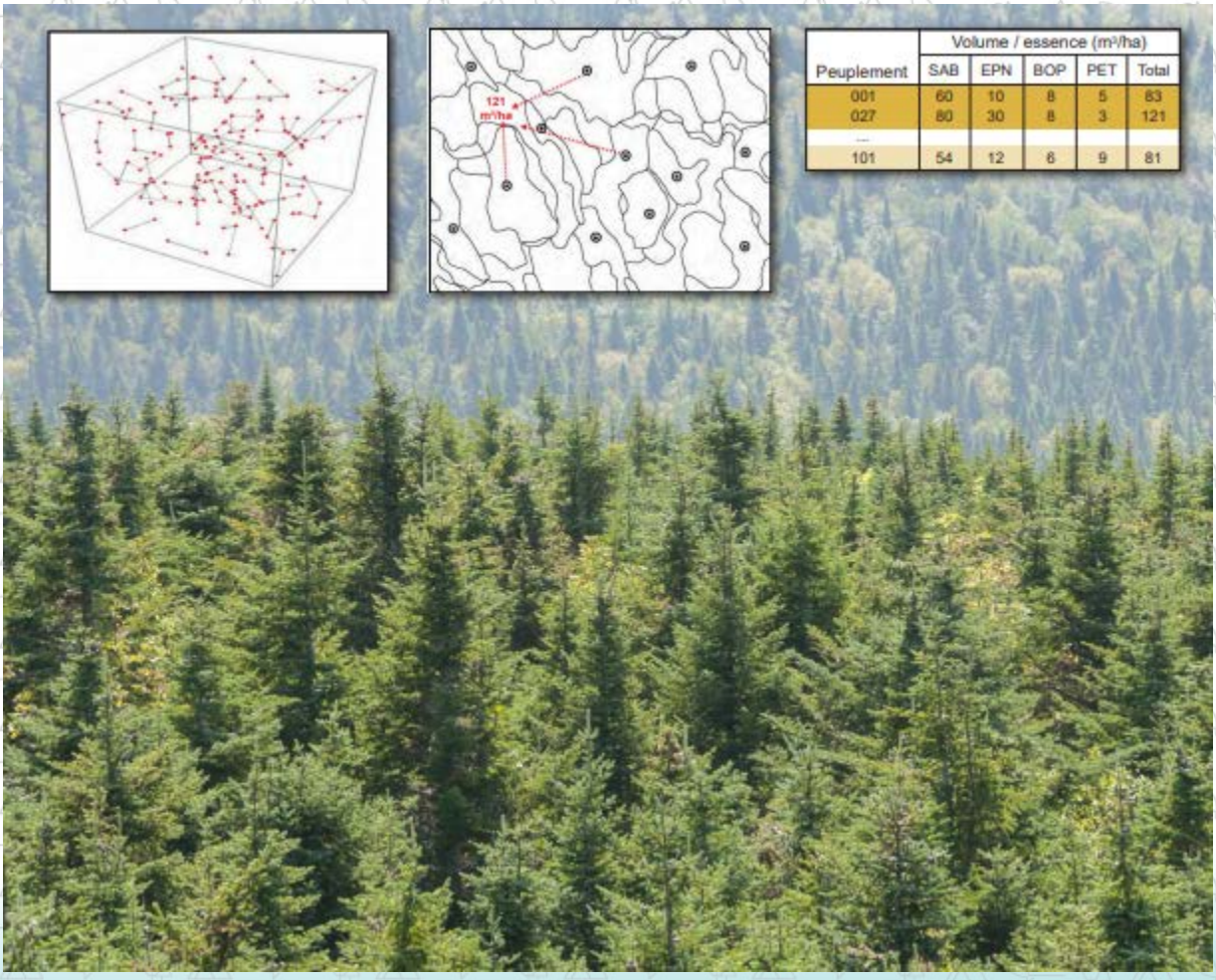
Méthodologie des compilations forestières du 4e inventaire écoforestier du Québec méridional : cas particulier des estimations k-NN

Mars 2017

MINISTÈRE DES FORÊTS, DE LA FAUNE ET DES PARCS



Votre gouvernement

Québec

Pour obtenir des renseignements additionnels, veuillez communiquer avec le ministère des Forêts, de la Faune et des Parcs du Québec :

Direction des inventaires forestiers
5700, 4e Avenue Ouest, A-108
Québec (Québec) G1H 6R1
Téléphone : 418 627-8669
Sans frais : 1 877 936-7387
inventaires.forestiers@mffp.gouv.qc.ca
mffp.gouv.qc.ca/les-forets/inventaire-ecoforestier/

Ministère des Forêts, de la Faune et des Parcs
Dépôt légal – Bibliothèque et Archives nationales du Québec, 2017
ISBN 978-2-550-78110-3 (1re édition, mars 2017) *

* Modifications mineures effectuées (mise en page, liens hypertextes, etc.) en janvier 2022

Rédaction
Jean-Gabriel Élie, ing.f., M. Sc.[1]

Coordination
Anne Morissette, géomorphologue, M. Sc.[1]

Collaboration
Carl Bergeron, ing.f., M. Sc. [1]
Sylvain Bernier, stat., M. Sc. [1]
Bastien Ferland-Raymond, M. Sc. Stat., M. Sc. Biol. [1]

Graphisme
Valérie Roy, techn. géo. [1]

Photographie de la page couverture
Robin Lefrançois, techn. forest. [1]

Mise en page
Magdalena Jacques, agente de secrétariat[1]
Kim Dussault, agente de secrétariat[1]

Révision linguistique
Pierre Sénéchal, réviseur linguistique

Référence
Ministère des Forêts, de la Faune et des Parcs, 2017. *Méthodologie des compilations forestières du 4e inventaire écoforestier du Québec méridional : cas particulier des estimations k-NN*, Québec, Ministère des Forêts, de la Faune et des Parcs, secteur des forêts, Direction des inventaires forestiers, 45 p.

---

[1] Direction des inventaires forestiers, ministère des Forêts, de la Faune et des Parcs

AVANT-PROPOS

Dans le cadre de la production des compilations forestières du 4e inventaire écoforestier du Québec méridional (IEQM), plusieurs méthodes ont été utilisées au cours des années. D'une part, les besoins des diverses clientèles ont évolué, celles-ci désirant obtenir des données de qualité à des échelles de plus en plus fines (opérationnelles) et, d'autre part, le nombre de placettes-échantillons à établir devait être réduit tout en maintenant la qualité globale des résultats. C'est ainsi que des changements majeurs aux processus de travail et à la méthode statistique en particulier ont été faits. Ce document de référence, qui correspond au premier volet d'un document plus général décrivant toutes les méthodes de compilation du 4e inventaire, expose la méthode *k*-NN utilisée dans le contexte des compilations forestières par peuplement. Cette méthode est utilisée dans les compilations forestières disponibles depuis 2013. Le tableau de l'avancement des activités de l'inventaire précise les différents territoires où cette méthode a été appliquée.

Ce document s'adresse à toute personne qui désire en connaître davantage sur les fondements de base et le fonctionnement de la méthode *k*-NN, de même que sur son utilisation dans le contexte particulier de l'inventaire forestier québécois. Les utilisateurs des données de compilations forestières seront ainsi en mesure de mieux comprendre la façon dont sont produits les différents résultats de variables dendrométriques. La Direction des inventaires forestiers (DIF) a développé une méthode de travail rigoureuse et efficace qui permet de tirer le maximum des données disponibles dans le but de produire des estimations de variables dendrométriques de qualité.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Processus général et activités

L'inventaire écoforestier du Québec méridional (IEQM) vise à acquérir et à diffuser des connaissances sur les écosystèmes forestiers québécois. Il permet notamment de qualifier et de quantifier la superficie des peuplements forestiers et les volumes marchands bruts de bois sur pied. Le processus de réalisation des activités comporte quatre grandes étapes qui se déroulent sur environ quatre ans (figure 1) :

Figure 1 : Étapes du processus d'inventaire écoforestier du Québec méridional

Voici une brève description de chacune de ces étapes :

Étape 1 : Acquisition des photographies aériennes numériques

Survol du territoire et prise des photos à interpréter;

Étape 2 : Cartographie écoforestière originale

Produite par photo-interprétation des images numériques, elle consiste à délimiter, qualifier et évaluer les superficies des peuplements écoforestiers selon des critères précis;

Étape 3 : Sondage terrestre

Consiste à établir des placettes-échantillons temporaires (PET) dans le but d'acquérir des mesures de variables dendrométriques dans les peuplements cartographiés;

Étape 4 : Compilation forestière

Consiste à associer des variables dendrométriques mesurées dans les placettes à différentes échelles d'agrégation des peuplements de la carte écoforestière, allant des peuplements individuels jusqu'à l'unité de sondage.

Dans la figure 1, l'ajout des données des perturbations naturelles et anthropiques permet de créer la carte écoforestière dite « à jour » qui correspond à la carte écoforestière originale sur laquelle on superpose les différentes couches de perturbations et d'interventions survenues entre l'année de la prise de vue et l'année en cours.

Dans le cadre de la révision du processus global d'inventaire qui s'est échelonnée de 2009 à 2012 avec le projet de l'Approche d'inventaire par peuplement écoforestier (AIPF), on a revu et amélioré plusieurs activités dans le but, notamment, d'augmenter l'efficacité des activités de l'inventaire et d'adapter les produits de l'inventaire aux besoins actuels. Cette révision du processus global de l'inventaire a eu un effet majeur sur le processus de compilation forestière et une refonte de la méthodologie a été menée.

La grande nouveauté dans le processus de compilation est sans contredit le fait que les résultats sont maintenant produits à l'échelle du peuplement écoforestier (compilation par peuplement). Dans l'ancienne approche, la méthodologie utilisée produisait des résultats de compilation par strate regroupée. Ces strates étaient formées en regroupant plusieurs strates cartographiques dont l'appellation cartographique était semblable. Cette méthode était basée sur des fondements statistiques solides, mais a été altérée au fil du temps par l'ajout de diverses procédures qui rendaient laborieuse, voire impossible l'estimation de la précision formelle. On a ainsi revu l'aspect statistique de tout le processus de compilation de façon à respecter de nouveau les fondements de base liés à l'inventaire forestier. On utilise donc maintenant la méthode statistique *k*-NN pour produire les résultats de compilation. Il est à noter que la compilation forestière par peuplement s'applique directement à la carte écoforestière originale, telle qu'elle est diffusée par la Direction des inventaires forestiers (DIF). Aucune modification n'est apportée à cette carte lors du processus de compilation et aucun élément de nature territoriale n'y est intégré.

La nouvelle méthodologie de compilation est applicable aux territoires qui correspondent aux unités de sondage échantillonnées depuis 2011 (voir le Guide d'utilisation de la carte écoforestière et des résultats d'inventaire écoforestier du Québec méridional [MFFP-DIF, 2021]). Pour obtenir de l'information sur la façon d'utiliser les données dendrométriques des territoires sondés avant 2011, veuillez consulter le Guide d'utilisation des données des projets de compilation — Projets des unités de sondage des années 2004-2010 (MRNF-DIF, 2011).

Le présent document s'en veut un de référence qui décrit la méthodologie utilisée pour la réalisation des compilations forestières originales à la DIF. On y décrit la méthode *k*-NN et son application propre aux inventaires forestiers québécois.

1. DESCRIPTION GÉNÉRALE DE LA MÉTHODE *k*-NN

L'acronyme « *k*-NN » signifie « *k Nearest Neighbors* » (*k* plus proches voisins). Au cours des dernières décennies, cette méthode statistique a connu un important gain de popularité en foresterie pour l'estimation d'attributs dendrométriques de peuplements forestiers (Malinen, 2003; Temesgen et coll., 2003; Mäkelä et Pekkarinen, 2004; LeMay et Temesgen, 2005; Packalén et Maltamo, 2007; LeMay et coll., 2008) ainsi que pour l'estimation et la cartographie de divers attributs forestiers à l'aide de photographies aériennes, d'imagerie satellitaire, de données LiDAR, de données géographiques ou d'une combinaison de ces sources (Franco-Lopez et coll., 2001; McRoberts et coll., 2002; Tomppo et Halme, 2004; Maltamo et coll., 2006; Packalén et Maltamo, 2006; McRoberts et coll., 2007; Hudak et coll., 2008). Cette méthode est aujourd'hui employée dans le cadre des inventaires forestiers nationaux, notamment en Suède (Reese et coll., 2003) et en Finlande (Tomppo, 2006).

