu des R^2 plus faibles avec des différences entre 0.02 et 0.15. Dans le cas de Mean2, QcNL_EPN avait des performances similaires à l'aide de M_{als} et M_{comb}. Les différences entre les R^2 des deux groupes de métriques sont inférieure à 0.05. Les deux groupes d'essence QcNL_SAB et QcNL_Comb ont tous obtenu de meilleurs R^2 en utilisant M_{tex} ou M_{comb}, avec des différences de 0.02 à 0.25 avec les

valeurs obtenues avec M_{als}. En ce qui concerne le paramètre std.dev1, les métriques M_{als} et M_{comb} ont fourni des performances similaires. Encore une fois, M_{tex} et M_{comb} ont permis d'obtenir des modèles plus performants qu'en utilisant M_{als} pour QcNL_SAB et QcNL_Comb. Finalement, M_{als} et M_{comb} ont également obtenu des performances semblables. Cependant, QcNL_SAB a obtenu des R^2 très faibles inférieurs à 0.09. Au contraire, les modèles de QcNL_Comb performaient mieux avec M_{comb} pour les trois techniques, alors que l'usage des M_{tex} seules a produit les modèles avec les R^2 les plus faibles.

Tableau 3-4: Coefficient de détermination (R^2) des modèles de prédiction des paramètres de formes des distributions des diamètres des arbres (DDA) différentiées bimodales pour les placettes du Québec et de Terre-Neuve dominées par l'épinette noire (QcNL_EPN), le sapin baumier (QcNL_SAB) et la combinaison des essences (QcNL_Comb).

		RF			LEAP			GLMNET		
		M_{als}	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**
Prop1-2	QcNL_EPN	0,09	-0	0,05	-0,01	-0,08	-0,01	-0,01	-0,05	-0,05
	QcNL_SAB	0,22	0,00	0,20	0,14	0,28	0,14	0,23	0,00	0,23
	QcNL_Comb	0,15	0,00	0,09	0,01	0,12	0,01	0,05	0,00	0,05
Mean1	QcNL_EPN	0,00	-0,82	-0,04	-0,26	-0,62	-0,06	0,14	0,07	0,20
	QcNL_SAB	0,00	0,06	0,01	0,09	0,20	0,03	0,03	0,16	0,06
	QcNL_Comb	0,09	0,12	0,11	0,12	0,25	0,26	0,10	0,25	0,21
Mean2	QcNL_EPN	0,50	0,15	0,55	0,02	-0,42	-0,47	0,47	-0,08	0,46
	QcNL_SAB	0,19	0,20	0,21	0,09	0,50	0,34	0,29	0,45	0,41
	QcNL_Comb	0,31	0,29	0,40	0,29	0,43	0,52	0,36	0,40	0,53
Std. Dev1	QcNL_EPN	0,23	-0,14	0,23	0,02	-0,42	-0,47	-0,01	-0,02	0,22
	QcNL_SAB	0,12	0,20	0,14	0,12	0,40	0,12	0,20	0,31	0,19
	QcNL_Comb	0,20	0,15	0,23	0,13	0,34	0,27	0,26	0,29	0,33
Std. Dev2	QcNL_EPN	0,44	0,19	0,42	0,50	0,24	0,46	0,47	0,20	0,64
	QcNL_SAB	0,00	0,00	0,00	0,12	0,00	0,12	0,09	0,00	0,00
	QcNL_Comb	0,35	0,11	0,36	0,24	0,09	0,43	0,41	0,08	0,43

3.3.4 Prédiction des paramètres des DDA sans différentiation de modalité

Pour les modèles sans classification de la DDA, les modèles ont relativement bien performé (Tableau 3-5). Seulement un seul modèle a obtenu un R^2 négatif lors de la prédiction du *shape* pour QcNL_EPN. Encore une fois, QcNL_EPN présentait un R^2 moyen de 0.16,

plus faible que QcNL_SAB et QcNL_Comb avec des moyennes de 0.33 et 0.37. Par ailleurs, QcNL_Comb a obtenu le R^2 le plus élevé à 0.52, comparativement à 0.43 pour QcNL_SAB et 0.32 pour QcNL_EPN. Dans le cas du paramètre *scale*, QcNL_SAB présentait un R^2 moyen plus faible à 0.48 et QcNL_Comb avait le plus élevé à 0.59. Par contre, les valeurs de R^2 maximales étaient similaires entre QcNL_EPN et QcNL_Comb à 0.71 et 0.73 respectivement.

Dans le cas de l'impact des métriques de texture sur la prédiction du paramètres *shape*, M_{comb} a produit les meilleures performances pour les trois groupes d'essences. Les améliorations entre M_{als} et M_{comb} sont comprises entre 0.01 et 0.13. Encore une fois, les R^2 maximums sont obtenu avec M_{comb} pour l'ensemble des techniques de modélisation et des groupes d'essences. Similairement, le paramètres *scale* est également mieux prédit avec M_{comb} pour l'ensemble des groupes d'essences et des techniques de modélisation. Les différences de R^2 entre M_{als} et M_{comb} se retrouvent entre 0.03 et 0.09. Cette amélioration est moins marquée que pour le *shape.*

Tableau 3-5: Coefficient de détermination (R^2) des modèles de prédiction des paramètres de formes des distributions des diamètres des arbres (DDA) sans différentiation de la modalité, pour les placettes du Québec et de Terre-Neuve dominées par l'épinette noire (QcNL_EPN), le sapin baumier (QcNL_SAB) et la combinaison des essences (QcNL_Comb).

