UNIVERSITÉ DU QUÉBEC À MONTRÉAL

OPTIMISATION DE LA RÉSILIENCE DE LA FORÊT URBAINE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

RAOUF MONCEF BELBAHAR

SEPTEMBRE 2021

REMERCIEMENTS

Mes sincères remerciements sont dédiés en premier lieu à ma directrice de recherche la professeure Marie-Jean Meurs, qui a été d'un soutien et d'une attention exceptionnels. Sa confiance, sa disponibilité ainsi que ses conseils précieux m'ont permis d'accumuler des expériences professionnelles et personnelles notables, qu'elle trouve ici l'expression de ma profonde et sincère reconnaissance. Je tiens également à exprimer ma gratitude à toute l'équipe SylvCiT notamment : Christian Messier, Arcady Gascon-Afriat, Fanny Maure, Bastien Lecigne, Sylvia Wood et enfin Annick Saint-Denis qui ont été aussi disponibles que sympathiques. Sans leur savoir, leur passion et leur enthousiasme, ce travail de recherche n'aurait jamais pu aboutir. Je tiens aussi à remercier mes collègues de laboratoire et amis, en particulier, Khalid, Marc et Diego pour les bons moments passés avec eux durant mon cursus de maîtrise. Je désire aussi remercier toutes les personnes que j'ai pu rencontrer au courant de nos diverses collaborations interdisciplinaires. Je tiens à remercier mes amis et tout d'abord Mehdi, qui a été d'une aide, d'une gentillesse et d'un support inestimable, ma minouche pour sa présence et son soutien, Rafik, Reda, Nail, Karim, Samy et bien d'autres pour leur rôle exceptionnel dans ma vie. Pour finir, je dédie ce mémoire à ma mère et mon père, mes piliers dans ma vie, sans qui rien n'aurait été possible et à qui j'exprime tous mes sentiments de respect, d'amour, de gratitude et de reconnaissance.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

DHP Diamètre à Hauteur de Poitrine

IN Infrastructures Naturelles

LR Régression linéaire

MAE Erreur Absolue Moyenne

MSE Erreur quadratique moyenne

NN Réseau de neurones

ReLU Unité linéaire rectifiée

RF Forêt d'arbres décisionnels

RMSE Racine carrée de l'erreur quadratique moyenne

SE Services Écosystémiques

SGD Descente de gradient stochastique

UHI Îlot de chaleur

USDA Département de l'Agriculture des É-U

RÉSUMÉ

L'urbanisation en constante augmentation a de nombreux effets négatifs sur les humains et sur les écosystèmes. Bien qu'ayant fait progresser le bien-être humain, les villes sont devenues des sources d'impact importantes sur l'environnement. Réduire cet impact représente donc un des grands défis actuels en matière de développement et d'optimisation de l'utilisation des ressources planétaires.

Aujourd'hui, les arbres urbains suscitent un intérêt croissant pour leur contribution à l'atténuation de certains effets de l'urbanisation. En effet, ils fournissent de nombreux services écosystémiques pouvant par exemple améliorer la qualité de l'air, atténuer les îlots de chaleur ou encore apporter une plus-value culturelle et esthétique. La reconnaissance de ces services a conduit à l'établissement de programmes de plantation d'arbres dans plusieurs villes du monde. Cependant, les approches traditionnelles favorisent souvent l'aspect esthétique rendant le couvert forestier susceptible d'être affecté par diverses menaces. Une méconnaissance de l'importance de la diversité des arbres à planter menace leur résilience ainsi que la dynamique de délivrance de ces services écosystémiques.

Notre recherche permet de créer un système de recommandation des essences d'arbres adéquates à planter en fonction de la diversité des traits fonctionnels de ceux-ci, s'appuyant sur des méthodes d'optimisation et d'apprentissage automatique. Le logiciel créé utilise les données relatives à l'emplacement, aux arbres mitoyens et au climat futur dans le but de prescrire la plantation des bons arbres au bon endroit afin de maximiser la diversité fonctionnelle de ceux-ci à différentes échelles spatiales. Une telle approche permet « d'immuniser » la couverture forestière urbaine face aux menaces climatiques et biotiques (insectes et maladies) en augmentation constante et d'optimiser les services fournis par ces arbres en fonction des besoins des citoyens. Un accent particulier est mis sur les interfaces personne-machine, notamment les facteurs qui affectent l'expérience utilisateur, l'acceptation des recommandations et la transparence.

Mots-clés : Système de recommandation, système d'aide à la décision, évaluation des services écosystémiques, apprentissage automatique, réseaux de neurones, arbres urbains

INTRODUCTION

La croissance des populations urbaines est de plus en plus rapide. En effet, selon un rapport du département des affaires économiques et sociales des Nations Unies (DESA, 2018), plus de 55% de la population mondiale vit dans les villes et ce chiffre devrait atteindre 68% à l'horizon 2050. Cette urbanisation massive est reconnue pour être la cause, entre autres, de la dégradation de la qualité de l'air urbain (Han *et al.*, 2014), de la présence de nombreux îlots de chaleur et de risques d'inondation (Eigenbrod *et al.*, 2011; Lemonsu *et al.*, 2015). Ceci entraîne une détérioration des conditions de vie en milieu urbain et a un impact négatif sur la population urbaine et l'environnement (Bodnaruk *et al.*, 2017). Établir des stratégies d'urbanisation durable devient plus que jamais une nécessité et est souvent considéré comme la clé d'un développement urbain réussi (DESA, 2018). Dans ce sens, les approches fondées sur des solutions naturelles dans les processus de planification et de décision urbaine se multiplient. C'est ainsi qu'aujourd'hui les arbres urbains suscitent un intérêt croissant pour la restauration des écosystèmes urbains et l'atténuation de certains effets de l'urbanisation (Bodnaruk *et al.*, 2017; Escobedo *et al.*, 2011; Parsa *et al.*, 2019). En effet, la reconnaissance des avantages et des services écosystémiques procurés par la canopée urbaine a conduit à l'établissement de programmes de plantation d'arbres à travers plusieurs villes du monde (Pataki *et al.*, 2011). Cependant, une méconnaissance des mécanismes et de la dynamique de délivrance des services écosystémiques, notamment l'interaction entre certaines espèces ou encore la priorisation d'une espèce par rapport à une autre pour des raisons esthétiques par exemple, peut réduire l'impact de ces programmes de plantation. En effet, en parallèle des effets positifs

mentionnés ci-dessus, les arbres urbains peuvent avoir des effets négatifs tels la libération d'allergènes dans l'air ou encore le blocage de la lumière pouvant ainsi nuire au bien-être humain (Delshammar *et al.*, 2015; Cariñanos *et al.*, 2017). En outre, les arbres urbains sont de plus en plus exposés à diverses menaces telles que le changement climatique, les invasions d'insectes et les maladies exotiques dues aux échanges commerciaux de plus en plus importants. La diversification des populations d'arbres devient d'une importance capitale. De ce fait, une gestion stratégique des forêts urbaines permettrait d'optimiser les services qu'elles rendent et assurerait la résilience des écosystèmes urbains, c'est-à-dire qu'ils seront capables d'absorber ou de tolérer les perturbations futures, continuant ainsi à fournir les fonctions et services qu'ils rendent à la société. Cependant, les services écosystémiques délivrés par un arbre ou un ensemble d'arbres varient considérablement en fonction des caractéristiques environnementales et socio-économiques de chaque site de plantation mais aussi des caractéristiques de l'arbre.

À cet égard, plusieurs méthodologies (Isely *et al.*, 2010; Walker, 2017) ont été développées et appliquées dans le but de comprendre comment gérer les forêts urbaines afin de produire des services écosystémiques plus bénéfiques. Une des approches développées permet de mettre en évidence les emplacements de plantation optimaux, en fonction des services écosystémiques et des paramètres économiques et écologiques, dans le but d'atténuer la pollution atmosphérique et l'îlot thermique urbain (Bodnaruk *et al.*, 2017). En terme d'application, nous pouvons citer par exemple, i-Tree (i-Tree, 2021), suite logicielle contenant des outils d'inventaires et d'analyse de la forêt urbaine permettant de mesurer les avantages fournis par cette dernière.

L'objectif de notre travail de recherche consiste en la conception et la mise en place d'une solution permettant de visualiser la distribution géographique des arbres, de quantifier les services rendus par chaque espèce et enfin de générer des recommandations de plantation des essences d'arbres adéquates.

Ces recommandations seront générées en fonction de l'emplacement, des arbres mitoyens et des données climatiques actuelles et prévisionnelles dans le but d'immuniser l'ensemble contre les facteurs environnementaux et climatiques d'une part, et de maximiser les apports économiques et sociaux de la canopée urbaine d'autre part.