Un des avantages associés à cette méthode d'imputation est de permettre la prédiction simultanée d'un ensemble de variables réponses (p. ex., volumes par essence) (Moeur et Stage, 1995), ce qui s'avère très utile dans le contexte de l'inventaire forestier pour répondre aux besoins des clientèles. Un autre avantage associé à cette méthode est qu'elle est indépendante de la distribution des variables utilisées, ce qui en fait une méthode non paramétrique (Haara et coll., 1997; Lemay et Temesgen, 2005). De plus, elle permet facilement l'intégration de nouvelles variables explicatives dans les analyses.

Un désavantage important de cette méthode est que certaines analyses peuvent nécessiter beaucoup de temps de calcul (McRoberts et coll., 2006), particulièrement lorsque le nombre d'unités pour lequel des imputations sont désirées est élevé. De plus, le *k*-NN ne permet pas d'extrapoler des valeurs lorsqu'elles sont à l'extérieur de l'étendue des placettes de référence ni d'interpoler des valeurs lorsque des portions de la population sont sous-représentées (Stage et Crookston, 2007). Ainsi, pour obtenir de bonnes estimations dans les peuplements rares, il faut les avoir échantillonnés au préalable. Finalement, les propriétés statistiques de l'approche *k*-NN sont encore mal comprises (Ferland-Raymond, 2010), ce qui rend complexe la formalisation de l'évaluation de l'incertitude des estimations produites.

1.1 Contexte forestier

Dans le contexte de l'inventaire écoforestier du Québec méridional, la méthode *k*-NN consiste à sélectionner, pour chaque peuplement de la carte écoforestière du territoire d'intérêt, les « *k* » peuplements sondés les plus semblables sur la base d'une analyse de similarité entre les peuplements sondés et le peuplement d'intérêt. Ensuite, on calcule la moyenne des données dendrométriques des placettes-échantillons contenues dans les peuplements sondés retenus (*k* voisins) pour estimer chacune des variables dendrométriques d'intérêt (p. ex., volume, surface terrière, nombre de tiges, etc.) dans le peuplement. Une pondération est appliquée lors du calcul de la moyenne des données dendrométriques afin d'accorder une plus grande importance aux

placettes-échantillons localisées dans les peuplements sondés les plus semblables au peuplement d'intérêt.

1.2 Notions de base

Une analyse *k*-NN nécessite toujours la préparation de deux ensembles de données : l'ensemble de référence et l'ensemble cible. L'ensemble de référence correspond aux peuplements sondés localisés sur le territoire d'intérêt. Les variables explicatives sont connues pour toutes les unités de cet ensemble, de même que les variables dendrométriques à estimer (variables réponses). Ces dernières sont obtenues à partir des données des placettes-échantillons implantées sur le territoire. L'ensemble cible, quant à lui, correspond à tous les peuplements de la carte écoforestière pour lesquels on désire obtenir des estimations de variables dendrométriques, y compris les peuplements sondés de l'ensemble de référence. Les variables explicatives associées aux peuplements de cet ensemble sont également connues, mais les variables réponses ne le sont pas. L'objectif de la méthode *k*-NN est d'estimer les variables réponses dans les peuplements de l'ensemble cible à partir des peuplements sondés de l'ensemble de référence.

1.3 Méthode de calcul de la distance

La similarité entre les peuplements se mesure en calculant la « distance » entre chacun des peuplements de la carte et les peuplements sondés. Différentes méthodes de construction de l'espace spectral des variables explicatives faisant partie de l'analyse peuvent être utilisées pour le calcul de cette distance qui n'a rien à voir avec une distance en mètres ou en centimètres.

De façon très générale, les distances « euclidienne », « *raw* » et « de Mahalanobis » sont comparables entre elles en ce sens qu'elles accordent la même importance à toutes les variables explicatives dans l'analyse. Cependant, la distance « *raw* » est calculée à partir des variables brutes, tandis que les distances « euclidienne » et « de Mahalanobis » sont calculées à partir des variables standardisées, ce qui permet généralement d'éliminer l'effet d'échelle entre les variables.

La distance « MSN » (*most similar neighbours*) a, quant à elle, été développée par Moeur et Stage (1995) pour tenir compte du fait que les variables explicatives n'ont généralement pas toutes le même pouvoir explicatif et que l'importance accordée à chacune devrait être variable. Ainsi, en résumant en un nombre réduit de nouvelles variables (variables canoniques) l'information contenue dans les variables explicatives à l'aide de l'analyse de corrélation canonique (ACC) (Rencher, 2002), on réussit à éliminer l'effet des variables ayant un faible pouvoir prédictif. Ce processus s'apparente donc à une sélection de variables, puisque le poids attribué aux variables moins utiles est minimal dans le calcul de distance.

La distance « RandomForest » se différencie des autres mesures de distance par le fait qu'elle se base sur une matrice de similarité et non sur une matrice de distance. Cette méthode permet aussi de pondérer les variables explicatives en fonction de leur importance. Une procédure de base est d'ailleurs disponible pour l'évaluation de l'importance des variables, mais seulement

dans un contexte univarié. Une approche a cependant été développée dans le but d'obtenir une mesure d'importance des variables lorsque la distance « RandomForest » est utilisée dans un *k*-NN en mode multivarié. Par ailleurs, cette méthode est également fonctionnelle avec des variables catégoriques, contrairement aux autres méthodes décrites précédemment.

Finalement, la méthode « *ik*-NN » utilise un algorithme génétique qui permet de définir un vecteur de poids qui sera appliqué aux variables explicatives afin de réduire le biais et l'erreur quadratique moyenne (RMSE). Cette méthode est cependant relativement complexe à implanter et nécessite des temps de traitement informatique qui sont non négligeables.

2. CAS PARTICULIER DE LA DIRECTION DES INVENTAIRES FORESTIERS

Dans le contexte de refonte de la méthode de compilation forestière, différents programmes et fonctions utilisés pour la production des estimations *k*-NN ont été développés dans le langage du logiciel statistique R (R Core Team, 2007). Cela a permis de bénéficier du travail conjoint du Service des forêts du ministère de l'Agriculture des États-Unis (Forest Service — USDA) et de l'Université du Michigan qui ont développé, dans le même langage, la bibliothèque de fonctions « *yaImpute* » destinée à réaliser des imputations *k*-NN dans un contexte d'inventaire forestier (Crookston et Finley, 2008). Cette bibliothèque offre la possibilité d'utiliser différentes méthodes de calcul de l'espace spectral et utilise un algorithme efficace pour la recherche des plus proches voisins. Pour les besoins de la DIF, certaines fonctions de cette bibliothèque ont donc été empruntées et adaptées au contexte forestier québécois. Pour faciliter son utilisation, la bibliothèque de fonction développée par la DIF a été structurée de façon modulaire, c'est-à-dire que chacune des fonctions peut être utilisée indépendamment et au moment voulu par l'utilisateur, dans la mesure où la structure des jeux de données est la bonne. Cela donne beaucoup de flexibilité à l'outil, en plus de le rendre plus convivial et relativement facile à suivre et à comprendre.

Puisque l'objectif est de produire des estimations de variables dendrométriques à l'échelle des peuplements écoforestiers, les jeux de données des ensembles de référence et cible ont été produits à cette échelle, selon l'attribut « géocode » (identifiant unique d'un peuplement écoforestier). Dans les cas où plus d'une placette-échantillon a été implantée dans un même peuplement, on a calculé la moyenne des données des placettes pour chacune des variables réponses.

Concernant la méthode de calcul de la distance, c'est la méthode « MSN » qui a été retenue, car elle offre un outil de sélection de variables qui permet de pondérer l'importance des variables explicatives. Le choix de cette méthode repose sur les résultats d'un rapport interne produit par Ferland-Raymond (2011) dans lequel les différentes méthodes de calcul de la distance ont été comparées. Les résultats démontrent que la méthode « MSN » est l'une des plus précises, tout en produisant des biais qui sont relativement faibles.

2.1 Variables explicatives

L'analyse de similarité entre les peuplements sondés et les peuplements de la carte écoforestière pour lesquels on désire obtenir des estimations de variables dendrométriques s'effectue sur la base d'un ensemble de variables explicatives. Pour que l'analyse *k*-NN soit fonctionnelle, toutes ces variables doivent être disponibles pour tous les peuplements de l'ensemble cible, lesquels incluent les peuplements de l'ensemble de référence (peuplements sondés). Il ne peut donc pas y avoir de données manquantes dans les différents jeux de données. Dans les analyses, on utilise plusieurs types de variables explicatives, soit les variables associées à la carte écoforestière

(variables de base), les variables associées aux données de télédétection et les variables auxiliaires qui regroupent les variables géographiques, climatiques et écologiques.

2.1.1 Carte écoforestière

Les variables de la carte écoforestière correspondent aux variables déduites des photos aériennes par le photo-interprète. Ces variables sont décrites dans la « Norme de stratification écoforestière — Quatrième inventaire écoforestier du Québec méridional » (MFFP-DIF, 2015). Comme la stratification a évolué au cours du quatrième inventaire, certains territoires ont été cartographiés suivant la norme de stratification « initiale » et d'autres, suivant la norme de stratification « AIPF ». Les principales différences entre ces stratifications concernent les essences, la densité du couvert et la hauteur des peuplements. Il est à noter que seulement trois territoires ont été cartographiés suivant la norme de stratification « initiale » et compilés avec la méthode *k*-NN. Par ailleurs, deux territoires ont été cartographiés suivant une stratification dite « mixte », qui est une stratification transitoire entre les stratifications « initiale » et « AIPF ». Tous les autres territoires qui ont été ou seront compilés avec la méthode *k*-NN (à partir du sondage 2011) ont été ou seront cartographiés suivant la stratification « AIPF » (voir la carte de disponibilité des résultats de compilations forestières du 4e inventaire).

Dans la norme de stratification « initiale », la composition du couvert forestier se définit à partir du groupement d'essences, lequel permet de nommer jusqu'à un maximum de trois essences. La détermination du groupement d'essences d'un peuplement est fonction de la proportion de la surface terrière totale (surface terrière relative) de chacune des essences qui le compose et suit une série de règles qui sont décrites dans la norme. Par ailleurs, toujours selon cette norme, la densité du couvert forestier est évaluée en classes de 20 % (variable catégorique), et la hauteur des peuplements est évaluée en classes de 5 m (variable catégorique).

Dans la norme de stratification AIPF, les trois attributs mentionnés au paragraphe précédent sont définis avec plus de précision. Ainsi, le concept de « groupement d'essences » n'existe plus. La composition de chaque peuplement est plutôt décrite en identifiant chaque essence et en déterminant la surface terrière relative que chacune d'elle représente (10 % près). Selon cette norme, un maximum de sept essences différentes peut être identifié. Quant à la densité du couvert, elle est évaluée en classes de 10 % (variable numérique), tandis que la hauteur des peuplements est évaluée au mètre près. Par ailleurs, chacun des deux étages des peuplements de structure étagée est décrit individuellement.

2.1.1.1 Conversion en variables numériques

La méthode de calcul de la distance retenue pour les analyses exige que les variables explicatives soient en format numérique plutôt qu'en format catégorique comme le sont, à titre d'exemples, les groupements d'essences ainsi que les classes de densité et de hauteur dans la stratification « initiale » du quatrième inventaire. Afin de satisfaire à cette exigence, un travail de conversion en valeurs numériques des variables représentant le groupement d'essences, la densité, la

hauteur, l'âge (pour les codes de peuplements inéquiens ou irréguliers), la présence de perturbations et l'épaisseur du dépôt de surface, a été nécessaire.