		RF			**LEAP**			**GLMNET**		
		M_{als}	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**	**M_{stand}**	**M_{tex}**	**M_{comb}**
Shape	QcNL_EPN	0,29	-0,03	0,32	0,14	0,08	0,25	0,17	0,08	0,18
	QcNL_SAB	0,40	0,09	0,42	0,34	0,26	0,43	0,35	0,32	0,40
	QcNL_Comb	0,45	0,12	0,50	0,35	0,25	0,48	0,40	0,25	0,52
Scale	QcNL_EPN	0,64	0,25	0,68	0,62	0,32	0,71	0,65	0,32	0,68
	QcNL_SAB	0,50	0,39	0,54	0,48	0,42	0,56	0,50	0,40	0,54
	QcNL_Comb	0,64	0,37	0,69	0,65	0,42	0,73	0,67	0,42	0,73

3.3.5 Comparaison des modèles spécifique

Le développement de modèles spécifiques à l'essences n'a pas permis d'améliorer les performances. Certes, QcNL_SAB a présenté des R^2 plus élevés, mais QcNL_EPN a

obtenu des résultats inférieurs à la combinaison des essences. Cette différence est peut-être causée par la différence de structure des essences. La prédiction d'attributs liés à la structure est dépendant de l'espèce. Par exemple, la structure des sapins baumiers est plus régulière avec des formes coniques. Au contraire, la structure des épinettes noires est plus hétérogène, ce qui augmente la variabilité des paramètres possibles à prédire. Les forêts retrouvées sur le site d'étude de Terre-Neuve sont majoritairement dominées par le sapin baumier, mais contiennent également une présence d'épinette noire. Cette codominance est également observée pour les forêts du site d'étude du Québec. Il est donc avantageux d'utiliser les modèles développés avec QcNL_Comb, car ils performent très bien pour l'ensemble des paramètres à prédire.

4 Discussion

À la lumière des différentes réalisations, la recherche présentée comporte plusieurs aspects intéressants. Premièrement, la méthodologie utilisée pour différencier la modalité des DDA et pour développer des modèles de classification des DDA peuvent s'adapter aux besoins opérationnels des forestiers. Ces méthodes permettent de mieux identifier les différents types de structures de peuplement et ces modèles permettent d'estimer la DDA sur de larges couvertures BLA. Deuxièmement, nous avons démontré que les modèles développés avec les groupes de métriques M_{tex} ou M_{comb} produisaient des R^2 plus élevés et des RMSD% plus faibles, en comparaison aux modèles employant M_{als} seul. Ces résultats indiquent que les métriques de texture contiennent une information additionnelle utile à l'estimation de la DDA. La différentiation de la modalité des DDA a obtenu des valeurs de précision globale moyenne presque identique entre les quatre techniques statistiques sélectionnées (M_{als} = 71,8%, M_{tex} = 71.3%, M_{comb} = 71.1%). Dans le cas de la prédiction des paramètres *shape* et *scale* des DDA unimodales, la moyenne des R^2 pour les métriques M_{comb} était de 0.42 et de 0.35 pour M_{als}. Toutefois, les RMSD relatifs n'ont pas présenté une différence aussi marquée, soit 15% pour M_{comb} et 16% pour M_{als}. Cette tendance est également retrouvée pour les modèles sans différentiation de modalité. La moyenne des R^2 pour les paramètres *shape* et *scale* se situe de 0.53 pour M_{als} et 0.61 pour M_{comb} avec des

valeurs moyennes de RMSD relatif de 15% pour M_{als} et 13% pour M_{comb}. Dans le cas des placettes bimodales, les performances des modèles étaient moins performantes. Néanmoins, la moyenne des R^2 de tous les paramètres (*proportion*, *mean, standard deviation, shape* et *scale*) était plus élevée avec les métriques M_{comb} (R^2 moyen = 0.23) qu'avec les métriques M_{als} (R^2 moyen = 0.17). En revanche, les valeurs de RMSD relatif moyen étaient très similaires, soient 60% pour M_{comb} et 61% pour M_{als}. Même si le groupe de métrique Mc_{omb} comprenait toutes les métriques retrouvées dans M_{als}, les techniques de modélisations ont permis d'évaluer l'importance des métriques individuelles dans la prédiction des différents paramètres. Pour la majorité des modèles utilisés pour la reconstruction des DDA, le modèle final comprenait au moins une métrique de texture dans les 10 métriques les plus importantes. Sur 15 modèles finaux (différentiation de la SDD, paramètres de la DDA unimodale, paramètres des DDA bimodales et DDA sans différentiation), 11 modèles comprenaient des métriques de texture dans leurs top 10 prédicteurs les plus importants. Ce qui indique que les métriques de texture apportent une information de plus dans les modèles prédictifs de la DDA. Par exemple, les modèles prédictifs des paramètres *shape* et *scale* des placettes unimodales ne contenaient aucune métrique de texture. Or, les modèles prédictifs des paramètres *shape* et *scale* des placettes sans différentiation des modalités ont une métrique de texture comme prédicteur le plus important en plus d'utiliser 3 et 4 métriques de texture dans le top 10. Les placettes sans différentiation de modalité comprennent une plus grande hétérogénéité dans la structure de la forêt, ce qui se traduit par une différence dans la canopée. Les métriques de texture issues du MHC permettent de capter cette différence comparativement aux métriques BLA.