Ce mémoire est constitué de plusieurs chapitres. Étant donnée la nature pluridisciplinaire du projet, le chapitre 1 expose des concepts préliminaires nécessaires à la compréhension de ce travail de recherche. Le chapitre 2 présente un état de l'art des systèmes de recommandations et de la quantification des services écosystémiques. Le chapitre 3 décrit en détail les données utilisées ainsi que la méthodologie suivie dans l'objectif d'une part d'effectuer un regroupement fonctionnel des espèces d'arbres urbains à Montréal, et d'autre part, de développer un modèle de prédiction de croissance des diamètres à hauteur de poitrine (DHP) des arbres publics de Montréal. Le chapitre 4 est consacré à la présentation et la description de l'outil SylvCiT. Nous y aborderons dans un premier temps le calcul des services écosystémiques et des indices biologiques, puis, nous évoquerons la mise en œuvre de la recommandation. Enfin, nous présenterons l'implémentation et les technologies utilisées pour le développement de l'outil. Les résultats obtenus sont présentés, analysés et discutés dans le chapitre 5. Enfin, la conclusion de ce mémoire offre une vue d'ensemble du projet, de ses enjeux et de ses perspectives futures.

CHAPITRE I

CONCEPTS PRÉLIMINAIRES

1.1 La foresterie urbaine

Les forêts urbaines sont aujourd'hui une composante importante et de plus en plus précieuse de l'environnement urbain. La forêt urbaine est définie de plusieurs manières dans la littérature. Ainsi, la définition la plus commune considère l'ensemble des arbres dans une cité, une ville ou une banlieue comme constituant une forêt urbaine. Arbres Canada (Arbres Canada, 2019) va plus loin dans la définition : la forêt urbaine désigne les arbres, les forêts, les espaces verts et les éléments abiotiques (caractéristiques du milieu de nature physique ou chimique), biotiques (influences entre êtres vivant dans le même milieu) et culturels connexes qui se trouvent dans les zones allant du noyau urbain à la limite péri-urbaine.

Les forêts urbaines sont constituées d'arbres de diverses provenances. Certains sont des vestiges ou du moins dérivent de forêts présentes avant l'accroissement urbain, d'autres sont des espèces indigènes ou exotiques plantées ou transplantées. Certains poussent sans attention, d'autres sont gérés de manière intensive. Ils sont plantés seuls, en grappes, lignes, bandes (en particulier le long des routes et des cours d'eau), dans des parcs ou dans des blocs forestiers denses (Oke, 1989). Les forêts urbaines en opposition aux forêts naturelles, en plus de souffrir d'une pauvreté en terme de diversité génétique, vivent dans des conditions écologiques

imposées par le mode de vie urbain et les activités urbaines. Elles subissent leurs effets défavorables spécifiques tels que la présence des sels de déglaçage et la compaction des sols, ce qui influence fortement le développement des arbres, affectés par la présence et le comportement des agents pathogènes (Tello *et al.*, 2005).

Au vu de l'importance des forêts en milieu urbain, notamment en ce qui concerne les nombreux bienfaits qu'elles nous offrent (*e.g.* la séquestration de polluants et particules atmosphériques, la conservation de l'énergie, la réduction des eaux de ruissellement, etc.) et au vu des stratégies mises en place par les villes pour la préservation et le développement durable des forêts (Network, 2016), la foresterie urbaine s'est imposée en tant que discipline pratique et sujet de recherche (Arbres Canada, 2019).

Ainsi, la foresterie urbaine est un secteur spécialisé des sciences forestières dont l'objectif est la culture et la gestion des arbres en vue d'assurer leur contribution actuelle et future au bien-être physiologique, social et économique de la société urbaine. Sont inclus dans cette contribution, les bienfaits environnementaux, les activités récréatives et l'utilité publique des arbres (Jorgensen, 1974). De plus, selon Arbres Canada, la foresterie urbaine consiste en la planification, la plantation, la protection, l'entretien, la gestion et le soin durable des arbres, des forêts, des espaces verts et des ressources connexes dans les villes et collectivités ainsi qu'en périphérie de celles-ci pour fournir aux gens des bienfaits associés à l'économie, à l'environnement, à la société et à la santé publique (Arbres Canada, 2019). La foresterie urbaine est donc un élément important de toute stratégie de développement urbain durable.

1.2 Urbanisation et services écosystémiques

Aujourd'hui, plus de la moitié de la population mondiale vit dans les villes et, d'ici à 2050, 68% de la population mondiale vivra en milieu urbain. Bien que les villes n'occupent que 3% de la surface de la Terre, elles consomment 78% de l'énergie et émettent 60% du dioxyde de carbone (DESA, 2018). L'urbanisation entraîne une transformation profonde des milieux naturels. Cela se traduit par une perte de la végétation originelle en déboisant les forêts naturelles au profit du cadre bâti (immeubles et surfaces imperméables), entraînant des problèmes environnementaux divers tels que l'augmentation de la température urbaine, la diminution de la qualité de l'air ou encore les problèmes liés au ruissellement de l'eau. Les infrastructures naturelles en zones urbaines sont une des clés permettant de résoudre ces problèmes en contribuant au bien être, à la santé et à la qualité de vie des populations. Ainsi, les infrastructures naturelles (IN) sont définies comme un réseau interconnecté d'espaces verts et bleus, tels que les parcs urbains et les plans d'eau, qui préservent les valeurs et les fonctions des écosystèmes naturels en fournissant des bénéfices aux populations humaines (Benedict *et al.*, 2012). Dans les parties suivantes nous aborderons les principaux problèmes induits par l'urbanisation ainsi que le rôle des arbres urbains dans leur mitigation.

1.2.1 La chaleur urbaine

Le développement des zones urbaines se traduit par l'augmentation de la circulation automobile, l'intensification des activités industrielles et la prédominance du milieu bâti induisant de ce fait une élévation des températures urbaines. Ce phénomène est couramment nommé « îlots de chaleur » (*Urban Heat Island (UHI)*) et est désormais considéré comme l'un des problèmes environnementaux urbains les plus graves au monde (Mochida et Lun, 2008).

L'intensité des îlots de chaleur est généralement définie comme la différence de température (ΔT) entre les zones urbaines et rurales environnantes. Cette différence de température peut être basée sur la température de surface ou la température de l'air (Li *et al.*, 2019). Ces écarts de températures peuvent être observés entre les zones fortement imperméabilisées et les espaces verts (parcs et terrains arborés) et zones rurales situés à proximité et peuvent varier entre 5 et 8 degrés Celsius (Hardin et Jensen, 2007). La figure 1.1 permet de représenter les îlots de chaleur et de fraîcheur urbains, ainsi que la température de surface dans la région de Montréal par un gradient de couleur. Les zones les plus chaudes sont représentées en rouge, tandis que celle les plus fraîches en vert. On peut observer des températures élevées dans les quartiers centraux et les zones industrielles. Cette température tend à diminuer plus on s'éloigne de la ville de Montréal et de sa zone péri-urbaine.

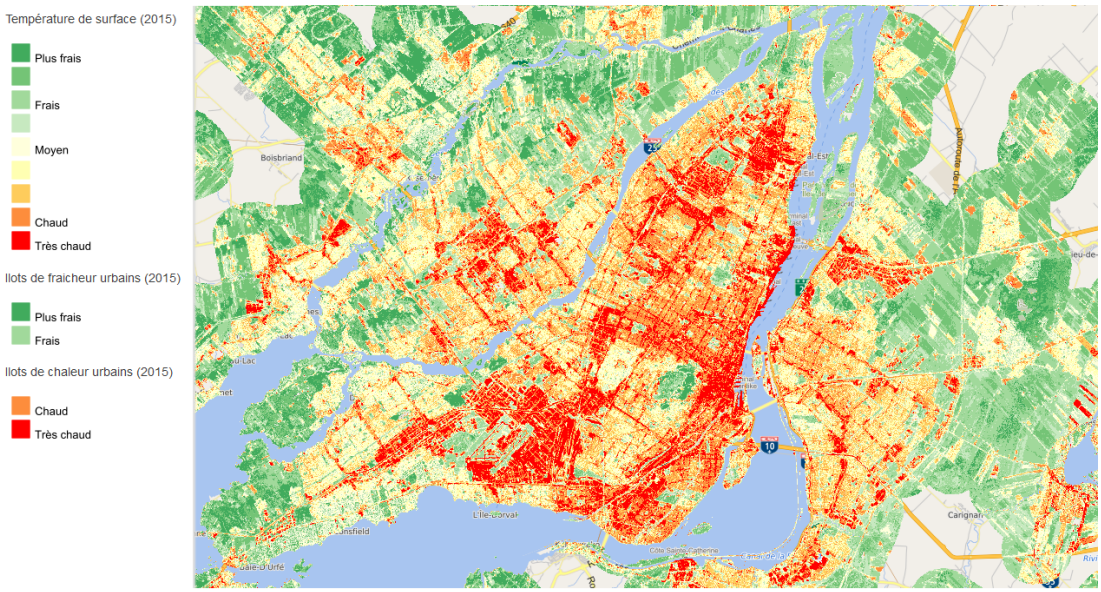


Figure 1.1: Carte des îlots de chaleur et de fraîcheur urbains, ainsi que de la température de surface à Montréal. Source : (Données Québec, 2012)

Ainsi, la hausse des températures peut entraîner pour la population un inconfort

allant jusqu'à provoquer des problèmes de santé tels que des coups de chaleur et des problèmes cardiaques (Lamothe *et al.*, 2019).