Dans le cas des groupements d'essences, ils ont été convertis en essences individuelles (maximum de trois essences) et on a attribué à chacune une proportion en surface terrière (milieu de classes) basée sur les règles d'attribution des groupements d'essences décrites dans la « Norme de stratification écoforestière — Quatrième inventaire écoforestier du Québec méridional » (MRNF-DIF, 2015). Dans le cas de l'âge, on réalise une analyse à partir des arbres-études implantés dans les peuplements inéquiens et dans les peuplements de structure irrégulière du territoire considéré afin de déterminer l'âge moyen de ces peuplements dont les classes d'âge sont « JIN », « JIR » ou « VIN », « VIR ». Les âges moyens calculés varient donc d'un territoire de compilation à l'autre. Par ailleurs, les perturbations moyennes naturelles ou anthropiques ont, quant à elles, été traduites par une variable binaire, soit 0 (absence) ou 1 (présence). Les perturbations d'origine ne sont pas considérées dans les analyses. Finalement, en ce qui a trait aux classes de densité et de hauteur et à l'épaisseur du dépôt de surface, des tableaux de conversion en valeurs numériques sont présentés à l'annexe I.

Dans la stratification AIPF, chacun des étages des peuplements de structure étagée est décrit individuellement. Les essences et leur surface terrière relative associée, la densité, la hauteur et l'âge sont donc définis pour chaque étage. La méthode *k*-NN requiert cependant que chaque unité d'analyse n'ait qu'une seule valeur pour chacun de ces attributs.

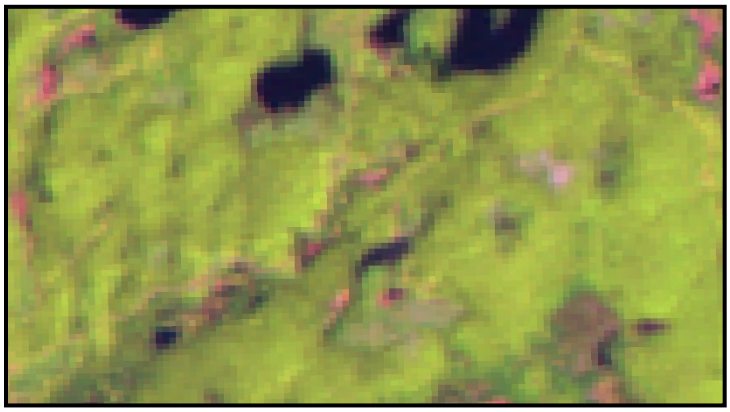
La conversion des peuplements de structure étagée en peuplements uniétagés s'effectue en utilisant une variable photo-interprétée (ET_DOMI) qui permet de désigner directement l'étage dominant en surface terrière et de déterminer son importance relative par rapport au peuplement global. Ainsi, si l'étage supérieur est désigné comme étant l'étage dominant, son importance relative est de 60 %, et l'importance de l'étage inférieur est donc de 40 %. À l'inverse, si l'étage inférieur est l'étage dominant, son importance relative est de 60 %. Dans le cas où le photo-interprète jugerait qu'aucun des deux étages ne domine en surface terrière, il attribue à chaque étage une importance relative égale de 50 %. Il ne reste plus qu'à ventiler la surface terrière relative de chaque essence en fonction de l'importance de chaque étage et à additionner les deux étages. La hauteur et l'âge des peuplements de structure étagée sont également convertis suivant cette même méthodologie.

2.1.2 Variables de télédétection

Les variables explicatives provenant des données de télédétection peuvent être de différents types. Certaines sont extraites d'images satellite (mosaïque d'images) provenant de différents types de satellites (Landsat, RapidEye, etc.), tandis que d'autres proviennent du LiDAR (*Light Detection and Ranging*). Jusqu'à présent, seules des données associées au Landsat ont été utilisées dans les compilations *k*-NN, principalement en raison de leur gratuité et de leur facilité d'accès. Les paragraphes suivants décrivent tout de même les principales caractéristiques des satellites Landsat et RapidEye ainsi que du LiDAR.

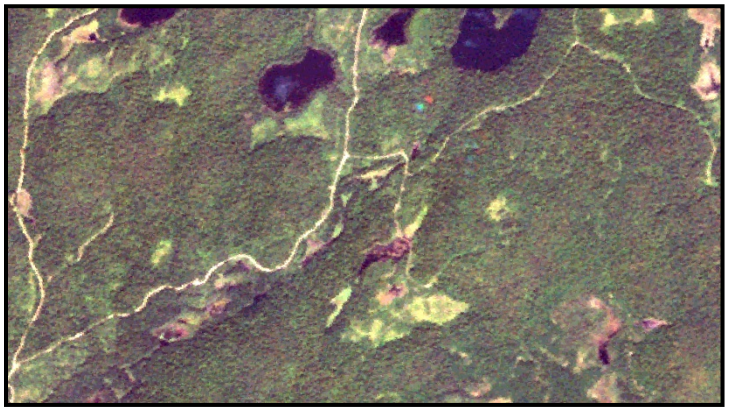
Au cours des dernières années, des changements importants ont été apportés en ce qui a trait à l'acquisition des données Landsat. En effet, depuis 1984, c'est le satellite Landsat 5 TM (*Thematic Mapper*) qui était utilisé. Il n'est cependant plus en service depuis novembre 2011 et a été remplacé par le satellite Landsat 8 au début de 2013. Dans le cadre des compilations forestières du quatrième inventaire utilisant la méthode *k*-NN, les deux versions du satellite ont été utilisées. Voici les principales caractéristiques de chacune. Le Landsat 5 TM produit des images dont la résolution des bandes spectrales est de 30 m (figure 2), sauf dans le cas de la bande 6 qui correspond à la bande thermique. Cette bande n'est généralement pas utilisée, car elle n'est pas très efficace pour distinguer la végétation et sa résolution est beaucoup plus grossière (120 m, rééchantillonnée à 30 m). Au total, le satellite Landsat 5 TM produit sept bandes spectrales. Le Landsat 8 OLI (*Operational Land Imager*), quant à lui, produit onze bandes spectrales dont la résolution est également de 30 m, sauf dans le cas des bandes 10 et 11 (100 m rééchantillonnée à 30 m) et de la bande panchromatique (bande 8) qui a une résolution de 15 m.

Figure 2 : Image Landsat



Quant au RapidEye, il est composé d'une constellation de cinq satellites en orbite autour de la Terre. Ce système a vu le jour en 2008 et est propriété de l'entreprise RapidEye AG. Les satellites sont placés à égale distance les uns des autres et l'enregistrement des données se fait sur une bande de 77 km de largeur, à 630 km d'altitude. Ce système permet d'obtenir rapidement des images de haute résolution à faibles coûts. Il se distingue du Landsat par sa résolution plus fine, 5 m, et par le fait que les images obtenues n'ont que cinq bandes spectrales (figure 3). Les données provenant de ce type de satellite pourraient ainsi être une option intéressante aux données Landsat dans le cas où la qualité de ces dernières ne serait pas satisfaisante (p. ex., trop de nuages).

Figure 3 : Image RapidEye



Les variables de télédétection calculées à partir de ces images sont la moyenne et l'écart-type des pixels dont le centroïde est situé au sein d'une unité d'analyse, soit le peuplement écoforestier. Des variables sont produites pour chacune des bandes spectrales à l'aide de l'outil ZonalStat du logiciel ArcGIS. Puisque le territoire ne contient pas seulement des pixels de forêts, une gestion particulière des pixels non forestiers (nuages, chemins, interventions) doit être faite afin d'exclure ces pixels des calculs de moyennes et d'écarts-types pour ne pas ajouter inutilement de bruit dans les résultats. Pour ce faire, on crée un masque en utilisant les différentes bandes disponibles. Ainsi, les pixels dont la signature spectrale correspond à celle de nuages, de leur ombre ou de chemins peuvent être exclus des calculs de moyennes et d'écarts-types. En plus des variables par bande, on calcule également l'indice de végétation NDVI (*Normalized Difference Vegetation Index*) à partir des bandes 3 (rouge) et 4 (infrarouge) dans le cas du Landsat 5 TM et des bandes 4 et 5 dans le cas du Landsat 8 OLI. Dans le cas du RapidEye, ce sont les bandes 3 (rouge) et 5 (infrarouge) qui sont utilisées. Le ratio de cet indice est de la forme :

$$NDVI = \frac{IR - R}{IR + R}$$

Le LiDAR, quant à lui, permet entre autres de déterminer de façon très précise la hauteur et la structure du couvert forestier en produisant un nuage de points en trois dimensions. Plusieurs variables explicatives telles que la moyenne, les valeurs minimale et maximale, les centiles, etc. peuvent être extraites de cette technologie pour une unité d'analyse donnée. Certains indices représentant par exemple le pourcentage de points dont la hauteur varie de 0 à 1 m, 1 à 2 m, etc. ou un indice de rugosité peuvent également être produits. Un programme d'acquisition est en cours afin d'obtenir la donnée sur l'ensemble du territoire forestier québécois d'ici à 2022. Le premier territoire qui utilisera cette donnée dans les compilations *k*-NN est l'unité d'aménagement (UA) 02451. Les compilations subséquentes utiliseront le LiDAR en fonction de la disponibilité de la donnée.

2.1.3 Variables auxiliaires

D'autres variables décrivant la localisation géographique et les conditions de croissance de

chacune des unités d'analyse sont également utilisées dans la modélisation. Le terme « variables auxiliaires » réfère aux variables autres que celles provenant de la stratification écoforestière et de la télédétection. Elles sont de trois types : géographique, climatique et écologique.

2.1.3.1 Variables géographiques

Les variables géographiques retenues sont généralement disponibles sous forme de rasters (images avec comme unité de base le pixel), ce qui implique qu'on doit calculer la moyenne des valeurs des pixels dont le centroïde est situé au sein d'un peuplement donné afin d'obtenir une seule valeur de la variable en question par peuplement.

C'est le cas pour les variables suivantes : altitude (m), pente (%), indice d'exposition au vent et deux variables décrivant l'effet combiné de la pente et de l'exposition (pente*cos[exposition] et pente*sin[exposition]) (Stage, 1976). Pour créer ces deux dernières variables, l'exposition (en degrés) est transformée en radians. Par ailleurs, l'indice d'exposition au vent, aussi nommé « topex », représente la somme des angles par rapport à l'horizon selon huit points cardinaux à une distance de 500 m d'un point donné (Ruel et coll., 2002). Lorsque cette somme est positive, l'unité d'analyse se trouve en bas de pente ou dans une vallée (moins exposée au vent), tandis que lorsque la somme des angles est négative, l'unité d'analyse se trouve en haut de pente ou sur un sommet (plus exposée au vent). Deux autres variables, soit la latitude et la longitude, sont également incluses dans les analyses. Dans ce cas, les valeurs correspondent au centroïde de chaque peuplement.

2.1.3.2 Variables climatiques

Trois variables climatiques, soit la température annuelle moyenne, les précipitations moyennes et les précipitations durant la saison de croissance, sont considérées dans les analyses. Ces variables sont calculées à l'aide de l'outil logiciel BioSIM version 9.5.2 (Régnière et Saint-Amant, 2008), pour chacune des unités d'analyse. Encore une fois, comme ces variables sont disponibles sous forme de rasters, on doit calculer la moyenne des valeurs des pixels qui sont localisés dans chaque peuplement pour obtenir une valeur par peuplement. Ces variables ont été retenues, puisqu'il a été démontré qu'elles avaient une influence importante sur la croissance des espèces forestières. En effet, ces trois variables ont été intégrées dans le modèle de croissance « Artémis » développé par la Direction de la recherche forestière (DRF) (Fortin et Langevin, 2010).

2.1.3.3 Variable écologique

Finalement, un indice de productivité est calculé en utilisant la méthode développée par l'équipe en écologie de la DIF dans le cadre du projet sur le zonage forestier (Cyr et coll., 2010). Ce projet, qui vise à répertorier les sites les plus productifs pour y faire de l'aménagement intensif, se base principalement sur les valeurs d'indice de qualité de station (IQS) et de croissance en surface terrière potentiels obtenues à partir des études d'arbre réalisées lors des sondages terrestres.