Lors de l'analyse des résultats, nous avons assumé que la différentiation de la modalité des DDA était adéquate. Tout comme d'autres études similaires (Mulverhill *et al*., 2018; Zhang *et al*., 2019), nous avons implémenté le coefficient de bimodalité (BC) comme méthode de différentiation des modalités. Cependant, cette méthode de différentiation des modalités est directement influencée par le kurtosis et encore plus par le *skewness* de la distribution (Pfister *et al*., 2013). Une distribution qui présente un *skewness* élevé et un kurtosis faible peut augmenter artificiellement les valeurs de BC et classifier une distribution comme bimodale. Les distributions qui présentent un *skewness* négatif sont généralement observés

lorsque des arbres à fort DHP dominent le peuplement, alors que les *skewness* positifs se retrouvent dans les peuplements dominés par des arbres à faible DHP. Ces deux situations produisent des valeurs de *skewness* supérieures à 0 et produisent une asymétrie dans la distribution. Plus une valeur de *skewness* s'approche de 0, plus la DDA sera homogène et mieux elle caractérise les peuplements équiennes (Zhang and Liu, 2006). De plus, Freeman and Dale (2013) ont testé différents paramètres qui pouvaient affecter la différentiation de la modalité à l'aide du BC en simulant une multitude de distributions. Ils ont évalué les impacts du *skewness*, de la proportion et de la distance entre les modes sur les valeurs obtenues du BC. Dans leurs résultats, le BC produisait 21% de faux-positifs où des distribution unimodales simulées obtenaient des valeurs de BC supérieures au seuil de bimodalité de 5/9. Le calcul du BC repose sur la supposition que la bimodalité implique une augmentation de l'asymétrie de la distribution ce qui implique qu'une augmentation du *skewness*. Dans un contexte unimodal, un fort *skewness* peut causer une augmentation de la valeur du BC et produire un faux positif. En ajout, le BC n'est pas calibré selon différentes valeurs de proportion entre les deux modes. Une proportion faible pour un des deux modes, ainsi qu'une petite distance entre les deux moyennes des modes peut entrainer une classification erronée. Pourtant, les placettes classifiées bimodales ont également présenté certains éléments qui pourraient expliquer la mauvaise performance du BC. Un des paramètres importants d'une DDA bimodale est la proportion de chaque Weibull qui compose le *finite mixture model*. Naïvement, il est plausible de penser qu'une DDA classée bimodale comprenne deux modes contenant une proportion semblable d'observations. Or, les placettes de Terre-Neuve présentaient des valeurs de la proportion associée à la seconde composante de la distribution bimodale avec des valeurs faibles (proportion < 25%). En moyenne, plus de 75% des DHP mesurés se retrouvaient dans la première composante. De plus, les distances entre les moyennes des modes étaient également faibles avec 5 cm de moyenne. Cette combinaison de caractéristiques a pu causer une inflation du BC et générer une mauvaise différentiation de la modalité des DDA, étant donné que la prédiction des DDA avec les modèles spécifiques aux modalités n'ont produit qu'une faible amélioration. Ces effets des caractéristiques de la DDA sur le BC indiquent qu'il ne faut pas se fier uniquement sur une méthode de différentiation des modalités et qu'une même méthode n'est pas nécessairement efficace sur tous les types de forêts. L'approche de différentiation

des modalités utilisant des caractéristiques de la courbe de Lorenz (Zhang *et al*., 2019) incorpore des mesures directement liées à l'hétérogénéité du peuplement et du *skewness* de la DDA (Bouvier *et al*., 2015; Thomas *et al*., 2008). Cependant, il est nécessaire d'effectuer davantage de recherche pour déterminer la méthode de différentiation des modalités de la DDA selon le type de forêt à l'étude.

En plus d'être relativement simple à calculer, les métriques de textures ont permis d'améliorer les modèles prédictifs d'un attribut forestier relativement complexe à prédire. En effet, ce projet se concentrait sur les forêts boréales dominées par les conifères. Cependant, les études similaires se retrouvaient dans des forêts mixtes dominées par des feuillus. Il serait intéressant de tester l'approche développée sur des forêts dominées par les feuillus pour comparer l'effet de la structure des peuplements et les performances de la prédiction de la DDA. Malheureusement, la DDA est une caractéristique complexe de la forêt et elle ne peut être représentée spatialement sous forme d'une carte avec les résultats de ce projet. Au contraire, des attributs forestiers comme la surface terrière ou la hauteur sont couramment représentés spatialement en format matriciel. Actuellement, il serait possible de représenter spatialement la forme générale de la DDA (unimodale ou bimodale), et ensuite appliquer les bons modèles prédictifs des paramètres (ex : *scale* et *shape*). Mais, la reconstruction de la DDA nécessite des étapes de plus et est très difficile à représenter, ce qui limite l'accès et le transfert de connaissances aux décideurs et gestionnaires du milieu. Pour pallier cette difficulté de mise en opération, une approche plus grossière, mais plus robuste pourrait servir à identifier des secteurs d'intérêts, puis d'affiner les prédictions de la DDA à l'aide de l'approche de la présente recherche. Par exemple, une approche alternative consisterait à grouper les mesures de DHP d'une placette en *n* intervalles, puis de prédire la fréquence de chaque intervalle à l'aide de métriques BLA et de texture. La définition des intervalles peut être propre aux intérêts de l'industrie. Donc, la DDA ne serait pas représentée par une fonction continue, mais un histogramme comprenant un nombre restreint de classes. Il serait donc possible de générer des produits cartographiques avec le nombre d'arbre prédits dans chaque intervalle prédéterminé et cibler les sites d'intérêts.