On distingue diverses causes à l'augmentation de la température dans les milieux urbains. D'abord, il est reconnu à travers la littérature que le contexte climatique de réchauffement planétaire contribue à l'accentuation des îlots de chaleur et à l'exacerbation des conséquences néfastes de ces derniers. Ensuite, des facteurs naturels influent sur la formation des îlots de chaleur. On peut par exemple citer la variabilité climatique inhérente aux saisons, le niveau d'ensoleillement et la vitesse du vent. Ainsi, un îlot de chaleur sera particulièrement important en période de canicule (Vergriete et Labrecque, 2007). Enfin, des facteurs anthropiques, relatifs à l'activité humaine influent également sur la formation des îlots de chaleur. En effet, la densité du couvert végétal, la proportion de surface construite et la dimension d'une ville ont un impact sur la température. De plus, la formation d'îlots de chaleur est fortement liée à la production de flux de chaleur par la consommation d'énergie. Par ailleurs, la circulation routière, l'activité industrielle et la climatisation, activités reconnues comme étant fortement émettrices de polluants et de chaleur, impactent le milieu ambiant et l'ampleur de la hausse de température.

Plusieurs stratégies ont été proposées afin de permettre la mitigation des effets des îlots de chaleur et ainsi favoriser la fraîcheur dans les milieux urbains tout en contribuant à la réduction de la demande énergétique et de la pollution de l'air et de l'eau. Ces mesures ont un effet positif sur le climat tant au niveau local que global. Parmi ces stratégies nous pouvons en citer quatre (Lamothe *et al.*, 2019) :

- Réduction de la chaleur anthropique,
- Gestion des eaux pluviales et perméabilité des sols,
- Gestion de l'architecture et de l'aménagement urbain,
- Augmentation du couvert arboré.

Nous nous attarderons particulièrement sur la dernière stratégie qui est directe-

ment reliée à l'objet de notre étude. Aujourd'hui de nombreux travaux démontrent l'importance des arbres urbains dans la réduction des effets des îlots de chaleur urbains (McPherson *et al.*, 2005; Turner-Skoff et Cavender, 2019; Hardin et Jensen, 2007). En effet, les arbres offrent des avantages intéressants en terme de création de fraîcheur et de mitigation des îlots de chaleur via deux processus, l'évapotranspiration et l'ombrage des surfaces minéralisées. Des arbres plantés et placés correctement peuvent atténuer les températures dans les environnements urbains (Turner-Skoff et Cavender, 2019). Ces arbres fournissent non seulement de l'ombre en interceptant et en absorbant la lumière (Huang *et al.*, 2008) mais grâce au processus physiologique de la transpiration, qui consiste en la perte d'eau par les feuilles de l'arbre et à son évaporation, ils permettent de refroidir activement l'air des villes (Georgi et Zafiriadis, 2006). Cette réduction de température, notamment dans les grandes villes, peut à terme contribuer à atténuer l'impact du changement climatique sur la santé humaine (McDonald *et al.*, 2007; Turner-Skoff et Cavender, 2019).

1.2.2 La qualité de l'air

L'expansion rapide des zones urbaines et la multitude d'activités tenues dans ces dernières, génèrent une grande production de polluants atmosphériques tels le dioxyde de carbone (CO_2), les oxydes d'azote (NO_3), le méthane (CH_4) et l'ozone (O_3). À cela s'ajoutent les particules fines appelées matières particulaires (*Particulate Matter*) en suspension telles que $PM_{2,5}$ ou PM_{10} (l'indice indique le diamètre en micromètres). Ces particules atmosphériques, en particulier celles dont le diamètre est $< 10\mu m$ (PM_{10}), constituent une menace à long terme pour la santé, en particulier pour les fonctions respiratoires humaines (McDonald *et al.*, 2007). Une étude réalisée en 2015 (Lelieveld *et al.*, 2015) estime la mortalité globale attribuable à la pollution atmosphérique (liée aux $PM_{2,5}$ et à l'O_3) à

3,3 millions de personnes en 2010. Un rapport de Santé Canada publié en 2019 estime que le nombre de morts par année au Canada attribuables à la pollution atmosphérique d'origine anthropique est de 14 600, selon les concentrations de polluants atmosphériques de 2014 à 2016 et une projection démographique de 2015 (Santé Canada, 2019).

De nombreuses études ont démontré la capacité des arbres à absorber le CO_2, notamment en l'absorbant et le stockant dans les tissus de l'arbre, tels que le tronc, les branches et les feuilles. Les arbres contribuent aussi à éliminer des quantités importantes de polluants atmosphériques et par conséquent à améliorer la qualité de l'environnement et la santé humaine.

Les arbres éliminent la pollution atmosphérique gazeuse principalement par l'absorption des contaminant gazeux par les stomates des feuilles et le dépôt des éléments polluants particulaires à leur surface. Les taux de séquestration et d'absorption des différents polluants dépendent fortement de l'espèce et de ses caractéristiques. Les polluants particulaires, quant à eux, sont retirés de l'air à travers les branches, les feuilles ou encore les aiguilles des arbres qui forment des filtres naturels. Le taux de capture des particules, dépend de la surface foliaire (feuilles), de la taille et de la densité de la couronne et enfin de la concentration des particules.

Il convient cependant de noter que les effets bénéfiques du retrait des polluants atmosphériques par les arbres pourraient être réduits par les émissions polluantes relatives à leur plantation et leur entretien.

1.2.3 Le ruissellement des eaux

La construction d'infrastructures urbaines et le recouvrement du sol par des surfaces minéralisées provoque l'imperméabilisation de ces derniers, empêchant la pénétration des eaux de pluie et de la fonte des neiges et augmentant donc le

ruissellement de surface. Ainsi, l'imperméabilisation du sol a de multiples conséquences, dont notamment, la diminution de l'alimentation des sols, une baisse du niveau des cours d'eau et un accroissement des concentrations de polluants dans ces derniers. En effet, lors d'épisodes de précipitations importantes, la capacité de captage et de rétention d'eau est souvent dépassée. Le ruissellement des eaux sur les surfaces minéralisées qui accumulent une quantité importante de polluants (huiles et graisses de voiture, pesticides et détritus divers) peut mener à des débordements en rive entraînant une pollution importante des cours d'eau (Conseil Régional Environnement Montréal, 2007). Malgré les efforts des villes pour développer des infrastructures destinées à la gestion des eaux de pluie – réseaux de drainages, bassins de rétention d'eau et stations d'épuration – dans le but de contrôler les volumes d'eau pour éviter des problèmes d'inondation, ces solutions se révèlent insuffisantes. Ainsi, selon (Ville de Montréal, 2021b), le coût de traitement des eaux usées de Montréal pour l'année 2019 était de 62 853 000$.

Plusieurs études ont démontré le rôle des arbres urbains dans la mitigation des effets du ruissellement des eaux en (1) interceptant les eaux de pluie avant qu'elle n'atteignent le sol (Livesley *et al.*, 2014; Xiao et McPherson, 2016), (2) augmentant la vitesse à laquelle les pluies s'infiltrent dans le sol et la capacité du sol à stocker l'eau grâce aux racines, (3) réduisant l'érosion du sol en diminuant l'impact des gouttes de pluie sur les surfaces stériles, et (4) réduisant l'humidité du sol et augmentant de ce fait la capacité de stockage du sol via le processus d'évapotranspiration (McPherson *et al.*, 2007). L'augmentation du couvert forestier constitue donc un moyen simple qui permet d'aider à contrôler le ruissellement et à réduire sa quantité globale.