3. QUALITÉ DES RÉSULTATS

L'évaluation de la qualité des résultats se fait par l'entremise de deux critères statistiques, soit le biais et le T^2 (McRoberts, 2012). Le biais, qui est un indicateur d'une tendance à surestimer ou sous-estimer, correspond à la somme des erreurs d'estimation (différence entre la valeur estimée [$\hat{y}$] et la valeur observée [y] d'une variable dendrométrique donnée). On calcule cette statistique pour chacune des placettes (et pour chaque essence), car c'est l'unité la plus proche du peuplement forestier pour laquelle les valeurs observées des différentes variables dendrométriques sont connues. Quant au T^2, il est une approximation de la proportion de variance expliquée par le modèle *k*-NN. Il s'apparente donc à la valeur de R^2 dans une régression classique. Lorsque le T^2 est égal à 0 %, cela indique que le modèle ne fait pas mieux que de prédire la moyenne des valeurs observées. À l'inverse, lorsque le T^2 est égal à 100 %, cela indique que l'on prédit exactement la valeur observée de la variable dendrométrique dans chacune des placettes.

Le biais moyen (m^3/ha) se calcule donc à l'aide de :

$$\overline{B} = \frac{\sum_{i=1}^{I} (\hat{y}_i - y_i)}{I}$$

où i = (1, 2, 3, …, I) correspond à la placette cible. Le T^2 se calcule quant à lui par :

$$T^2 = \left(\frac{SS_{mean} - SS_{res}}{SS_{mean}} \right) \times 100$$

où l'erreur quadratique moyenne (MSE) provient de :

$$SS_{res} = \sum_{i=1}^{I} (y_i - \hat{y}_i)^2$$

et l'erreur d'imputation par la moyenne :

$$SS_{mean} = \sum_{i=1}^{I} (y_i - \bar{y})^2$$

où $\bar{y}$ correspond à la moyenne des observations.

Enfin, on calcule également le T^2 généralisé, qui permet d'obtenir une valeur de T^2 pour l'ensemble des essences. Cette statistique se définit par :

$$T^2 = \sum_{j=1}^{J} (T_j^{\,2} \times w_j)$$

où j = (1, 2, 3, …, J) est une essence parmi J essences et w correspond au volume observé (poids).

Il est important de noter que la valeur minimale du T^2 est contrainte à 0 lorsque le résultat de la formule est négatif.

4. DÉFINITION DES PARAMÈTRES DE LA MÉTHODE *k*-NN

Lors de l'application de la méthode *k*-NN, plusieurs paramètres doivent être déterminés afin de permettre la production des estimations *k*-NN. Ces paramètres sont : les variables d'optimisation incluses dans les analyses, le nombre de « plus proches voisins » à considérer (*k*) et la valeur de *d* (exposant de l'inverse de la distance), qui permet de mettre davantage de poids sur les peuplements sondés les plus semblables parmi les *k* plus proches voisins sélectionnés dans le calcul des estimations de variables dendrométriques (moyenne pondérée).

4.1 Variables d'optimisation

Les variables d'optimisation correspondent aux variables réponses qui seront considérées dans l'analyse de corrélation canonique réalisée avec la méthode de calcul de la distance retenue (« MSN »). Puisque, finalement, on vise principalement à produire des estimations de volumes marchands bruts moyens par essence, on détermine les variables d'optimisation en fonction de cet attribut. L'objectif est donc de déterminer quelles sont les essences les plus importantes sur le territoire afin d'éviter de tenir compte d'essences qui ne représentent qu'une proportion négligeable du volume total, ce qui entraînerait une réduction de la qualité des estimations des essences principales. Pour ce faire, on calcule à partir des données des placettes (ensemble de référence) l'importance relative de chacune des essences en fonction du volume marchand brut et on ne retient que celles qui forment la majorité du volume. À titre indicatif, une essence qui représente moins de 1 % (essence rare ou marginale) du volume total n'est généralement pas retenue dans les analyses. Par contre, dans certains cas, elle pourrait l'être si une révision à la hausse (ajout de placettes-échantillons) avait été faite pour cette essence lors de la planification du sondage terrestre.

À titre d'exemple, le tableau suivant présente l'importance relative en volume des essences dans l'unité d'aménagement 08665 (tableau 1). Dans cet exemple, seules les cinq essences les plus importantes ont été retenues comme variables d'optimisation. Ces cinq essences représentent près de 99 % du volume total.

Tableau 1 : Importance des essences de l'UA 08665 en matière de volume marchand brut moyen et nombre de fois que l'essence est présente dans une PET

Essence	Volume moyen		Nombre de placettes avec l'essence (sur 300 PET)
	(m³/ha)	%	
EPN	75,40	69,5	287
PET	14,12	13,0	98
PIG	10,96	10,1	96
SAB	4,58	4,2	112
BOP	2,27	2,1	87
MEL	0,94	0,9	23
EPB	0,27	0,3	8

4.2 Paramètres *k* et *d*

Les deux autres paramètres (*k* et *d*) sont déterminés simultanément. En effet, une fois que toutes les données des ensembles de référence et cible ont été structurées à l'aide des fonctions contenues dans le « *package* » R développé par la DIF, on réalise les imputations *k*-NN pour une série de combinaisons de valeurs de *k* et de *d*. On reporte ensuite les valeurs des statistiques désirées (biais, erreur quadratique moyenne [MSE ou EQM] ou T^2) obtenues sur des graphiques de façon à pouvoir déterminer la combinaison de paramètres qui permet de minimiser (biais, MSE) ou maximiser (T^2) les critères statistiques pour chacune des essences (p. ex., figure 4 et figure 5). Par défaut, la fonction présente le biais et le MSE. Lors de l'analyse visuelle des graphiques par essence, on doit accorder une plus grande importance aux essences principales sur le territoire pour ne pas biaiser le choix des paramètres *k* et *d*, car les mêmes valeurs seront appliquées à toutes les essences dans les analyses subséquentes.

Afin de faciliter la prise de décisions, on produit également un graphique qui présente le résultat global (« cote ») des différentes combinaisons de paramètres *k* et *d* (figure 6). Cet indice combine les résultats des différentes essences et est pondéré pour tenir compte à la fois de l'importance de chacune d'elles (volume marchand brut moyen) et du poids que l'on veut donner à chacun des critères statistiques dans la sélection des paramètres *k* et *d*. L'objectif ici est de déterminer la combinaison de paramètres qui maximise la cote.

Il est à noter que la détermination de ces deux paramètres s'effectue une deuxième fois dans le processus de compilation, à la suite du processus de sélection de variables (voir section suivante). Cela permet d'ajuster ces paramètres afin de prendre en considération le nouvel ensemble de variables explicatives qui a été conservé dans les analyses.

Figure 4 : Exemple de sélection des paramètres *k* et *d* pour l'épinette noire dans l'UA 08665

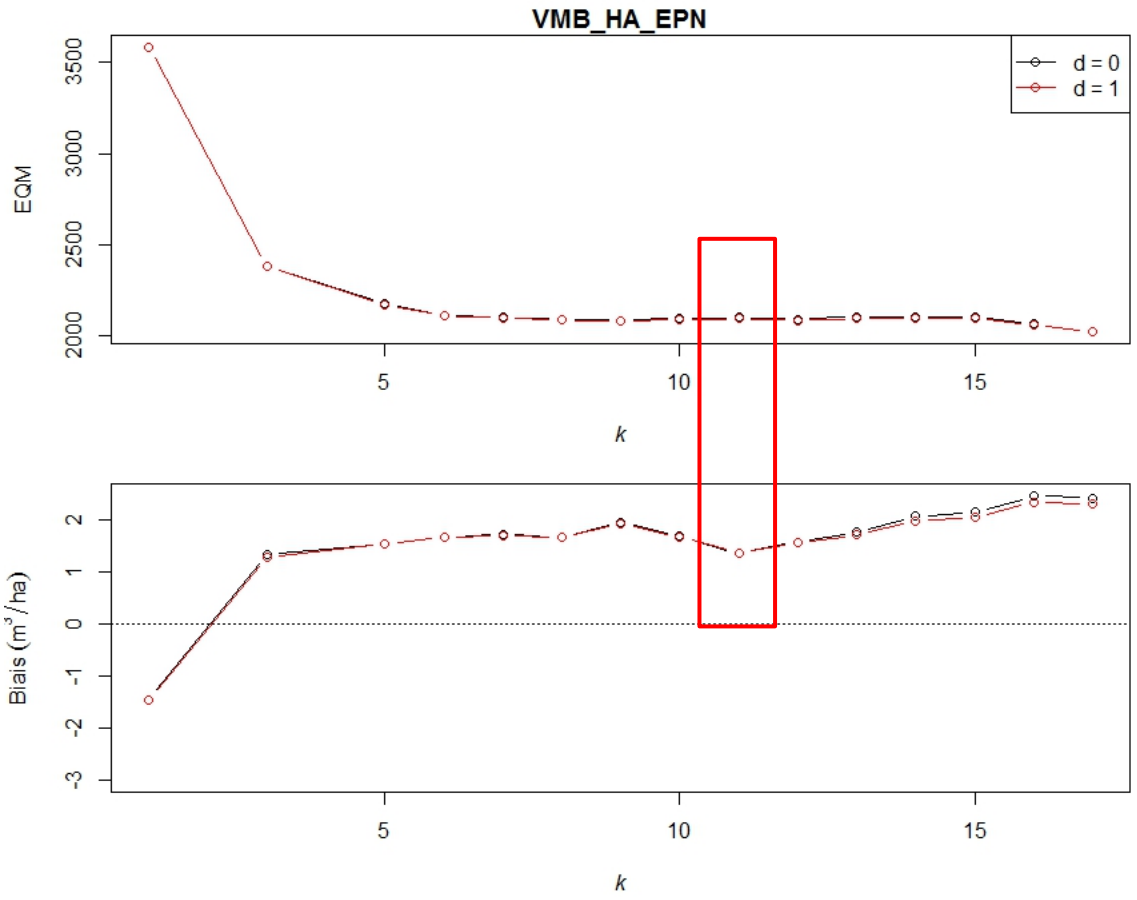


Figure 5 : Exemple de sélection des paramètres *k* et *d* pour le peuplier faux-tremble dans l'UA 08665