Également, cette recherche a montré les métriques de textures présentent un avantage à être intégrées dans les modèles prédictifs d'attributs forestiers. Ces métriques de textures pourraient être incluses dans des modèles prédictifs de l'essence d'arbre et peut-être ainsi bonifier les performances. De plus, la recherche a confirmé que le choix de technique de classification de la DDA est primordial. Il faut donc tester plus d'une technique pour trouver une tendance dans les données et confirmer la présence de DDA bimodales. Une technique peut fonctionner pour un site, mais pas pour un autre, dû à la structure des peuplements. De même, elle a permis de développer une méthodologie complète de prédiction de la DDA pour des forêts boréales ayant des structures hétérogènes et de mesurer les limites de la méthode. Pour terminer, il aurait été intéressant de tester des fonctions autres que la Weibull, comme une fonction log-normale ou gamma. Tant pour la placette unimodale que bimodale. Ces tests approfondis pourraient servir à déterminer quelles PDF permet de mieux représenter les peuplements avec un fort *skewness*. Cet aspect se répercute sur la prédiction du paramètres *shape* des fonctions Weibull. Le paramètre *shape* est plus difficile à prédire, car il est généralement plus sensible au bruit dans les données. Le paramètres *shape* correspond à la mesure de la vitesse à laquelle la probabilité d'un DHP *i* diminue lorsque le DHP augmente. Il s'avère donc difficile de capter les variations subtiles dans les valeurs du paramètre *shape*. Plusieurs DDA présentaient une forte densité de mesures dans les DHP les plus faible et ensuite, la densité diminuait pour générer une longue queue à la DDA. Cette longue queue engendre un *skewness* élevé. Cependant, le *shape* contrôle d'une certaine façon la pente de la distribution. Or, les résultats montrent une très grande différence entre la pente de la Weibull lorsque la densité augmente en même temps que le DHP augmente, comparativement à la portion où la densité diminue. Par exemple, dans la Figure *1-1*, la pente de la Weibull entre 9 cm 14cm de DHP semble similaire à la pente retrouvée entre le DHP de 14 cm à 21 cm. C'est pour cette raison que la Weibull représente mieux la DDA que les deux autres fonctions. En revanche, la DDQA retrouvée à la Figure *1-2*, comprend une seule pente. Étant donné que la plus forte densité est retrouvée au plus faible DHP, la DDA ne comprend qu'une seule pente, soit la diminution de la densité en fonction de l'augmentation du DHP. Or, le paramètre *shape* de la Weibull ne permet pas de bien capter l'ensemble des faibles DHP. Pour ce genre de DDA, la médiane est plus faible que la

moyenne, ce qui découle de sa forte densité de mesure dans les DHP faibles et la moyenne gonflée par les quelques mesures plus élevées.

Finalement, différents auteurs (Hou *et al*., 2016; Leclère *et al*., 2022; Packalén and Maltamo, 2007; Zhang *et al*., 2019) ont obtenus de meilleures prédictions de la DDA en développement des modèles spécifiques aux essences dominantes des peuplements. Or, les modèles spécifiques au épinettes noires (QcNL_EPN) ont obtenus les performances les plus faibles pour les paramètres *shape* et *scale* des DDA unimodales, de même que pour le *shape* des DDA sans différentiation. Également, QcNL_EPN avait les R^2 moyens les plus faibles pour la plupart des paramètres des DDA bimodales, sauf pour l'écart-type de la deuxième composante (std.dev2). Ces mauvaises performances sont observées à la fois pour les placettes unimodale ou bimodale. Il se peut que la fonction Weibull ne soit pas la plus adaptée pour représenter la forme générale des DDA pour les peuplements dominés par l'épinette noire. Étant donné que les paramètres des DDA ne permettent pas une représentation adéquate de la forme générale, les métriques ne permettent pas d'expliquer la variance des paramètres. Au contraire, le paramètres *shape* des placettes QcNL_SAB est le mieux prédit, ce qui porte à croire que la structure de la DDA est mieux représentée par une fonction de Weibull. Or, la combinaison des essences (QcNL_Comb) semble bénéficier aux performances du modèle. En moyenne, les R^2 obtenus par QcNL_Comb avaient des valeurs légèrement inférieures à celles de QcNL_SAB ou supérieures. Pour presque tous les modèles, les valeurs de R^2 de QcNL_Comb étaient plus élevée que celles de QcNL_EPN. Cette différence majeure avec les autres études pourrait provenir de la plus grande complexité des structures des peuplements, ainsi que du choix de la fonction ajustée sur la DDA. En addition, la structure conique caractéristique des sapins baumier permet une meilleure pénétration du BLA, ce qui permet de capturer une plus grande portion de la structure verticale. À l'inverse, les peuplements d'épinettes noires ont une structure plus hétérogène, ce qui limite la pénétration du BLA, mais augmente la variabilité des paramètres à prédire. La corrélation entre les paramètres de la DDA et les métriques dérivées est donc plus faible.