1.3 L'approche fonctionnelle

Jusqu'à récemment le choix des essences d'arbres ainsi que leur emplacement de plantation dans les villes étaient motivés par des critères d'esthétisme, d'acceptabilité par les populations et de tolérance envers certains stress rencontrés en milieux urbains tels que la présence de sels de déglaçage, la compaction du sol ou encore la pollution (Paquette *et al.*, 2016). Cette stratégie est problématique. En effet, seulement quelques espèces sont utilisées de façon répétitive et composent la vaste majorité des arbres en ville. À Montréal, par exemple, les érables constituent 41% des arbres en ville suivis par les frênes avec 21% (Paquette *et al.*, 2016). De plus, on constate une grande ressemblance entre les caractéristiques (traits fonctionnels) de ces arbres ayant pour conséquences une faible résilience des arbres et une sensibilité à des menaces similaires.

Les traits fonctionnels représentent des caractéristiques morphologiques, physiologiques ou phénologiques d'un organisme ayant un effet sur sa performance individuelle en termes de délivrance de services écosystémiques et déterminant sa réponse face à un ou plusieurs facteurs environnementaux (Cameron et Paquette, 2016). Les traits fonctionnels comprennent, entre autres : le type de feuilles, la densité du bois ou encore la tolérance aux inondations et à la sécheresse. Ces traits, indépendants du genre ou de la famille, permettent de montrer, par exemple, que deux érables forment une communauté moins diversifiée qu'un érable et une épinette. Plus la diversité fonctionnelle est importante, mieux la communauté d'arbres étudiée réagit face à une altération de l'environnement. Cette quantification permet de refléter la diversité pour un peuplement donné mais aussi de mieux définir des objectifs de diversification et de quantifier la différence à combler pour les atteindre (Cameron et Paquette, 2016).

Cependant, l'analyse de la diversité fonctionnelle demande des connaissances des

valeurs quantitatives des traits d'arbres et des analyses spécifiques (Paquette *et al.*, 2016). De plus, le calcul des indices de diversité fonctionnelle est peu utile pour des fins de communication (Cameron et Paquette, 2016).

Une autre technique, basée sur la diversité fonctionnelle est l'utilisation de groupes fonctionnels (Mason *et al.*, 2005). Ces groupes sont basés sur les traits fonctionnels et sont formés à l'aide de techniques de regroupement hiérarchique.

Ces groupes fonctionnels, une fois formés et rendus disponibles sous forme de listes, ont pour objectif d'aider les gestionnaires et citoyen.e.s dans leur choix de plantation en facilitant la détection des problèmes d'abondance de groupes nuisant à la diversité. Ils permettent ainsi d'informer les pépinières des intentions d'achat des villes pour les prochaines années (Cameron et Paquette, 2016).

Dans ce chapitre, nous avons introduit les principales notions relatives au domaine de la foresterie urbaine auxquelles nous nous référerons dans ce travail. Ce chapitre a pour but de faciliter la lecture et la compréhension de ce mémoire.

CHAPITRE II

ÉTAT DE L'ART

2.1 Les systèmes de recommandation

Les systèmes de recommandation (*Recommender systems* ou *RS*) jouent un rôle de plus en plus important dans les applications en ligne caractérisées par une très grande quantité de données (Cremonesi *et al.*, 2011). Ils reposent sur plusieurs domaines de recherche, notamment le filtrage et la recherche d'informations, l'interaction humain-machine (*Human-Computer Interaction*) et l'intelligence artificielle. Les systèmes de recommandation sont utilisés dans diverses applications en aidant les personnes utilisatrices à trouver de nouveaux articles ou services susceptibles de les intéresser (Cremonesi *et al.*, 2011). Ces systèmes jouent également un rôle important dans la prise de décision en aidant à maximiser des profits (Qu *et al.*, 2014) ou à minimiser des risques (Bouneffouf *et al.*, 2013).

Il existe aujourd'hui plusieurs stratégies pour le développement de systèmes de recommandation (Lu *et al.*, 2015a). Dans ce chapitre nous décrivons les approches les plus populaires dans la littérature ainsi que celles que nous estimons les plus adéquates pour nos travaux.

Ricci, Rokach et Shapira (Ricci *et al.*, 2011) proposent une introduction complète aux systèmes de recommandation qui rassemble les acquis nécessaires pour élabo-

rer des algorithmes performants dont l'usage est adapté à une situation donnée. Les auteurs présentent les techniques relatives aux systèmes de recommandation, leurs applications et évaluations, l'interaction avec ces systèmes, les communautés et les systèmes de recommandation et enfin, certaines techniques et algorithmes avancés.

Le travail de Melville et Sindhwani (Melville et Sindhwani, 2017) permet d'entrer dans le détail des systèmes de recommandation et de découvrir les trois principales catégories d'approches :

— Filtrage collaboratif (*Collaborative filtering*),
— Filtrage basé sur le contenu (*Content based filtering*),
— Approches hybrides (*Hybrid approaches*).

Le filtrage collaboratif est l'approche la plus populaire (Jannach *et al.*, 2012), notamment parce qu'elle est indépendante du domaine et ne nécessite pas d'informations spécifiques concernant les éléments à recommander. Cette approche utilise les données de personnes utilisatrices similaires dites « voisines » afin de déterminer si une recommandation pourrait être pertinente pour une personne donnée. La similitude des préférences de deux personnes est calculée en fonction de la similitude de l'historique des évaluations des personnes utilisatrices.

Le filtrage basé sur le contenu utilise les caractéristiques et les préférences de chaque utilisateur, en tenant compte des anciennes sélections de cette personne. La similitude des éléments est calculée en fonction des caractéristiques associées aux éléments comparés. Par exemple, si une personne a évalué positivement un film appartenant au genre science fiction, le système apprendra à lui recommander d'autres films du même genre. En résumé, les techniques classiques de recommandation basées sur le contenu ont pour objectif de faire correspondre les attributs du profil utilisateur avec les attributs des éléments.

Bien que les techniques citées ci-dessus constituent les deux approches les plus populaires, beaucoup de techniques « hybrides » ont vu le jour dans le but de surpasser les limitations des méthodes de filtrage collaboratif et de filtrage basé sur le contenu.

Les limitations les plus évidentes sont le démarrage à froid (*cold start*), les matrices creuses (*sparse matrix*) ou encore la nécessité d'avoir une quantité suffisante de méta-données relatives aux objets à recommander (Sharma et Mann, 2013). Le démarrage à froid se produit lorsqu'aucune information de préférence n'est disponible pour un utilisateur ou un élément donné (Volkovs *et al.*, 2017). Les matrices creuses sont dues au fait qu'une personne utilisatrice n'interagit qu'avec un très petit sous-ensemble des éléments recommandables par le système (Shi *et al.*, 2014). Enfin, le filtrage basé sur le contenu nécessite une description détaillée des éléments (méta-données) et un profil utilisateur très bien organisé avant que la recommandation ne puisse être faite (Isinkaye *et al.*, 2015). Les approches hybrides vont combiner les deux types de filtrage pour atteindre une plus grande efficacité.

Cependant, ces dernières années, la communauté de recherche a réalisé que l'efficacité des systèmes de recommandation ne réside pas seulement dans la précision des recommandations (McNee *et al.*, 2006). Cette focalisation du domaine sur les mesures de précision a même déjà été considérée comme potentiellement nuisible (McNee *et al.*, 2006). En effet, les recommandations les plus précises selon les métriques standards ne sont parfois pas les recommandations les plus utiles aux personnes utilisatrices. Par exemple, la similarité entre les éléments d'une liste de recommandation ou encore la sérendipité ne peuvent être mesurées à travers l'évaluation de la précision et pourraient donc impacter la qualité des recommandations (Ge *et al.*, 2010). Il est donc nécessaire de développer de nouvelles approches permettant aux communautés d'usage de contribuer à améliorer les recommanda-

tions (Bakalov *et al.*, 2013; He *et al.*, 2016), et ce, en mettant davantage l'accent sur l'expérience utilisateur des systèmes de recommandation (Konstan et Riedl, 2012).

Lu *et al.* (Lu *et al.*, 2015b) ont passé en revue les développements applicatifs des systèmes de recommandation et résument les techniques de recommandations les plus utilisées dans 8 catégories distinctes allant du e-business à l'e-gouvernance. Leur étude porte sur quatre dimensions : les méthodes de recommandation, les logiciels de recommandation, le domaine d'application et les plateformes d'application (tel que le mobile). Leurs travaux donnent un bon aperçu des différentes techniques de recommandations appliquées à des situations réelles et permettent de mettre en évidence les principaux problèmes dont souffrent les systèmes tels que les problèmes de confiance, de contexte, mais aussi de diversification des recommandations.

Les travaux de Adomavicius et Tuzhilin (Adomavicius et Tuzhilin, 2005) passent en revue plusieurs approches de recommandation, mettent en évidence les limites des méthodes de recommandation actuelles et discutent des extensions possibles afin de fournir de meilleures recommandations. Ces extensions incluent, entre autres, la modélisation des communautés d'usage et des éléments recommandés, l'intégration d'informations contextuelles dans le processus de recommandation et la flexibilité des systèmes de recommandation.