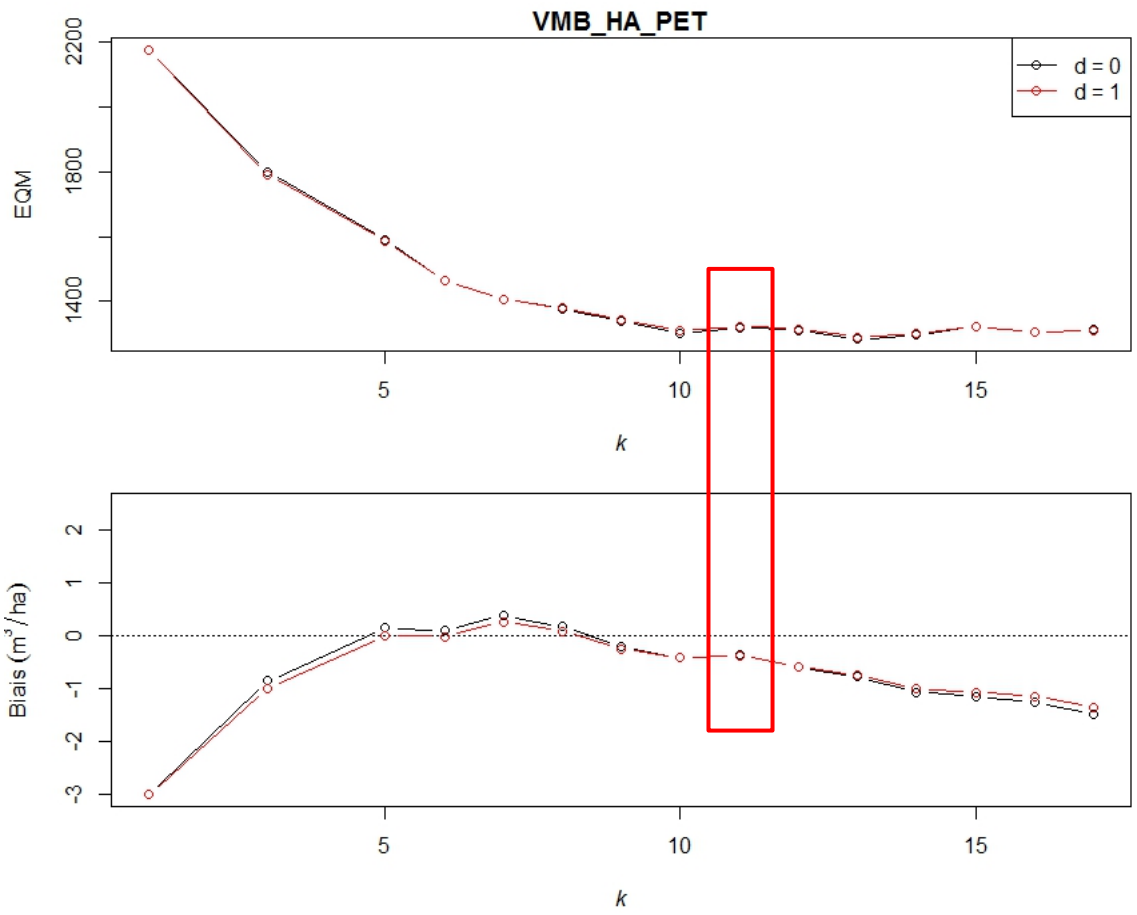
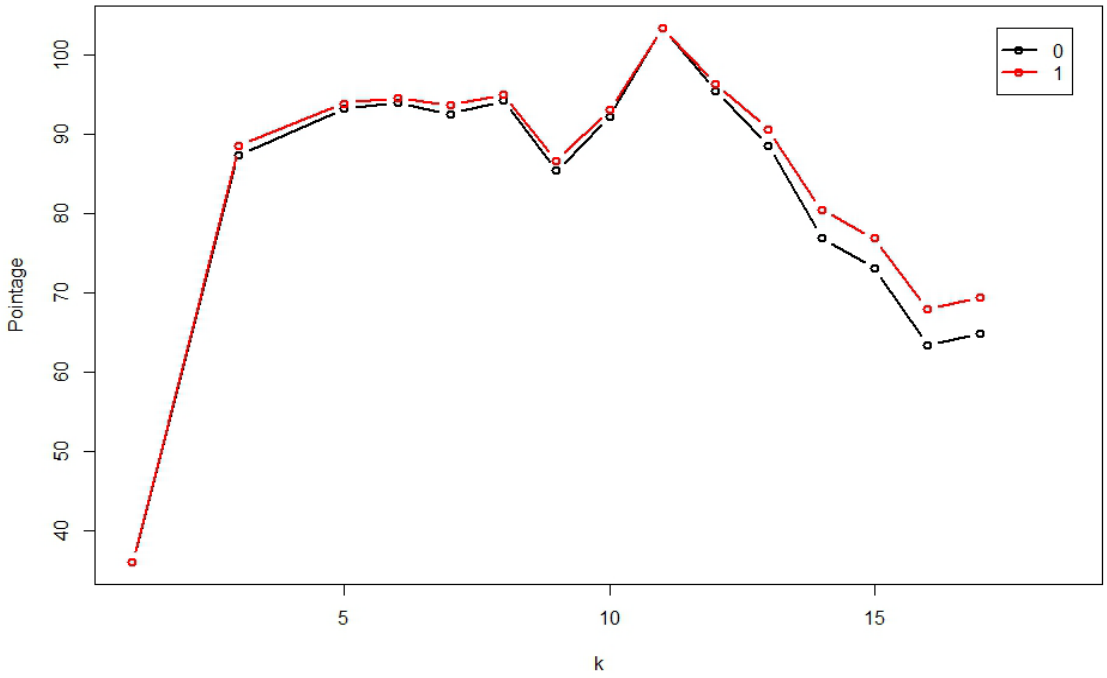


Figure 6 : Indice global des résultats liés aux combinaisons des paramètres k-d (essences et statistiques de comparaison)

Dans ce cas-ci, des poids égaux de 0,5 ont été appliqués aux deux critères statistiques (biais et erreur quadratique moyenne).



5. OPTIMISATION DE LA MÉTHODE *k*-NN

Deux techniques principales d'optimisation sont utilisées dans le processus de compilation afin d'optimiser la qualité des résultats produits par la méthode *k*-NN. Ces techniques sont simples et plutôt objectives, ce qui facilite grandement leur application. Cependant, préalablement à l'utilisation de ces techniques, certaines analyses sont effectuées afin d'éliminer *a priori* les variables qui n'ont aucun pouvoir explicatif ou qui présentent des problèmes de corrélation. Ainsi, on effectue une analyse qui vise à déterminer les variables (explicatives ou réponses) qui n'ont pas de variance (p. ex., la surface terrière relative du tilleul d'Amérique est la même pour tous les géocodes) et, donc, pas de pouvoir prédictif de façon à les éliminer des jeux de données. Par ailleurs, on réalise également une analyse qui permet de détecter la multicolinéarité dans les données, c'est-à-dire qu'il existe une relation linéaire entre plusieurs variables. Comme ces variables sont très corrélées entre elles et donc expliquent la même chose dans l'analyse, on élimine l'une d'elles pour ainsi éliminer les problèmes de multicolinéarité.

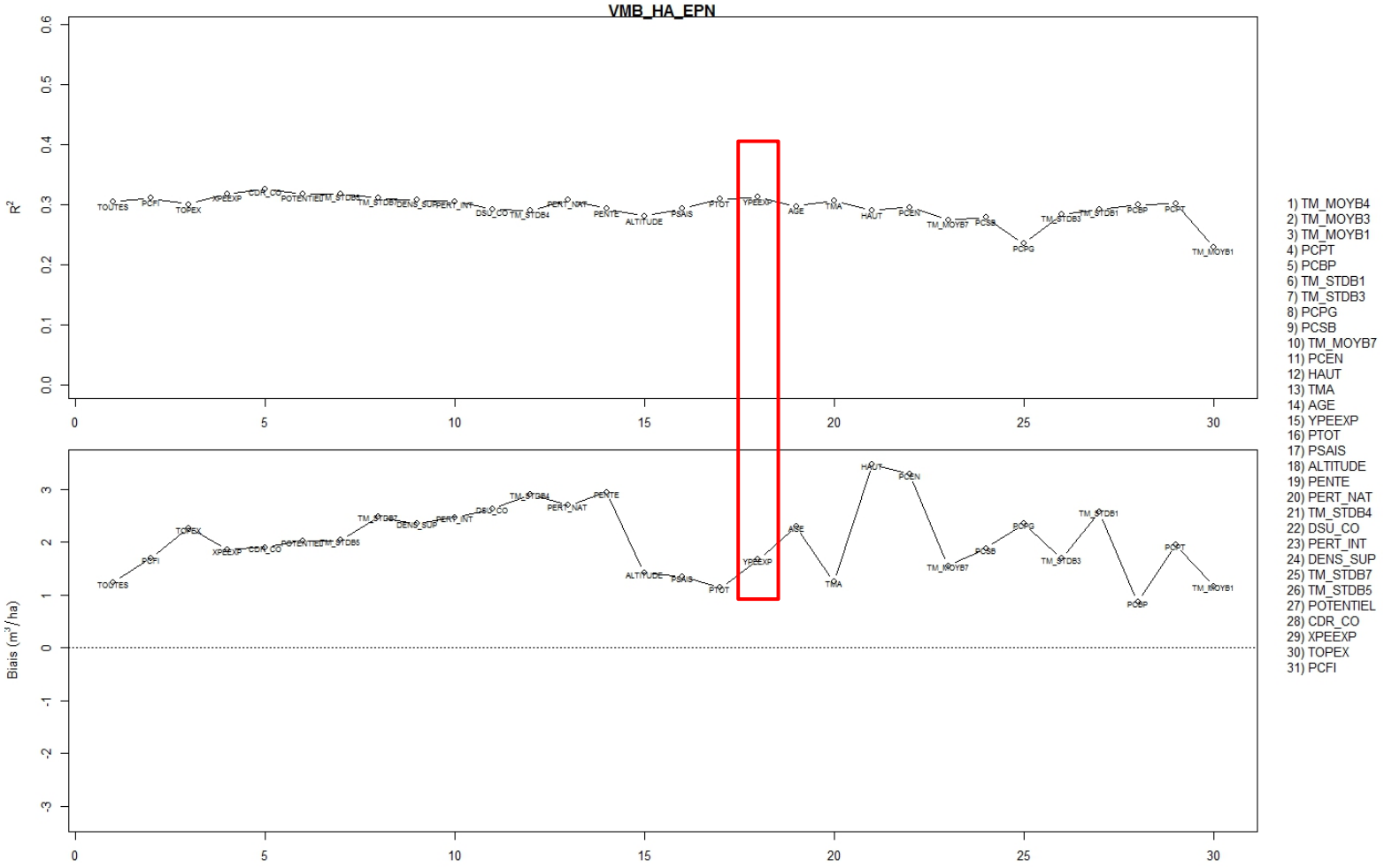
5.1 Sélection de variables

Cette technique d'optimisation consiste à sélectionner des variables explicatives en utilisant la méthode d'exclusion, qui élimine les variables qui semblent peu importantes pour la prédiction des variables dendrométriques d'intérêt. Même si la méthode de calcul de la distance utilisée lors de l'application de la méthode *k*-NN (« MSN ») prévoit déjà une procédure qui s'apparente à une sélection de variables en attribuant un poids à chacune des variables explicatives en fonction de leur pouvoir de prédiction, on juge qu'il est tout de même avantageux d'essayer de réduire le nombre de variables explicatives au lieu de simplement attribuer un poids minimal à un grand nombre de variables qui ont un faible pouvoir explicatif. En fait, un nombre élevé de variables explicatives n'améliore pas toujours les capacités de prédiction du *k*-NN. Dans certains cas, il est même possible d'observer une augmentation de l'erreur de prédiction (McRoberts et coll., 2002).

Plus concrètement, la méthode d'exclusion est un processus itératif qui permet d'éliminer au fur et à mesure les variables les moins importantes dans l'analyse. Cette méthode commence par faire une analyse *k*-NN avec le modèle complet, c'est-à-dire en incluant toutes les variables explicatives. À la suite des résultats obtenus à l'aide de l'ACC, la variable la moins importante est retirée et une autre analyse *k*-NN est effectuée avec les variables explicatives restantes. Ce processus est répété jusqu'à ce qu'il ne reste que deux variables dans l'analyse. Deux graphiques présentant l'évolution des statistiques (biais et T^2) en fonction de l'ordre d'élimination des variables explicatives sont ensuite produits pour chacune des variables d'optimisation (p. ex., figure 7). L'analyse de ces graphiques permet de déterminer l'endroit où le biais est minimisé et le T^2 est maximisé, et ainsi de déterminer les variables à conserver dans l'analyse *k*-NN finale. Encore une fois, lors de cette analyse visuelle, on accorde davantage de poids aux variables d'optimisation qui sont les plus importantes en matière de volume. La liste de variables à droite de la figure indique l'ordre d'exclusion des variables, la dernière dans la liste étant la première à avoir été éliminée.

Figure 7: Exemple de sélection de variables à l'aide de la méthode d'exclusion dans l'UA 08665

Il est à noter que le R^2 sur l'axe des y correspond en réalité à la statistique T^2 décrite à la section 3.