5 Conclusion

Cette étude comparait deux approches pour prédire la DDA. La première approche consistait à prédire les paramètres de la DDA avec des modèles spécifiques à la modalité de la distribution. La deuxième ne nécessitait aucune différentiation de modalité et le même modèle était employé pour prédire les paramètres des DDA. Aussi, nous avons comparé les performances des modèles à l'aide de différents groupes de métriques qui comportaient, soit uniquement des métriques BLA (M_{als}), soit uniquement des métriques de textures dérivées du MHC (M_{tex}) ou une combinaison des différentes sources de métriques (M_{comb}). La comparaison des performances a permis de constater que la prédiction de la DDA présentait de meilleures performances lorsque les modèles utilisaient des groupes contenant des métriques de texture (M_{tex}, M_{comb}). Les R^2 obtenus ont présenté dans augmentations situées entre 0.03 et 0.25, selon le paramètres prédit. Par ailleurs, nous avons démontré que la classification des DDA selon leur modalité était améliorée lorsque réalisée avec M_{tex}, grâce à l'information additionnelle contenue dans les métriques.

Nous avons également démontré que le développement de modèles spécifiques à la modalité améliorait uniquement les prédictions des DDA bimodales. Au contraire, l'utilisation de modèles spécifiques à la modalité repose étroitement sur la différentiation des modalités. Cette étape clé doit être bien réalisée, sinon les performances des modèles sont fortement impactées comme observé avec les paramètres des DDA bimodales (R^2 entre 0.09 et 0.64 selon l'essence dominante et le paramètre). Par ailleurs, nous avons également mis l'accent sur le choix de la fonction pour représenter la DDA. Dans notre cas, la fonction Weibull ne captait pas adéquatement les DDA dominées par les arbres à faible DHP. Des études plus approfondies comparant d'autres options de fonctions sera nécessaire pour sélection celle qui est le mieux adaptée aux types de forêt rencontrés. Nous avons également démontré l'importance de tester plusieurs techniques de classification de la modalité, étant donné les nombreuses limitations du BC pouvant facilement générer des faux-positifs. De plus, la différentiation de la modalité des DDA devrait être évaluée pour chaque type de peuplement, étant données les grandes variabilités

structurelles qui découlent des perturbations et des essences présentent et sélectionner la fonction la plus représentative.

Finalement, nous avons démontré l'utilité des métriques de textures dérivées du MHC. La simplicité des métriques de textures présente un avantage aux métriques BLA, qui nécessitent plus de connaissances et de puissances de calcul pour leur utilisation opérationnelle. Par ailleurs, nous avons également abordé le potentiel des métriques de texture dans la prédiction d'attributs forestiers plus complexes. Également, nous avons abordé l'utilité des métriques de texture dans le développement de modèles prédictifs d'attributs forestiers en combinant les essences dominantes. Cet avantage permet de réduire le nombre de modèles nécessaires pour traiter une zone désirée. L'ajout de ces métriques de texture ne se limite pas uniquement la prédiction de la DDA, mais également à n'importe quel modèle prédictif d'attributs forestiers nécessaire à la gestion optimisée des ressources forestières.

6 Références hors article

Beets, P. N., Reutebuch, S., Kimberley, M. O., Oliver, G. R., Pearce, S. H., & McGaughey, R. J. (2011). Leaf area index, biomass carbon and growth rate of radiata pine genetic types and relationships with LiDAR. *Forests*, *2*(3), 637–659. https://doi.org/10.3390/f2030637

Bouvier, M., Durrieu, S., Fournier, R. A., & Renaud, J.-P. P. (2015). Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. *Remote Sensing of Environment*, *156*, 322–334. https://doi.org/10.1016/j.rse.2014.10.004

Cheng, M.-Y., & Hall, P. (1999). Mode testing in difficult cases. *The Annals of Statistics*, *27*(4), 1294–1315. https://doi.org/10.1214/aos/1017938927

Coomes, D. A., Duncan, R. P., Allen, R. B., & Truscott, J. (2003). Disturbances prevent stem size-density distributions in natural forests from following scaling relationships. *Ecology Letters*, *6*(11), 980–989. https://doi.org/10.1046/j.1461-0248.2003.00520.x

Cordonnier, T., Dreyfus, P., & Trouvé, R. (2012). Quelles dimensions et quels indices d'hétérogénéité privilégier pour l'expérimentation dans les peuplements forestiers mélangés ou irréguliers? *Revue Forestiere Francaise*, *64*(6), 773–787. https://doi.org/10.4267/2042/51115

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4), 1–34. https://doi.org/10.18637/jss.v064.i04