Dans la continuité, les travaux de Parra, Brusilovsky et Trattner (Parra *et al.*, 2014) se concentrent sur l'amélioration de la transparence et de la contrôlabilité par la personne utilisatrice au cours du processus de recommandation dans le contexte d'un système hybride. Les auteurs ont développé une nouvelle interface de recommandation hybride visuelle permettant aux utilisateurs de fusionner et de contrôler manuellement l'importance des stratégies de recommandation. Les

personnes utilisatrices peuvent inspecter les résultats de la fusion à l'aide d'une visualisation interactive de diagrammes de Venn pour examiner et filtrer les éléments recommandés par une ou plusieurs méthodes. Ces travaux présentent les résultats de deux études explorant les apports de l'approche proposée dans l'amélioration de l'expérience utilisateur. Ces études ont démontré que l'interface contrôlable améliorait l'expérience utilisateur et que la version visuelle pouvait offrir de meilleures performances de classement.

Les travaux de Steck, Van Zwol et Johnson (Steck *et al.*, 2015) sous forme de didacticiel décrivent les différents aspects cruciaux pour une expérience utilisateur fluide et efficace dans le contexte des systèmes de recommandation interactifs. Ainsi, ils mettent en évidence les différents défis spécifiques à ces derniers tels la façon de guider la personne utilisatrice de manière intuitive dans le processus de recommandation, le choix du bon équilibre entre l'exploitation des retours des personnes utilisatrices et la possibilité d'explorer le catalogue de recommandations.

He, Parra et Verbert (He *et al.*, 2016) offrent un état de l'art des systèmes de recommandation interactifs. Leur étude s'intéresse à la visualisation d'informations de recommandation, qui consiste en l'utilisation de représentations visuelles interactives des données pour faciliter leur compréhension. Ils proposent de plus un cadre de développement pour les systèmes de recommandation interactifs en intégrant les techniques de visualisation et de recommandation afin de favoriser le retour d'information. Sur la base du cadre développé, ils analysent les systèmes de recommandations interactifs existants, évaluent leurs avantages et inconvénients. Enfin, plusieurs perspectives de recherches intéressantes sont présentées telles l'incorporation des émotions dans le processus de recommandation ou les moyens d'acquisition d'informations contextuelles.

Les précédents travaux ont été confirmés par ceux de Valdez, Ziefle et Verbert (Val-

dez *et al.*, 2016). Ils ont analysé toutes les publications sur les systèmes de recommandation de la base de données Scopus [1], et en particulier les articles présentant des approches mettant en avant les interactions humains-machines. Les résultats de cette étude montrent que les futures recherches dans le domaine doivent être orientées vers :

- Une prise en charge de niveaux d'interactions plus avancés (*i.e.* contrôles définissant quelles données peuvent être suivies et prises en compte et à quelles fins),
- La nécessité d'adapter les systèmes de recommandation et leurs interfaces utilisateur à différentes caractéristiques personnelles et contextuelles,
- Le développement de systèmes de recommandation incorporant à la fois des données acquises automatiquement et des révisions par les personnes utilisatrices, afin de personnaliser les recommandations en fonction de leurs besoins.

Selon les études les plus récentes dans le domaine des systèmes de recommandation (Karimi *et al.*, 2018; Eirinaki *et al.*, 2018; Villegas *et al.*, 2018), plusieurs approches gagnent en popularité dans la littérature, notamment les :

- Systèmes de recommandation basés sur les connaissances (*Knowledge-based recommender system*) (Burke, 2000). Ce type d'approche a contrario des approches classiques citées précédemment est basé non pas sur l'historique des préférences utilisateurs mais sur une connaissance explicite du domaine. La manière dont certaines caractéristiques d'un élément recommandable répondent aux besoins et aux préférences de la personne utilisatrice, le contexte et l'utilité de l'élément pour la personne utilisatrice sont pris en compte,
- Systèmes de recommandation démographique (*Demographic-Based recom-*

1. `https://www.scopus.com`. Visité le 21 Mars 2021

mender systems) (Safoury et Salah, 2013). Ces systèmes sont généralement considérés comme une sous-catégorie des systèmes basés sur le contenu. Dans cette approche, chaque profil utilisateur est catégorisé en fonction de ses attributs dans une classe démographique. Ainsi, les recommandations sont générées en fonction de la classe démographique de la personne utilisatrice, ce qui permet ainsi de s'affranchir de la nécessité d'avoir un historique des préférences/évaluations des utilisateurs,

— Systèmes de recommandation contextuel (*Context-aware recommender system*). Ce type d'approche génère des recommandations plus pertinentes en les adaptant à la situation contextuelle spécifique de la personne utilisatrice (Adomavicius et Tuzhilin, 2011). L'approche contextuelle combine des informations provenant de plusieurs sources pour affiner l'espace contextuel et résoudre les problèmes majeurs du système de recommandation tels que le passage à l'échelle et le problème du démarrage à froid (Song *et al.*, 2014).

Ces dernières années les techniques d'apprentissage automatique ont connu beaucoup de succès en résolvant des tâches complexes dans divers domaines tels le traitement du langage naturel ou encore la vision par ordinateur. Portugal, Alencar et Cowan (Portugal *et al.*, 2018) ont effectué une revue systématique de 121 études afin d'analyser entre autres l'utilisation des algorithmes d'apprentissage automatique dans les systèmes de recommandation et les mesures de performance principales et alternatives. Ainsi, les auteurs montrent que les algorithmes d'apprentissage supervisé ou non supervisé sont de plus en plus utilisés dans la littérature, notamment, les algorithmes de regroupement (*clustering*), les méthodes d'ensembles et les séparateurs à vaste marge (*Support vector machine*) (Cortes et Vapnik, 1995). Enfin, les auteurs de l'étude rapportent que l'erreur absolue moyenne (équation 3.3.3), la précision (équation 4.6), le rappel (équation 4.7) et la F-mesure (équation 4.8) sont les mesures de performance les plus utilisées

pour évaluer les algorithmes d'apprentissage automatique, tandis que l'évaluation de la couverture est la mesure alternative la plus utilisée pour refléter comment les recommandations générées couvrent le catalogue des éléments disponibles à la recommandation (Herlocker *et al.*, 2004).

Dans la continuité, l'étude de Zhang, Yao, Sun et Tay (Zhang *et al.*, 2019) confirme cette tendance à l'utilisation de systèmes de recommandation basés sur l'apprentissage profond qui ont attiré une attention considérable en surmontant les obstacles des modèles conventionnels et en obtenant une qualité de recommandation élevée. En effet, les modèles d'apprentissage profond sont capables de capturer efficacement des relations utilisateur/élément non-linéaires et non triviales. Ils permettent la codification d'abstractions plus complexes en tant que représentations de données.

2.2 La quantification des services écosystémiques

Au cours des dernières années, les effets causés par le changement climatique se sont fait de plus en plus ressentir à travers le monde. Des organismes tels que les Nations unies recommandent l'établissement de stratégie d'urbanisation durable afin de contrer ces effets (DESA, 2018). Dans ce contexte, les avantages des arbres en milieu urbain sont de plus en plus reconnus notamment comme étant un moyen de lutte contre les effets des changements globaux à travers les services écosystémiques délivrés tels que l'amélioration de la qualité de l'air, la conservation de l'énergie, l'interception des eaux pluviales, et la réduction du dioxyde de carbone atmosphérique (McPherson *et al.*, 2007). La reconnaissance de ces avantages a motivé les décideurs en matière d'infrastructures vertes à mettre en oeuvre des programmes ambitieux de plantation d'arbres, en particulier en Amérique du Nord. Cependant, si les programmes de plantation ambitieux sont

louables, leur succès dépend fortement de la sélection d'arbres bien adaptés à leurs sites de plantation. En effet, la production de services écosystémiques est améliorée avec la croissance des arbres (Hirons et Sjöman, 2019), il faut donc s'assurer de la survie des espèces d'arbres jusqu'à maturité. Cela nécessite que les espèces soient soigneusement adaptées à leur site de plantation et que leurs vulnérabilités aux ravageurs et aux agents pathogènes soit minimisées (Sjöman *et al.*, 2018). Dans cette logique, plusieurs études récentes fournissent des méthodologies et outils d'évaluation et d'aide à la décision à destination des gestionnaires des forêts urbaines dans le but d'améliorer la planification, la conception et la gestion de ces forêts.