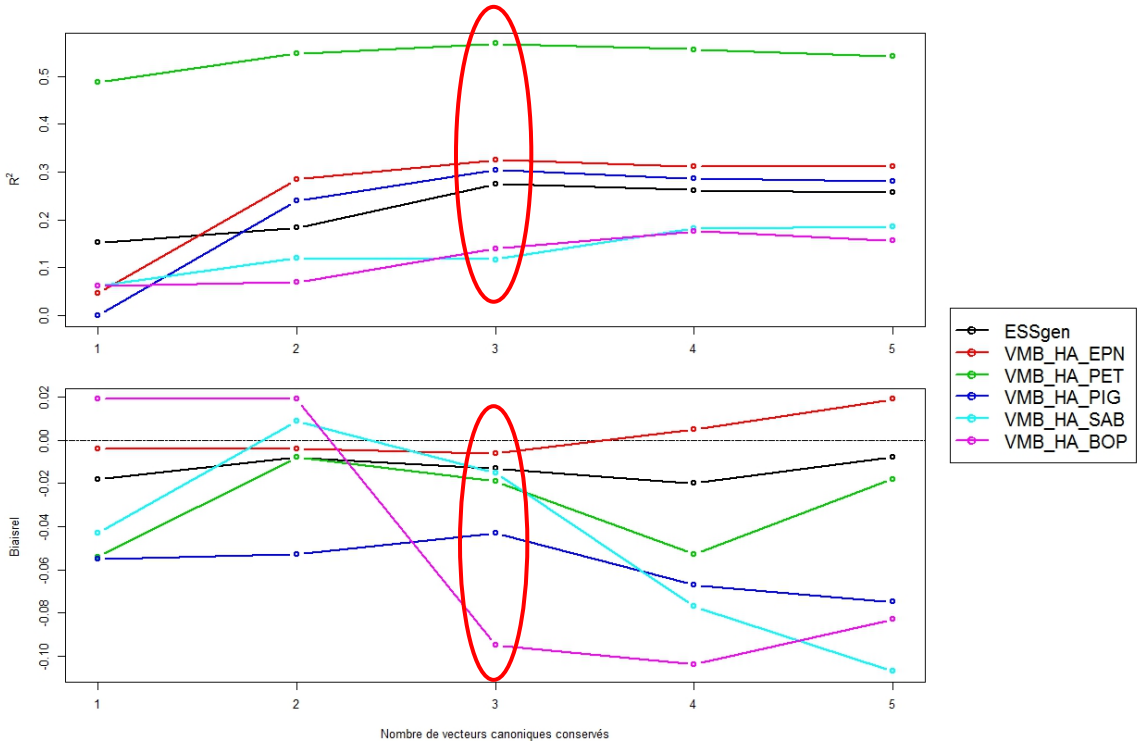


5.2 Nombre de variables canoniques conservées

La deuxième technique utilisée dans le but d'optimiser la méthode *k*-NN consiste à déterminer le nombre de variables canoniques à conserver dans l'analyse de corrélation canonique (ACC dans la méthode de calcul de la distance « MSN »). Une première possibilité serait de conserver l'ensemble des variables canoniques et prendre $C = \min(p, q)$, où p correspond au nombre de variables explicatives et q, au nombre de variables réponses utilisées dans l'analyse (variables d'optimisation). Toutefois, il est fort probable que très peu d'information soit contenue dans les dernières variables canoniques, ce qui les rend inutiles. À l'inverse, on pourrait ne conserver que la première variable canonique. Mais le risque de perdre certains renseignements contenus dans les autres variables est grand. Par défaut, la bibliothèque *yaImpute* (Crookston et Finley, 2008) fait des tests statistiques et ne conserve que les corrélations canoniques qui sont significativement supérieures à 0. Cette façon de faire ne semble cependant pas optimale considérant l'évolution du T^2 et du biais (biais relatif dans ce cas-ci) en fonction du nombre de variables canoniques conservées pour certaines essences (figure 8). En effet, bien que l'influence du nombre de variables canoniques sur la variabilité des valeurs de T^2 soit relativement faible, sauf pour des nombres de variables canoniques très faibles, on observe généralement une diminution du biais relatif avec une réduction du nombre de variables canoniques. Dans l'exemple de la figure 8, trois vecteurs canoniques ont été conservés à la suite de l'analyse visuelle des résultats. On observe en effet que les valeurs de T^2 demeurent stables pour de cinq à trois vecteurs, tandis que le biais relatif tant à diminuer pour trois des cinq essences principales (EPN, PIG et SAB). Le biais relatif du peuplier faux-tremble (deuxième essence en importance) demeure, quant à lui, stable pour de cinq à trois vecteurs canoniques.

Figure 8 : Exemple de graphique utilisé pour la détermination du nombre de vecteurs canoniques à conserver dans le modèle *k*-NN

Il est à noter que le R^2 sur l'axe des y correspond en réalité à la statistique T^2 décrite à la section 3.



5.3 Cas particuliers

Par ailleurs, en plus des deux techniques d'optimisation décrites ci-dessus, des analyses additionnelles peuvent être réalisées dans certains territoires, au besoin, en raison de caractéristiques particulières du milieu, telles que la dimension, la forme ou la localisation géographique du territoire à compiler. Par exemple, dans le cas d'UA de grandes superficies situées dans le nord de la zone d'inventaire intensive (p. ex., UA de la région 02 (Saguenay – Lac-Saint-Jean), on a réalisé plusieurs compilations *k*-NN en divisant le territoire sur la base de considérations écologiques. Comme la forme de ces UA est souvent très allongée (axe nord-sud) et recoupe plus d'un type de forêts (p. ex., sapinière et pessière), et sachant que l'écologie joue un rôle majeur dans la croissance des arbres, on a réalisé, en plus de la compilation *k*-NN couvrant l'ensemble du territoire, une compilation *k*-NN qui ne couvre que la portion nord du territoire (pessière) et une qui ne couvre que la portion sud (sapinière). On a ensuite comparé les critères statistiques associés aux estimations *k*-NN par essence, obtenus en utilisant toutes les placettes du territoire, avec ceux obtenus en utilisant seulement les placettes localisées dans la sapinière ou seulement celles localisées dans la pessière. Cette comparaison a permis de déterminer la méthode produisant les meilleurs résultats.

Dans d'autres cas, c'est le processus inverse qui a été analysé. Ainsi, au lieu de scinder un territoire en plusieurs parties, on a plutôt regroupé plusieurs territoires dans le but d'augmenter la taille de l'ensemble de référence (nombre de peuplements sondés) lors de la réalisation de la compilation. Ces analyses se font principalement dans le cas de territoires de petites superficies dans lesquels peu de placettes ont été établies. Les regroupements proposés doivent évidemment tenir compte de l'aspect écologique des différents territoires pour être considérés, sans quoi la qualité des estimations pourrait être altérée. On utilise encore une fois les critères statistiques associés aux estimations *k*-NN pour comparer les résultats obtenus en considérant les territoires de façon individuelle et regroupée.

6. IMPORTANCE DES VARIABLES

L'intérêt d'utiliser la méthode de calcul de la distance « MSN » est qu'il est possible de calculer l'importance de chacune des variables explicatives retenues dans le modèle *k*-NN final. Les méthodes « RandomForest » et « *ik*-NN » offrent également cette possibilité, mais un rapport produit par Ferland-Raymond (2011) montre que la méthode « MSN » permet de produire un gradient beaucoup plus marqué de l'importance des variables que les autres mesures de distance.

De façon très générale, c'est à partir des vecteurs canoniques générés dans l'ACC inclus dans la méthode « MSN » que l'on peut obtenir une idée de la pondération de chacune des variables dans le *k*-NN. Ces variables contiennent cependant un effet d'échelle qui doit être enlevé avant d'en faire l'interprétation. Rencher (2002) présente de façon détaillée la procédure de standardisation des vecteurs de poids à utiliser pour éliminer cet effet d'échelle. Les *c* vecteurs de poids standardisés obtenus doivent ensuite être combinés pour donner une idée globale de l'importance des variables. Pour ce faire, ils sont d'abord multipliés par leur corrélation canonique correspondante et les valeurs absolues du produit sont ensuite sommées de manière à obtenir un vecteur de poids de taille *p*, où *p* correspond au nombre de variables explicatives.

L'ordre final d'importance des variables peut varier de façon importante d'un territoire à l'autre, car il est déterminé lorsque tous les paramètres ont été fixés et toutes les techniques d'optimisation ont été réalisées. Il est donc peu probable, même dans des territoires adjacents, qu'exactement les mêmes variables explicatives soient retenues et, par surcroît, dans le même ordre, car trop d'éléments peuvent différer entre les territoires : variables d'optimisation retenues et leur importance relative, disponibilité de certaines variables explicatives (p. ex., LiDAR), etc.

À titre informatif, une synthèse globale des résultats a été produite afin d'obtenir l'importance des variables totale pour l'ensemble des territoires compilés avec la méthode statistique *k*-NN jusqu'à maintenant (compilations antérieures à 2017). Cette synthèse a été produite en séparant les territoires où dominent les essences feuillus (12 territoires) et résineuses (25 territoires), car le pouvoir explicatif des différentes variables explicatives varie potentiellement entre ces types de territoires. Les figures suivantes présentent donc l'ordre d'importance des variables (figure 9 et figure 11) et le nombre de fois que chacune a été utilisée dans les compilations (figure 10 et figure 12). On peut ainsi avoir une idée générale des variables importantes dans les analyses *k*-NN.

Figure 9 : Importance des variables retenues dans les territoires où dominent les essences feuillues compilées avec la méthode *k*-NN

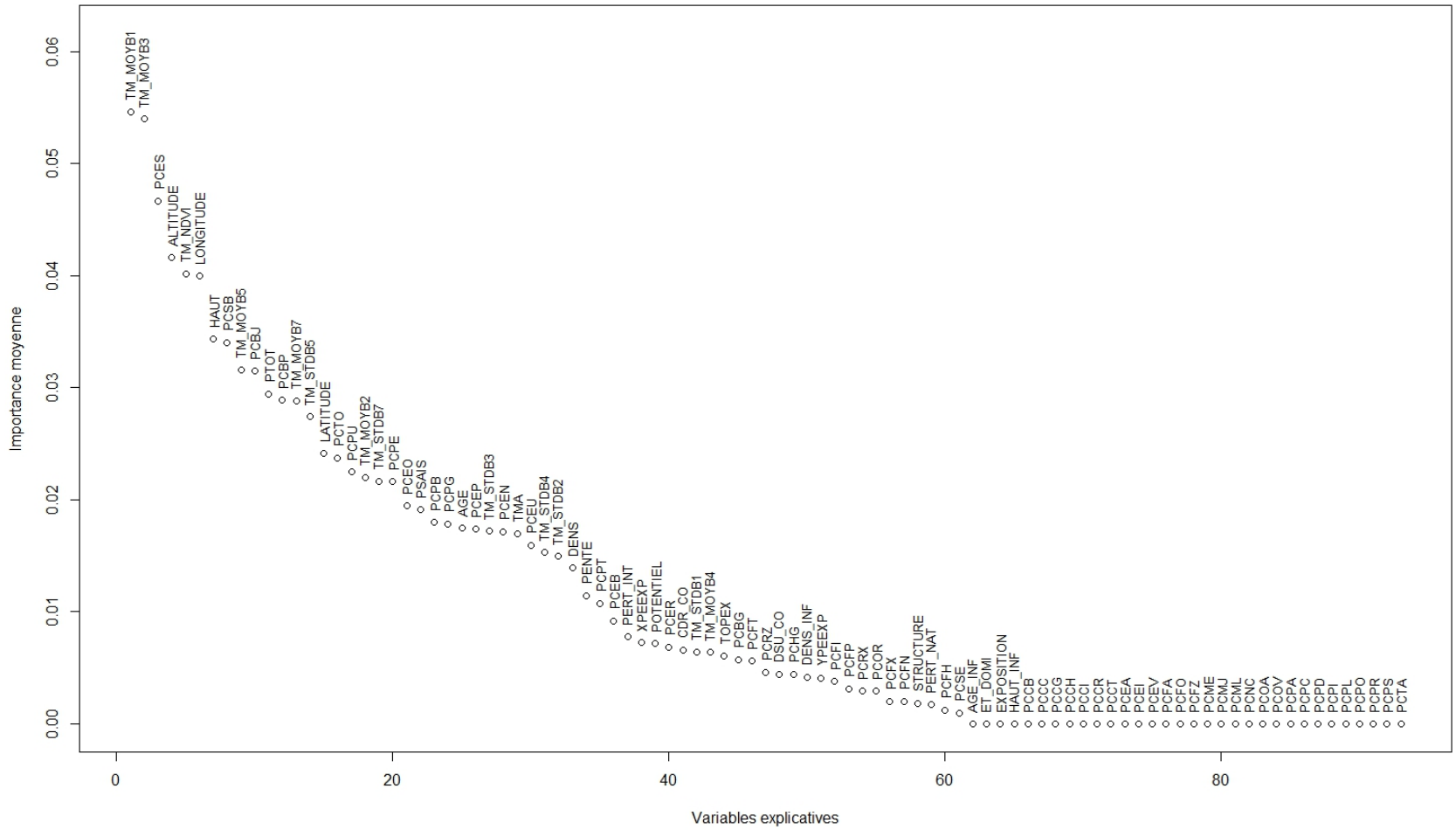


Figure 10 : Fréquence des variables retenues dans les territoires où dominent les essences feuillues compilées avec la méthode *k*-NN