Diao, J., Liu, J., Zhu, Z., Wei, X., & Li, M. (2022). Active forest management accelerates carbon storage in plantation forests in Lishui, southern China. *Forest Ecosystems*, *9*, 100004. https://doi.org/10.1016/j.fecs.2022.100004

Ellison, A. M. (1987). Effect of seed dimorphism on the density-dependent dynamics of experimental populations of Atriplex triangularis (Chenopodiaceae). *American Journal of Botany*, *74*(8), 1280–1288. https://doi.org/10.2307/2444163

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83–97. https://doi.org/10.3758/s13428-012-0225-x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Furnival, G. M., & Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics*, *16*(4), 499–511. https://doi.org/10.2307/1271435

Ginzler, C. (2019). *Remote Sensing Data Sources*. *May 2022*, 95–100. https://doi.org/10.1007/978-3-030-19293-8_3

Gobakken, T., & Næsset, E. (2004). Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. *Scandinavian Journal of Forest Research*, *19*(6), 529–542. https://doi.org/10.1080/02827580410019454

Goetz, S., Steinberg, D., Dubayah, R., & Blair, B. (2007). Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. *Remote Sensing of Environment*, *108*(3), 254–263. https://doi.org/10.1016/j.rse.2006.11.016

Groot, A., Cortini, F., & Wulder, M. A. (2015). Crown-fibre attribute relationships for enhanced forest inventory: Progress and prospects. *The Forestry Chronicle*, *91*(3), 266–279. https://doi.org/https://dx.doi.org/10.5558/tfc2015-048

Guo, H., Lei, X., You, L., Zeng, W., Lang, P., & Lei, Y. (2022). Climate-sensitive diameter distribution models of larch plantations in north and northeast China. *Forest Ecology and Management*, *506*, 119947. https://doi.org/10.1016/J.FORECO.2021.119947

Hall-Beyer, M. (2017). Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. *International Journal of Remote Sensing*, *38*(5), 1312–

1338. https://doi.org/10.1080/01431161.2016.1278314

Haralick, R. M., Shanmugam, K., & Its'Hak, D. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*, 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*, 14. https://doi.org/10.1214/aos/1176346577

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Springer Series in Statistics The Elements of Statistical Learning* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-21606-5

Hopkinson, C., & Chasmer, L. (2009). Testing LiDAR models of fractional cover across multiple forest ecozones. *Remote Sensing of Environment*, *113*(1), 275–288. https://doi.org/10.1016/j.rse.2008.09.012

Hou, Z., Xu, Q., Vauhkonen, J., Maltamo, M., & Tokola, T. (2016). Species-specific combination and calibration between area-based and tree-based diameter distributions using airborne laser scanning. *Canadian Journal of Forest Research*, *46*(6), 753–765. https://doi.org/https://doi.org/10.1139/cjfr-2016-0032

Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E., & Falkowski, M. J. (2008). Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, *112*(5), 2232–2245. https://doi.org/https://doi.org/10.1016/j.rse.2007.10.009

Jenness, J. S. (2004). Calculating landscape surface area from digital elevation models. *Wildlife Society Bulletin*, *32*(3), 829–839. https://doi.org/10.2193/0091-7648(2004)032[0829:clsafd]2.0.co;2

Johnsson, K., Linderoth, M., & Fontes, M. (2017). What is a "unimodal" cell population? Using statistical tests as criteria for unimodality in automated gating and quality control. *Cytometry Part A*, *91*(9), 908–916. https://doi.org/https://doi.org/10.1002/cyto.a.23173

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1–20. https://doi.org/10.18637/jss.v011.i09

Knoebel, B. R., & Burkhart, H. E. (1991). A bivariate distribution approach to modeling forest diameter distributions at two points in time. *Biometrics*, *47*(1), 241–253. https://doi.org/10.2307/2532509

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M. (2022). *caret: Classification and Regression Training*. https://cran.r-project.org/package=caret

Kvalvik, I., Solås, A. M., & Sørdahl, P. B. (2020). Introducing the ecosystem services concept in Norwegian coastal zone planning. *Ecosystem Services*, *42*(December 2019), 101071. https://doi.org/10.1016/j.ecoser.2020.101071

Land, A. H., & Doig, A. G. (1960). An Automatic Method of Solving Discrete Programming Problems. *Econometrica*, *28*(3), 497–520. https://doi.org/10.2307/2223855

Lechner, A. M., Foody, G. M., & Boyd, D. S. (2020). Applications in Remote Sensing to Forest Ecology and Management. *One Earth*, *2*(5), 405–412. https://doi.org/10.1016/j.oneear.2020.05.001

Leclère, L., Lejeune, P., Bolyn, C., & Latte, N. (2022). Estimating Species-Specific Stem Size Distributions of Uneven-Aged Mixed Deciduous Forests Using ALS Data and Neural Networks. *Remote Sensing*, *14*(6). https://doi.org/10.3390/rs14061362

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Lumley, T. (2020). *leaps: Regression Subset Selection, based on Fortran code by Alan Miller, R package version 3.1*. https://cran.r-project.org/web/packages/leaps/leaps.pdf