L'étude de Haase *et al.* (Haase *et al.*, 2014) offre une introduction à l'évaluation des services écosystémiques (SE) en milieu urbain, les types de SE les plus étudiés et les modèles les plus utilisés pour leur quantification et leur évaluation. Les auteurs ont effectué une revue de la littérature basée sur 217 études. La conclusion des travaux met en évidence (de manière non-exhaustive) la rareté des études traitant de la dynamique temporelle et spatiale des services écosystémiques malgré leur importance pour l'urbanisme. De plus, il existe un manque d'études mettant en évidence les compromis et synergies entre les différents SE. Enfin, les auteurs concluent qu'une évaluation spatialement explicite des SE à une résolution relativement élevée (au niveau des quartiers et rues) sera essentielle pour intégrer les évaluations et conclusions des différentes études dans les stratégies de planification et gestion urbaines.

La revue de littérature de Lin *et al.* (Lin *et al.*, 2019) montre l'intérêt croissant de la communauté de recherche pour la quantification des services écosystémiques. Les auteurs ont identifié 242 travaux et 476 cas d'études pour la période de 1996-2017 avec plus de la moitié des études publiées entre 2012 et 2017. Deux grandes catégories de modèles numériques sont utilisées pour l'analyse des forêts urbaines :

les modèles à usage général et les modèles spécifiques aux forêts urbaines. Parmi ces derniers, i-Tree (i-Tree, 2021) est probablement le calculateur de bénéfices d'arbres le plus connu et le plus utilisé. Il s'agit d'une suite d'outils conçus par *USDA Forest Service* pour quantifier les services environnementaux fournis par une forêt ou un groupe d'arbres, ainsi que la structure de cette forêt. Les auteurs identifient plusieurs outils appartenant à la même suite, mis en oeuvre à travers plusieurs études :

- *i-Tree Eco* (anciennement *UFORE*), qui quantifie la structure, les menaces, les avantages et les valeurs fournis par les arbres urbains, a été le plus souvent mis en oeuvre,
- *Streets* (anciennement *STRATUM*) permettant de quantifier la valeur monétaire des SE (aujourd'hui inclus dans i-Tree Eco),
- *Canopy* qui permet une estimation de l'étendue de la canopée de la forêt urbaine d'un territoire à l'étude à partir d'une photo aérienne de *Google Earth*,
- *Hydro* qui est un outil permettant d'effectuer des analyses comparatives de couverture forestière et d'évaluer leur impact hydrologique potentiel à différentes échelles,
- *Species* conçu pour aider les forestiers urbains à sélectionner les espèces d'arbres les plus appropriées en fonction des services environnementaux potentiels et de la zone géographique de l'espèce.

Les auteurs présentent par ailleurs d'autres modèles largement utilisés. Tout d'abord, on peut citer l'usage de ENVI-met (ENVI-met, 2021) permettant la modélisation et la comparaison de scénarios de paysages conçus ou réels (par exemple, avec/sans arbres, disposition des arbres et bâtiments). Puis, CFD (Buccolieri *et al.*, 2018) qui est un ensemble de modèles basés sur les lois fondamentales de la mécanique des fluides et de la thermodynamique, permet par exemple de quantifier les effets thermiques des arbres sur les bâtiments environnants. Ces modèles sont principa-

lement utilisés car ils sont disponibles gratuitement et contiennent divers modules pour différentes applications.

ENVI-met et CFD requièrent de nouveaux paramètres relatifs au site d'étude lors de leur application en dehors de leurs domaines de modélisation d'origine, contrairement à i-Tree, qui peut être utilisé dans de nouveaux emplacements ou de nouvelles conditions sans recalibrage des paramètres du modèle. Aussi, l'une des limites de CFD est la nécessité pour la personne utilisatrice d'avoir des connaissances en programmation.

Concernant les modèles statistiques, toujours selon Lin et al (Lin *et al.*, 2019), ces derniers sont souvent axés sur l'économie notamment l'estimation de l'impact des arbres urbains sur les valeurs immobilières ainsi que l'économie d'énergie. Il est à noter aussi que la majorité des articles mettant en oeuvre des modèles statistiques emploient des modèles de régression hédonique afin d'estimer, par exemple, la contribution des services écosystémiques à la valeur d'une propriété.

Une autre caractéristique ayant également été étudiée par les auteurs est l'évaluation de l'incertitude. Cette dernière est une caractéristique fondamentale de tout modèle et est généralement due à un manque de connaissance du domaine ou à des informations incomplètes. Ainsi, l'évaluation de l'incertitude des modèles est généralement négligée ou sous évaluée. En effet, aucune des études n'aborde l'incertitude dans la communication des résultats des modèles mis en oeuvre. Cela pourrait s'expliquer par les raisons suivantes : (1) le fait que les modèles évaluant les interactions écosystémiques complexes comme I-Tree ou ENVI-met devraient subir des changements importants dans leur architecture, notamment en ce qui concerne les formules utilisées et la paramétrisation, si une évaluation complète et approfondie de l'incertitude était faite ; (2) les méthodes d'évaluation de l'incertitude sont bien développées mais aucune méthode n'est universellement

applicable et efficace pour tous les modèles. Ainsi, Lin et al (Lin *et al.*, 2019) préconisent en particulier que pour les modèles axés sur l'aide à la décision, les futures modélisations devraient se concentrer sur l'amélioration de l'évaluation et de la communication de l'incertitude.

Dans ce chapitre, nous avons pu résumer différents travaux académiques permettant d'expliquer nos travaux de recherche. Nous constatons qu'à ce jour très peu de travaux — aucun à notre connaissance — ne se sont concentrés sur la création d'un système de recommandation à code source ouvert et permettant des analyses spatialement explicites dans le but d'optimiser la résilience de la forêt urbaine. Dans le chapitre suivant, nous aborderons la méthodologie que nous avons employée pour le développement de notre outil.

CHAPITRE III

REGROUPEMENT ET PRÉDICTIONS

Avant de présenter la méthodologie, il est important de comprendre dans quel contexte ce projet s'insère. Notre travail s'est effectué dans un contexte largement multidisciplinaire combinant foresterie urbaine, écologie fonctionnelle et informatique. Ainsi, nos travaux reposent principalement sur des approches informatiques et basées sur l'intelligence artificielle pour répondre aux différentes problématiques rencontrées.

Au cours de ce travail de recherche, plusieurs problématiques ont été abordées. Tout d'abord, nous commençons par une présentation des données dans la partie 3.1.

Ensuite, nous présentons la méthodologie mise en œuvre pour effectuer le regroupement fonctionnel dans la partie 3.2.

Enfin, nous avons également tenté de développer un modèle de prédiction de croissance d'arbres en milieu urbain que nous détaillons dans la partie 3.3.

3.1 Données

3.1.1 Arbres urbains de Montréal

Afin de procéder à nos expérimentations, nous avions besoin d'un inventaire des arbres urbains publics. Nous avons donc utilisé l'inventaire de la ville de Montréal, collecté et mis à jour par les inspecteurs et inspectrices de la ville dans chacun des arrondissements et disponible au format csv (*Comma-separated values*) sur le portail de données ouvertes de la ville (Ville de Montréal, 2021a). Cet inventaire contient des données des arbres de propriété municipale : les arbres de rue et hors rue (parcs et places publiques). Il recense 318 309 arbres (en date du 16 Mars 2021) et inclut 22 variables consignant des informations sur chaque arbre, telles que l'espèce, le diamètre à hauteur de poitrine (DHP), la date de mesure, la date de plantation et l'emplacement géolocalisé exprimé en latitude et longitude. Le tableau 3.1 offre un aperçu des champs disponibles.

Colonne	**Spécification**	**Valeur**
Invent	Type d'arbre inventorié.	H pour hors rue, R pour arbre de rue
Arrond	Numérotation de l'arrondissement dans la base de données.	De 1 à N
Nom_arrond	Nom de l'arrondissement.	
Rue	Nom de la rue sur laquelle l'arbre est positionné, pour les arbres des rues.	
Coté	Coté de la rue.	N = Nord, S = Sud, E = Est, O = Ouest.
No_civique	Numéro civique de la résidence ou du bâtiment sur lequel ou en face duquel l'arbre est situé, lorsque disponible.	