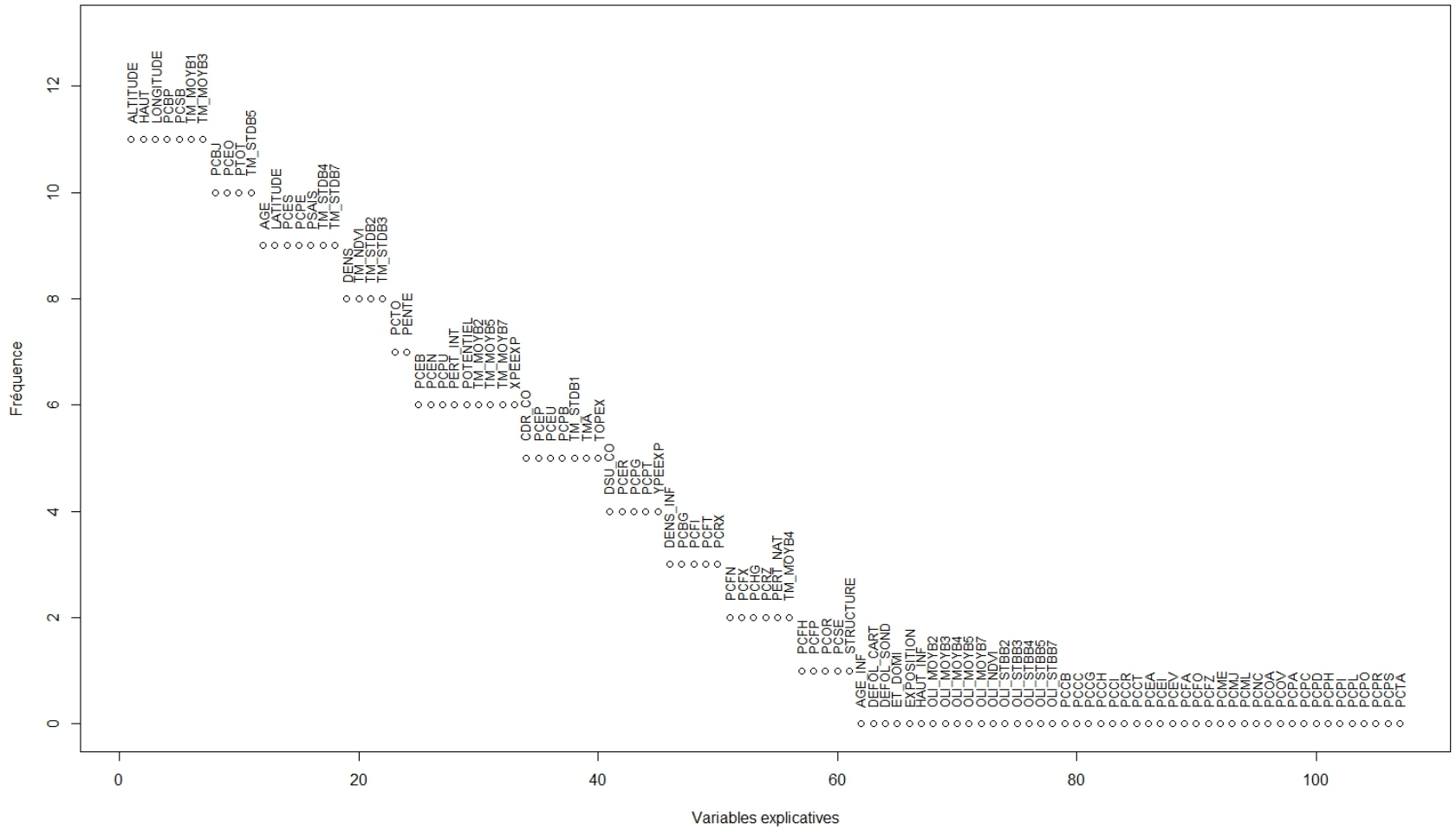


Figure 11 : Importance des variables retenues dans les territoires où dominent les essences résineuses compilées avec la méthode *k*-NN

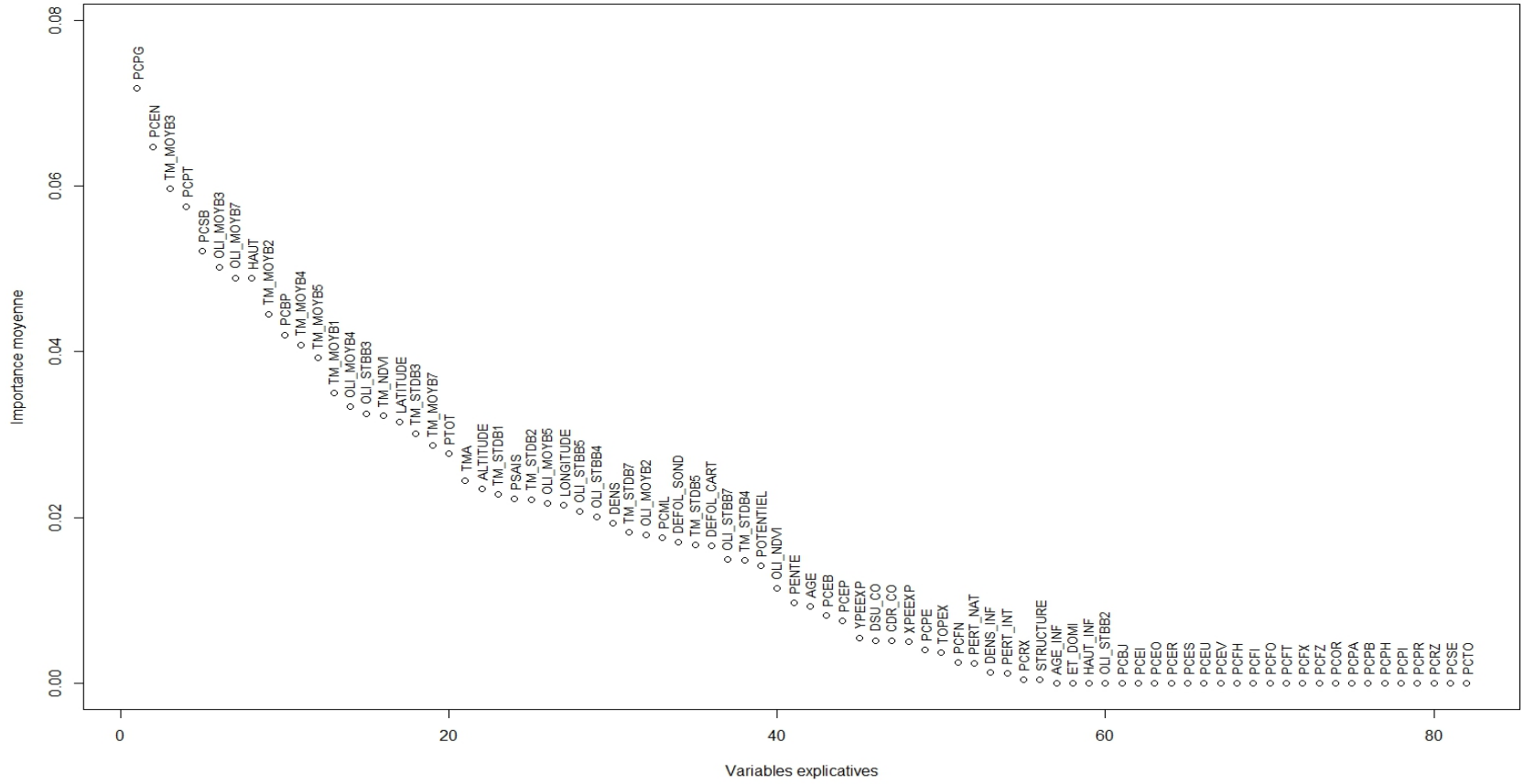
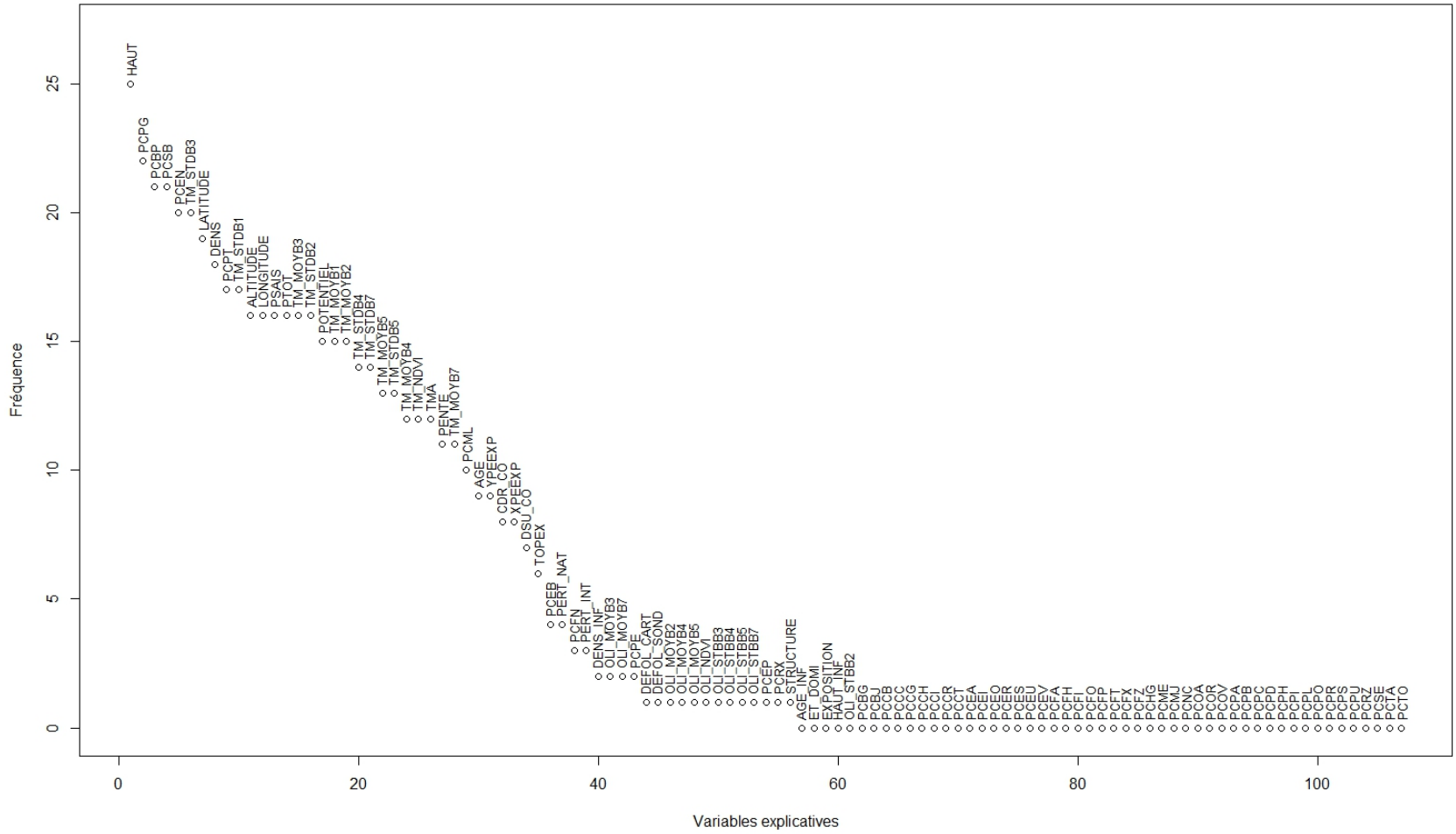


Figure 12 : Fréquence des variables retenues dans les territoires où dominent les essences résineuses compilées avec la méthode *k*-NN



7. VARIABLES DENDROMÉTRIQUES ESTIMÉES

Les estimations de variables dendrométriques sont obtenues en optimisant les analyses *k*-NN en fonction du volume marchand brut de certaines essences seulement (les plus importantes sur le territoire). À partir de la liste des *k* voisins sélectionnés, on produit les estimations pour trois variables d'intérêt, soit le nombre de tiges (tiges/ha), la surface terrière (m^2/ha) et le volume marchand brut (m^3/ha), et ce, par peuplement et par essence. On produit également des résultats agrégés selon les groupes d'attribution, le type d'essences et le total toutes essences. On utilise ensuite ces estimations pour calculer le volume marchand brut moyen par tige (dm^3/tige) et le diamètre moyen quadratique (cm). Dans le cas des gaules, on estime uniquement le nombre de tiges et la surface terrière, et ce, par type d'essences (résineuses ou feuillues) et pour le total toutes essences seulement, à partir de la même liste des *k* voisins sélectionnés par peuplement. Pour plus de détails sur la façon de bien utiliser les données de compilation, référez-vous au Guide d'utilisation de la carte écoforestière et des résultats d'inventaire écoforestier du Québec méridional (MRNF-DIF, 2021).