Mauro, F., García-Abril, A., Ayuga-Téllez, E., Rojo-Alboreca, A., Valbuena, R., & Antonio Manzanera, J. (2021). Comparison of two parameter recovery methods for the transformation of Pinus sylvestris yield

tables into a diameter distribution model. *Annals of Forest Science*, *78*(12), 15. https://doi.org/10.1007/s13595-021-01028-5ï

Mulverhill, C., Coops, N. C., White, J. C., Tompalski, P., Marshall, P. L., & Bailey, T. (2018). Enhancing the Estimation of Stem-Size Distributions for Unimodal and Bimodal Stands in a Boreal Mixedwood Forest with Airborne Laser Scanning Data. *Forests*, *9*(2), 95. https://doi.org/10.3390/f9020095

Nasrabadi, N. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4). https://doi.org/https://doi.org/10.1117/1.2819119

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., & Pereira, J. M. C. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, *275*, 117–129. https://doi.org/10.1016/j.foreco.2012.03.003

Packalén, P., & Maltamo, M. (2007). The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2007.01.005

Packalén, P., & Maltamo, M. (2008). Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. *Canadian Journal of Forest Research*, *38*(7), 1750–1760. https://doi.org/10.1139/x08-037

Palahi, M., Pukkala, T., Blasco, E., Trasobares, A., Palahí, M., Pukkala, T., Blasco, E., & Trasobares, A. (2007). Comparison of beta, Johnson's SB, Weibull and truncated Weibull functions for modeling the diameter distribution of forest stands in Catalonia (north-east of Spain). *European Journal of Forest Research*, *126*(4), 563–571. https://doi.org/10.1007/s10342-007-0177-3

Peduzzi, A., Wynne, R. H., Fox, T. R., Nelson, R. F., & Thomas, V. A. (2012). Estimating leaf area index in intensively managed pine plantations using airborne laser scanner data. *Forest Ecology and Management*, *270*, 54–65. https://doi.org/10.1016/j.foreco.2011.12.048

Peuhkurinen, J., Tokola, T., Plevak, K., Sirparanta, S., Kedrov, A., & Pyankov, S. (2018). Predicting Tree Diameter Distributions from Airborne Laser Scanning, SPOT 5 Satellite, and Field Sample Data in the Perm Region, Russia. *Forests*, *9*(10), 639. https://doi.org/10.3390/f9100639

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. (2013). Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in Psychology*, *4*, 700. https://doi.org/10.3389/fpsyg.2013.00700

Podlaski, R. (2016). Highly skewed and heavy-tailed tree diameter distributions: Approximation using the gamma shape mixture model. *Canadian Journal of Forest Research*, *46*(11), 1275–1283. https://doi.org/10.1139/cjfr-2016-0175

Pope, G., & Treitz, P. (2013). Leaf Area Index (LAI) estimation in boreal mixedwood forest of Ontario, Canada using Light detection and ranging (LiDAR) and worldview-2 imagery. *Remote Sensing*, *5*(10), 5040–5063. https://doi.org/10.3390/rs5105040

Poudel, K. P., & Cao, Q. V. (2013). Evaluation of Methods to Predict Weibull Parameters for Characterizing Diameter Distributions. *Forest Science*, *59*(2), 243–252. https://doi.org/10.5849/FORSCI.12-001

Pretzsch, H. (2010). Description and analysis of stand structures. In *Forest Dynamics, Growth and Yield.* (pp. 223–289). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88307-4_7

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Reynolds, M. R., Burk, T. E., & Huang, W.-C. (1988). Goodness-of-fit tests and model selection procedures for diameter distribution models. *Forest Science*, *34*(2), 373–399. https://doi.org/10.1093/forestscience/34.2.373

Roussel, J.-R., & Auty, D. (2021). *lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. https://cran.r-project.org/web/packages/lidR/index.html

Rowe, J. S. (1972). *Forest Regions of Canada. Based on W. E. D. Halliday's "A forest classification for*

Canada" 1937. PublicationNo 1300; Department of the Environment, Canadian Forestry Service: Ottawa, ON, Canada. https://publications.gc.ca/collections/collection_2019/eccc/Fo47-1300-eng.pdf

SAS Institute Inc. (1990). *SAS/STAT user's guide: version 6* (Vol. 2). Sas Inst.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). {mclust} 5: clustering, classification and density estimation using {G}aussian finite mixture models. *The {R} Journal*, *8*(1), 205–233. https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf

Shabani, N., Akhtari, S., & Sowlati, T. (2013). Value chain optimization of forest biomass for bioenergy production: A review. *Renewable and Sustainable Energy Reviews*, *23*, 299–311. https://doi.org/10.1016/j.rser.2013.03.005

Shahi, S., & Pulkki, R. (2015). ARTICLE A simulation-based optimization approach to integrated inventory management of a sawlog supply chain with demand uncertainty. *Canadian Journal of Forest Research*, *45*, 1313–1326. https://doi.org/10.1139/cjfr-2014-0373

Shang, C., Treitz, P., Caspersen, J., & Jones, T. (2017). Estimating Stem Diameter Distributions in a Management Context for a Tolerant Hardwood Forest Using ALS Height and Intensity Data. *Canadian Journal of Remote Sensing*, *43*(1), 79–94. https://doi.org/10.1080/07038992.2017.1263152