EMP_NO	Numérotation correspondant à l'emplacement de l'arbre dans la base de données. Elle est propre aux arbres de rue et hors rue.	
Emplacement	Lieu physique, extension de terrain où se trouve l'arbre.	Banquette gazonnée, Banquette asphaltée, Fond de trottoir, Parc, Parterre gazonné, Parterre asphalté, Parterre, Terre plein, Trottoir entre autres.
Coord_X et Coord_Y	Coordonnée x et y du point central de l'emplacement Projection : Québec Modified Transverse Mercator (NAD83) – Québec MTM Zone 8.	
Sigle	Acronyme composé des deux premières lettres du genre, de l'espèce et du cultivar, si appliquable du nom Latin.	Exemple : Sigle : GYDI provenant du nom latin : Gymnocladus dioicus
Essence_Latin	Nom latin de l'essence.	
Essence_Fr	Nom français de l'essence.	
Essence_Ang	Nom anglais de l'essence.	
DHP	Mesure du Diamètre du tronc d'un arbre à la Hauteur de la Poitrine qui équivaut à 1,4 m. à partir du plus haut niveau du sol.	
Date_relevé	La date de la dernière prise du DHP	
Date_plantation	Date de la plantation de l'arbre, lorsque disponible.	
Propriété	Propriétaire de l'arbre	Ville, Copropriété : Propriété repartie entre la Ville et une entité privée, Privé

Localisation	Position de l'arbre en fonction d'un immeuble ou chaînage et direction à partir du coin du dernier immeuble.	
Code_parc	No. d'index du parc. Numérotation unique qui représente un parc dans la Ville.	
Nom_parc	Nom du parc.	
Latitude et Longitude	Position géographique latitude et longitude (WGS 84)	

Tableau 3.1: Répertoire des données sur les arbres publics du territoire de la Ville de Montréal.

Il est à noter que cette base de données contient beaucoup de valeurs manquantes ainsi que des valeurs mal renseignées notamment pour les noms d'espèces d'arbres ou encore les dates de plantation souvent absentes, particulièrement pour les arbres ayant un DHP important. De plus, certaines données sont interverties notamment les dates de plantation et les dates de mesures. Afin de pouvoir exploiter cet ensemble, nous l'avons soumis à un processus de pré-nettoyage que nous détaillerons dans la partie 3.3.1.

3.1.2 Historique de la croissance des diamètres à hauteur de poitrine

Dans le but d'entraîner un algorithme d'apprentissage automatique pour la prédiction des diamètres à hauteur poitrine (DHP), nous avons utilisé l'inventaire de l'évolution de la croissance des arbres publics de la ville de Montréal disponible au format csv sur le portail de données ouvertes de la ville (Ville de Montréal, 2021a). Cet inventaire est constitué de 636 524 observations (en date du 16 Mars 2021) et de cinq variables comprenant les informations suivantes :

— une numérotation correspondant à l'emplacement de l'arbre,

— l'emplacement (coté rue ou hors rue),

— la mesure du diamètre à hauteur de poitrine,

— la date de prise de mesure,

— le nom latin de l'essence.

3.1.3 Traits fonctionnels

Colonne	**Spécification**	**Unité / valeur**
Code	Code d'identification de l'espèce	
SeedM	Masse sèche de 1000 graines	g
Wd	Densité du bois moyenne	g/cm3
LMA	Rapport entre la masse sèche des feuilles et la surface	g/m2
Nmass	Teneur en Azote	mg/g
Shade-T	Indice de tolérance à l'ombre	Relation LMA - Intensité lumineuse
Drought-T	Indice de tolérance à la sécheresse	
Waterlog	Indice de tolérance à la saturation en eau du sol	
AM	Association endomycorhizienne	1 = Observé 0 = Jamais observé
ECM	Association ectomycorhizienne	1 = Observé 0 = Jamais observé
A.G	Division phylogénétique	A = Angiosperme G = Gymnosperme
Disp	Moyen primaire de dispersion des graines	A = Animale W = Vent H = Eau, D = Méthodes provenant de la plante mère U = Dispersion non assistée

Tableau 3.2: Table des traits fonctionnels

Le tableau 3.2 représente les traits fonctionnels utilisés pour notre étude. Ces traits ont été mis à notre disposition par la chaire de recherche CRSNG/Hydro-Québec sur le contrôle de la croissance des arbres. Cette base de données est constituée de 11 traits pour 271 espèces provenant de la littérature, représentant les traits les plus documentés et sélectionnés en raison de leur importance écologique (Ruiz-

Benito *et al.*, 2014).

Parmi ces données figurent des indices de tolérance à l'ombre (*Shade-T*), à la sécheresse (*Drought-T*) et à l'inondation (*Waterlog-T*), liés à la performance des arbres en milieu urbain. Selon Cameron et Paquette (Cameron et Paquette, 2016), la tolérance à l'inondation représente la tolérance de l'arbre à des conditions anaérobiques (en l'absence d'oxygène) similaires à celles créées par la compaction des sols en milieu urbain.

La densité du bois (*Wd*) et la masse sèche des graines (*SeedM*) représentent quant à elles des caractéristiques écologiques fondamentales pouvant distinguer les espèces entre elles, notamment leur vitesse de croissance ou encore la tolérance à la compétition pour l'accès aux ressources nécessaires à leur évolution.

La base de données contient aussi la masse foliaire (*LMA*), représentant le rapport entre la masse sèche des feuilles et la surface, et la teneur en azote (*Nmass*) qui sont des traits foliaires, donc relatifs aux feuilles des arbres.

De plus, la base de données consigne également deux variables binaires relatives à la présence/absence d'endomycorhizes arbusculaires (*AM*) ou d'ectomycorhizes (*ECM*) qui sont des types de champignons colonisant les racines qui créent une association permettant une meilleure captation de l'eau et des nutriments par le système racinaire des arbres.

À ces traits, s'ajoute le mode de dispersion des graines (*Disp*) lié à la stratégie de vie de l'arbre. Selon Cameron et Paquette (Cameron et Paquette, 2016), ce trait n'est pas directement utile pour la gestion des arbres en milieu urbain, mais permet de révéler certaines stratégies adoptées par les arbres et pourrait donc être une variable discriminante.

Enfin, le dernier trait est la division phylogénétique (*A.G*) prenant la valeur A ou

G respectivement pour Angiospermes et Gymnospermes. Les angiospermes sont généralement des plantes à fleurs et représentent la plus grande proportion des espèces végétales terrestres tandis que les gymnospermes sont des plantes à graines nues, généralement des conifères.

3.2 Regroupement fonctionnel

Afin de formuler une stratégie de gestion des arbres urbains capable de cibler la diversité et la résistance des forêts urbaines, plusieurs concepts clés doivent être explorés et intégrés dans une approche unique et facile à utiliser qui permettra une prise de décision plus efficace. C'est dans ce sens que l'approche de regroupement fonctionnel proposée par Cameron et Paquette (Cameron et Paquette, 2016) a été développée. Cette approche est fondée sur l'hypothèse que les communautés fonctionnellement diverses ont une résistance plus élevée en raison de la probabilité qu'au moins certaines espèces réagissent différemment aux conditions variables et aux facteurs de stress. Cette différence de réaction et de sensibilités est reflétée à travers les groupes formés et leur diversification peut conduire à une meilleure résilience de la forêt urbaine. Enfin, l'avantage offert par les groupes fonctionnels est qu'ils permettent de fournir un moyen simple de catégoriser les espèces selon leur rôle fonctionnel plutôt que selon leur simple classification botanique.

Considérant les avantages cités ci-dessus, nous avons décidé de bâtir notre approche de recommandation sur la suggestion des groupes fonctionnels aux personnes utilisatrices plutôt que d'une liste d'espèces d'arbres. Ce regroupement nous permet aussi, lors de l'analyse de la forêt urbaine, de fournir une vue de la diversité fonctionnelle qui permet de résumer facilement la diversité effective. Recommander un ou plusieurs groupes fonctionnels, dans lesquels les personnes utilisatrices peuvent sélectionner des essences d'arbres à planter, plutôt qu'une

ou plusieurs espèces, permet de satisfaire les contraintes auxquelles les municipalités pourraient être soumises, *i.e.* la disponibilité des espèces en pépinière ou les contraintes provenant de l'environnement local. De plus, la maîtrise et l'intégration de la technique de regroupement fonctionnel au sein de notre outil nous permettent, lorsque les traits fonctionnels sont disponibles, d'effectuer un regroupement automatique pour de nouvelles régions du monde. Enfin, en collaboration avec l'équipe d'expert.e.s en foresterie urbaine, nous avons tenté d'autres techniques de regroupement afin de comparer les résultats avec l'approche de regroupement fonctionnel basée sur le regroupement hiérarchique.

Le regroupement hiérarchique est une méthode de classification itérative généralement implémentée comme suit (Govender et Sivakumar, 2020) :

1. Chaque observation est considérée comme un groupe.
2. Les distances entre les groupes sont calculées à l'aide d'une mesure de similarité.
3. Deux groupes étant à une distance minimale (en d'autres termes, les éléments les plus proches) sont combinés et remplacés par un seul groupe. La matrice de distance est ensuite recalculée afin de refléter ce processus de fusion.
4. Répétition des étapes 2 et 3 jusqu'à ce qu'il n'y ait qu'un seul groupe contenant toutes les observations.