8. BONNES PRATIQUES ET MISES EN GARDE

Un élément important à considérer dans le cadre de l'application de la méthode *k*-NN est que cette méthode ne produit pas des estimations exemptes de biais. En effet, aucune estimation ne peut être plus petite ou plus grande que la plus petite ou la plus grande observation de l'ensemble de référence. Ce phénomène devient encore plus important lorsque le nombre de voisins (*k*) est supérieur à 1, car les estimations correspondent alors à la moyenne pondérée des variables dendrométriques mesurées dans les *k* peuplements sondés. Le biais est cependant facilement quantifiable, ce qui permet de bien utiliser et interpréter les résultats obtenus pour l'une ou l'autre des essences sur le territoire.

Les estimations de variables dendrométriques sont produites pour tous les peuplements de 7 m et plus de hauteur de la carte écoforestière originale. Cependant, la qualité des estimations sur le territoire n'est pas la même partout. Plusieurs éléments contribuent à cette variabilité et il est important d'en tenir compte lorsque l'on utilise les résultats de compilation. Premièrement, la qualité des estimations peut être influencée par le plan d'échantillonnage réalisé lors du sondage terrestre. En effet, un des objectifs du sondage est d'échantillonner les peuplements forestiers, productifs, accessibles, de 7 m et plus de hauteur et pouvant être aménagés. Cet ensemble de peuplements correspond à la population cible. Le territoire n'est cependant pas entièrement affecté à des activités liées à l'aménagement forestier. En effet, on détermine le territoire à sonder en utilisant les données territoriales associées aux modes de gestion, aux usages forestiers et aux zones d'application de modalités d'intervention (ZAMI). Ainsi, seuls certains modes de gestion sont sondés et, dans ces secteurs, certaines portions de territoire (usages et ZAMI) peuvent être exclues si les contraintes à l'aménagement sont jugées trop importantes. À titre d'exemple, on ne sonde pas les peuplements situés dans des aires protégées, des parcs provinciaux, des refuges biologiques, des centres de ski alpin ou des forêts d'expérimentation. On ne sonde donc qu'une partie des peuplements de 7 m et plus de hauteur. Sur le plan statistique, les estimations *k*-NN pour un peuplement donné sont réputées valides seulement si ce peuplement faisait partie de la population cible lors du sondage. Dans le cas contraire, les résultats de compilation ne sont là qu'à titre indicatif et doivent être utilisés avec précaution.

Par ailleurs, une des règles fondamentales de l'application de la méthode d'estimations *k*-NN dans le contexte forestier est que, pour être en mesure de bien estimer une essence donnée, il faut l'avoir observée au préalable sur le territoire lors du sondage terrestre. Dans le cas des essences principales, il n'y a généralement pas de problème, car l'échantillonnage est très souvent suffisant. En effet, l'allocation de placettes se fait proportionnellement à la superficie des strates formées et, comme ces essences sont partout sur le territoire, un nombre important de placettes sont établies dans ces superficies. Dans certains cas bien précis, le taux d'échantillonnage peut même être abaissé, par exemple, lorsqu'un nombre élevé de PET a été prévu dans des strates d'une très grande superficie, dont la composition en essences est très homogène. Le nombre de PET est ainsi réduit dans ces strates, puis est ensuite réparti dans d'autres strates de plus faible superficie ou de composition en essences plus hétérogène ou plus rare. Malgré ces modifications

de placettes, il arrive souvent que l'échantillonnage ne soit toujours pas adéquat (trop faible) dans le cas des peuplements composés d'essences rares. Il est donc difficile d'obtenir des estimations de qualité pour ces essences et les résultats doivent, encore une fois, être utilisés avec prudence.

9. RÉFÉRENCES

CROOKSTON, N. L. et A. O. FINLEY (2008). “yaImpute: An R package for *k*-NN imputation”, Journal of Statistical Software, 23(10) : 1-16.

CYR, G., J. GOSSELIN et V. LAFLÈCHE (2010). Guide d’identification des aires d’intensification de la production ligneuse : Méthodologie d’identification des aires potentielles pour l’intensification de la production de matière ligneuse et interprétation des résultats, ministère des Ressources naturelles et de la Faune, Forêt Québec, Direction des inventaires forestiers, p. 40-56.

FERLAND-RAYMOND, B. (2010). Application du *k*-NN à l’estimation de volumes forestiers — Cas particulier d’un sous-domaine de l’UAF 012-54, essai, Département de mathématiques et de statistique, Faculté des sciences et de génie, Université Laval, 52 p.

FERLAND-RAYMOND, B. (2011). Compilation du volume de bois par essence avec le *k*-NN : comparaison des méthodes et outils de validation, rapport interne, ministère des Forêts, de la Faune et des Parcs, Direction des inventaires forestiers, 42 p.

FORTIN, M. et L. LANGEVIN (2010). ARTÉMIS-2009 : un modèle de croissance basé sur une approche par tige individuelle pour les forêts du Québec, Mémoire de recherche forestière 156, ministère des Ressources naturelles et de la Faune, Direction de la recherche forestière.

FRANCO-LOPEZ, H., A. R. EK et M. E. BAUER (2001). “Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method”, Remote Sensing of Environment, 77 : 251-274.

HAARA, A., M. MALTAMO et T. TOKOLA (1997). “The K-nearest-neighbour method for estimating basal-area diameter distribution”, Scandinavian Journal of Forest Research, 12(2) : 200-208.

HUDAK, A. T., N. L. CROOKSTON, J. S. EVANS, D. E. HALL et M. J. FALKOWSKI (2008). “Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data”, Remote Sensing of Environment, 112 : 2232-2245.

LEMAY, V. et H. TEMESGEN (2005). “Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables”, Forest Science, 51(2) : 109-119.

LEMAY, V., J. MAEDEL et N. C. COOPS (2008). “Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery”, Remote Sensing of Environment, 112 : 2578-2591.

MÄKELÄ, H. et A. PEKKARINEN (2004). “Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data”, Forest Ecology and Management, 196 : 245-255.

MALINEN, J. (2003). “Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning”, Silva Fennica, 37(1) : 109-120.

MALTAMO, M., J. MALINEN, P. PACKALÉN, A. SUVANTO et J. KANGAS (2006). “Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data”, Canadian Journal of Forest Research, 36 : 426-436.

MCROBERTS, R. E. (2012). “Estimating forest attribute parameters for small areas using nearest neighbors techniques”, Forest Ecology and Management, 272 : 3-12.

MCROBERTS, R. E., M. D. NELSON et D. G. WENDT (2002). “Stratified estimation of forest area using satellite imagery, inventory data, and the *k*-Nearest Neighbors technique”, Remote Sensing of Environment, 82 : 457-468.

MCROBERTS, R. E., G. R. HOLDEN, M. D. NELSON, G. C. LIKNES et D. D. GORMANSON (2006). “Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service”, Canadian Journal of Forest Research, 36 : 2968-2980.

MCROBERTS, R. E., E. O. TOMPPO, A. O. FINLEY et J. HEIKKINEN (2007). “Estimating areal means and variances of forest attributes using the *k*-Nearest Neighbors technique and satellite imagery”, Remote Sensing of Environment, 111 : 466-480.

MFFP-DIF (2015). Norme de stratification écoforestière — Quatrième inventaire écoforestier du Québec méridional (réédition, septembre 2015), ministère des Forêts, de la Faune et des Parcs, Direction des inventaires forestiers, 101 p.

MFFP-DIF (2021). Guide d’utilisation de la carte écoforestière et des résultats d’inventaire écoforestier du Québec méridional (réédition, juillet 2021), ministère des Forêts, de la Faune et des Parcs, Direction des inventaires forestiers, 65 p.

MOEUR, M. et A. R. STAGE (1995). “Most similar neighbor: an improved sampling inference procedure for natural resource planning”, Forest Science, 41(2) : 337-359.

MRNF-DIF (2011). Guide d’utilisation des données des projets de compilation — Projets des unités de sondage des années 2004 à 2010 (quatrième inventaire écoforestier), ministère des Ressources naturelles et de la Faune, Direction des inventaires forestiers, 81 p.

PACKALÉN, P. et M. MALTAMO (2006). “Predicting the plot volume by tree species using airborne laser scanning and aerial photographs”, Forest Science, 52(6) : 611-622.

PACKALÉN, P. et M. MALTAMO (2007). “The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs”, Remote Sensing of Environment, 109 : 328-341.

R CORE TEAM (2016). “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria [https://www.R-project.org/].

REESE, H., M. NILSSON, T. G. PAHLÉN, O. HAGNER, S. JOYCE, U. TINGELÖF, M. EGBERTH et H. OLSSON (2003). "Countrywide estimates of forest variables using satellite data and field data from the National Forest Inventory", Ambio, 32(8) : 542-548.

RENCHER, A. C. (2002). "Methods of multivariate analysis (Second Edition)", John Wiley & Sons Inc., 708 p.

RÉGNIÈRE, J. et R. SAINT-AMANT (2008). BioSIM 9 – User's Manual, Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Qc. Inf. Rep. LAU-X-134E, 76 p.

RUEL, J.-C., S. J. MITCHELL et M. DORNIER (2002). "A GIS based approach to map wind exposure for windthrow hazard rating", Northern Journal of Applied Forestry, 19(4) : 183-187.

STAGE, A. R. (1976). "An expression for the effect of aspect, slope, and habitat type on tree growth", Forest Science, 22(4) : 457-460.

STAGE, A. R. et N. L. CROOKSTON (2007). "Partitioning error components for accuracy-assessment of near-neighbor methods of imputation", Forest Science, 53(1) : 62-72.

TEMESGEN, H., V. M. LEMAY, K. L. FROESE et P. L. MARSHALL (2003). "Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia", Forest Ecology and Management, 177 : 277-285.

TOMPPO, E. et M. HALME (2004). "Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach", Remote Sensing of Environment, 92 : 1-20.

TOMPPO, E. (2006). "The Finnish multi-source National Forest Inventory – small area estimation and map production (Chapter 12)", Finnish Forest Research Institute, Finland, 195-224.

ANNEXE I TABLEAUX DE CONVERSION

Tableau I : Conversion de la classe de densité du couvert en variable numérique (stratification « initiale »)

Classe de densité	Description (%)	Classe de densité numérique (%)
A	81-100	90,5
B	61-80	70,5
C	41-60	50,5
D	26-40	33,0

Tableau II : Conversion de la classe de hauteur en variable numérique (stratification « initiale »)

Classe de hauteur	Description (m)	Classe de hauteur numérique (m)
1	22 et +	24,5
2	17 à 22	19,5
3	12 à 17	14,5
4	7 à 12	9,5
5	4 à 7	5,5
6	2,1 à 4	3,0
7	0 à 2	1,0

Tableau III : Détermination de l'épaisseur du dépôt à partir du dépôt de surface (toutes les stratifications)

Codification du dépôt de surface	Exemple	Épaisseur du dépôt (cm)
'x'	'1a'	100,0
'x'Y	'1a'Y	75,0
'x'M	'1a'M	37,5
R'x'	R'1a'	30,0
M'x'	M'1a'	17,5
R	R	12,5

Forêts, Faune
et Parcs
Québec