Siipilehto, J., & Mehtätalo, L. (2013). Parameter recovery vs. parameter prediction for the Weibull distribution validated for Scots pine stands in Finland. *Silva Fennica*, *47*, 22 p. https://doi.org/10.14214/sf.1057

Solberg, S., Brunner, A., Hanssen, K. H., Lange, H., Næsset, E., Rautiainen, M., & Stenberg, P. (2009). Mapping LAI in a Norway spruce forest using airborne laser scanning. *Remote Sensing of Environment*, *113*(11), 2317–2327. https://doi.org/10.1016/j.rse.2009.06.010

Strunk, J. L., Gould, P. J., Packalen, P., Poudel, K. P., Andersen, H.-E. E., & Temesgen, H. (2017). An Examination of Diameter Density Prediction with k-NN and Airborne Lidar. *Forests*, *8*(11), 444. https://doi.org/10.3390/f8110444

Tarp-Johansen, M. J. (2002). Stem diameter estimation from aerial photographs. *Scandinavian Journal of Forest Research*, *17*(4), 369–376. https://doi.org/10.1080/02827580260138116

Thomas, V., Oliver, R. D., Lim, K., & Woods, M. (2008). LiDAR and Weibull modeling of diameter and basal area. *The Forestry Chronicle*, *84*(6), 866–875. https://doi.org/10.5558/tfc84866-6

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, *112*(5), 1982–1999. https://doi.org/10.1016/j.rse.2007.03.032

Torgo, L. (2017). *Data mining with R: learning with case studies.* (Second Edi). Chapman and Hall/CRC. https://doi.org/https://doi.org/10.1201/9781315399102

Valbuena, R., Maltamo, M., Martín-Fernández, S., Packalen, P., Pascual, C., & Nabuurs, G. J. (2013). Patterns of covariance between airborne laser scanning metrics and Lorenz curve descriptors of tree size inequality. *Canadian Journal of Remote Sensing*, *39*(SUPPL.1), 37–41. https://doi.org/10.5589/m13-012

Valbuena, R., Packalén, P., Martín-Fernández, S., & Maltamo, M. (2012). Diversity and equitability ordering profiles applied to study forest structure. *Forest Ecology and Management*, *276*, 185–195. https://doi.org/10.1016/j.foreco.2012.03.036

Valbuena, R., Vauhkonen, J., Packalen, P., Pitkänen, J., & Maltamo, M. (2014). Comparison of airborne laser scanning methods for estimating forest structure indicators based on Lorenz curves. *ISPRS Journal of Photogrammetry and Remote Sensing*, *95*, 23–33. https://doi.org/10.1016/j.isprsjprs.2014.06.002

van Ewijk, K., Treitz, P., Woods, M., Jones, T., & Caspersen, J. (2019). Forest Site and Type Variability in ALS-Based Forest Resource Inventory Attribute Predictions over Three Ontario Forest Sites. *Forests*, *10*(3), 226. https://doi.org/10.3390/f10030226

van Ewijk, K. Y., Treitz, P. M., & Scott, N. A. (2011). Characterizing forest succession in central Ontario using lidar-derived indices. *Photogrammetric Engineering and Remote Sensing*, *77*(3), 261–269. https://doi.org/10.14358/PERS.77.3.261

Vandendaele, B., Fournier, R. A., Vepakomma, U., Pelletier, G., Lejeune, P., & Martin-ducup, O. (2021). Estimation of northern hardwood forest inventory attributes using uav laser scanning (Uls): Transferability of laser scanning methods and comparison of automated approaches at the tree- and stand-level. *Remote Sensing*, *13*(14). https://doi.org/10.3390/rs13142796

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://doi.org/10.1007/978-0-387-21706-2

White, J. C., Coops, N. C., Wulder, M. A., Vastaranta, M., Hilker, T., & Tompalski, P. (2016). Remote sensing technologies for enhancing forest inventories: A review. *Canadian Journal of Remote Sensing*, *42*(5), 619–641. https://doi.org/10.1080/07038992.2016.1207484

Woods, M., Pitt, D., Penner, M., Lim, K., Nesbitt, D., Etheridge, D., & Treitz, P. (2011). Operational implementation of a LiDAR inventory in Boreal Ontario. *Forestry Chronicle*, *87*(4), 512–528. https://doi.org/10.5558/tfc2011-050

Yu, Y. (2022). mixR: An R package for Finite Mixture Modeling for Both Raw and Binned Data. *Journal of Open Source Software*, *7*(69), 4031. https://doi.org/10.21105/joss.04031

Zeileis, A. (2014). *ineq: Measuring Inequality, Concentration, and Poverty. R pacakge verion 0.2-13. https://CRAN.R-project.org/package=ineq*. 15. https://cran.r-project.org/:CRAN

Zhang, L., & Liu, C. (2006). Fitting irregular diameter distributions of forest stands by Weibull, modified Weibull, and mixture Weibull models. *Journal of Forest Research*, *11*(5), 369–372. https://doi.org/10.1007/s10310-006-0218-7

Zhang, Z., Cao, L., Mulverhill, C., Liu, H., Pang, Y., & Li, Z. (2019). Prediction of Diameter Distributions with Multimodal Models Using LiDAR Data in Subtropical Planted Forests. *Forests*, *10*(2), 125. https://doi.org/10.3390/f10020125

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(5), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x