Le résultat d'un algorithme de regroupement hiérarchique est un dendrogramme (illustré par la figure 3.1), qui est une structure arborescente représentant la séquence de groupes imbriqués (Dubes et Jain, 1976). La distance de chaque fusion (ou division) est également représentée sur la structure. Couper le dendrogramme (comme illustré par la ligne discontinue rouge dans la figure 3.1) à un niveau souhaité donne un ensemble de groupes disjoints (Govender et Sivakumar, 2020).

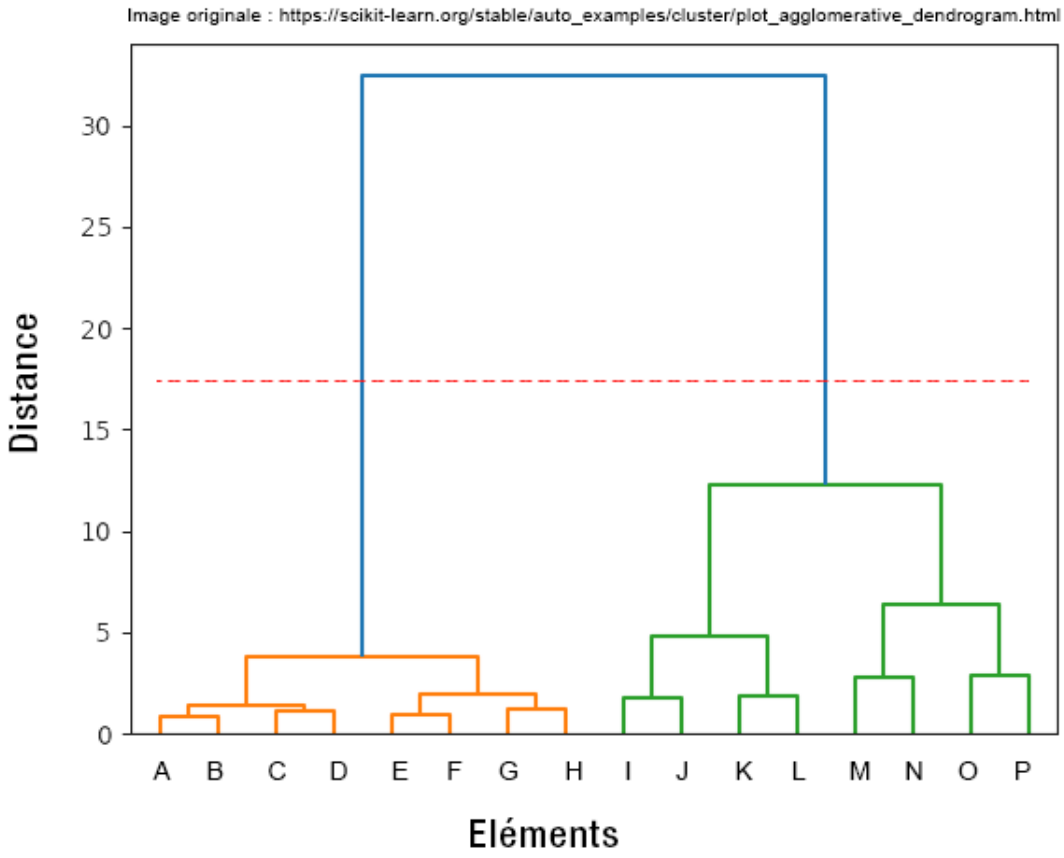


Figure 3.1: Exemple de dendrogramme

Ainsi, un algorithme de regroupement hiérarchique nécessite de : (1) définir une distance entre les individus puis construire leur matrice de distance. Le choix de la mesure de distance est à faire au préalable. Cette dernière dépend des données étudiées et des objectifs, et (2) choisir un critère d'agrégation.

Ainsi, la fonction de calcul de la matrice de distance est importante et influe sur les résultats du regroupement. Étant donné la nature de l'ensemble de données qui est composé de traits de différents types : ordinaux, nominaux et numérique, il est judicieux de calculer les distances de Gower (Gower, 1971). La distance de Gower est une des mesures de proximité les plus populaires pouvant être utilisées

pour calculer la distance entre deux entités dont les attributs sont composés de valeurs quantitatives, ordinales et nominales. Ainsi, pour chaque type de variable, une distance appropriée est calculée. La distance est toujours un nombre compris entre 0 (identique) et 1 (dissemblable). Puis, une combinaison linéaire utilisant des poids spécifiés par la personne utilisatrice est calculée afin de créer une matrice de distance finale qui sera utilisée pour le regroupement.

De plus, afin de construire le dendrogramme, nous avons choisi la méthode de Ward (Ward, 1963) comme critère d'agrégation. De manière simplifiée, cette méthode cherche à minimiser l'inertie intra-classe et à maximiser l'inertie inter-classe afin d'obtenir des classes les plus homogènes possibles. Intuitivement, l'inertie indique la distance qui sépare les points d'un groupe.

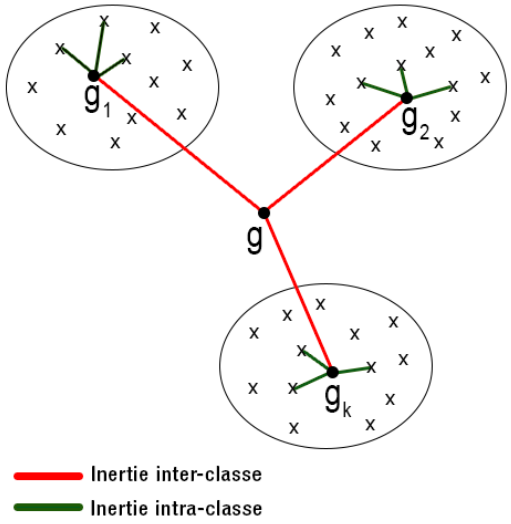


Figure 3.2: Illustration de l'inertie interclasse et intraclasse

Ainsi, si $G = \{e_i\ ,\ i \in \{1, ..., p\}\}$ est un groupe d'individus, de centre de gravité g, partitionné en k classes d'effectif n_1,n_2,..., n_k qu'on appellera G_1, G_2,..., G_k qui ont pour centres de gravité g_1, g_2, ..., g_k et d étant la distance entre les individus, alors :

L'inertie inter-classe (I_{inter}) est égale à :

$$I_{inter} = \frac{1}{N} \sum_{i=1}^{k} n_i \times d(g_i, g)^2 \tag{3.1}$$

L'inertie intra-classe (I_{intra}) est égale à :

$$I_{intra} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} d(e_j, g_i)^2 \tag{3.2}$$

Ensuite, pour obtenir une partition de la population, il suffit de découper le dendrogramme obtenu à une certaine hauteur. Dans notre cas, nous avons opté pour un découpage en 10 classes selon la méthodologie proposée par Cameron et Paquette (Cameron et Paquette, 2016).

3.2.1 Préparation des données

Dans le but de comparer plusieurs techniques de regroupement (présentées en 3.2.2) avec la technique de regroupement hiérarchique, nous devions effectuer une phase de préparation des données. En effet, chaque technique de regroupement requiert un formatage particulier des données.

Nous avons donc commencé par diviser notre ensemble de données en 3 :

- Ensemble de données contenant seulement les traits numériques : la masse sèche de 1 000 graines (SeedM), la densité du bois moyenne (Wd), le rapport entre la masse sèche des feuilles et la surface (LMA), la teneur en azote (Nmass), l'indice de tolérance à l'ombre (Shade-T), l'indice de tolérance à la sécheresse (Drought-T) et l'indice de tolérance à la saturation en eau du sol (Waterlog-T),
- Ensemble de données contenant seulement les traits catégoriques : l'association endomycorhizienne (AM), l'association ectomycorhizienne (ECM),

la division phylogénétique (A.G) et le moyen primaire de dispersion des graines (Disp),

— Ensemble de données complet, *i.e.* l'ensemble composé des sous-ensembles ci-dessus.

Cette division a été effectuée dans le but de répondre au format de données requis par les algorithmes de regroupement que nous avons sélectionné pour la comparaison.

Encodage à chaud. De nombreux algorithmes d'apprentissage automatique ne peuvent pas agir directement sur les données catégoriques. Ils exigent que toutes les variables d'entrée soient numériques. Cela signifie que les données catégoriques doivent être converties sous forme numérique. Afin de résoudre ce problème, nous avons donc encodé nos données catégoriques en « encodage à chaud » (*One-hot encoding*). Autrement dit, nous avons transformé nos variables catégoriques en une représentation binaire. Par exemple, pour la variable $A.G$, il y a deux catégories : Angiosperme (A) et Gymnosperme (G), deux variables binaires sont donc nécessaires. La valeur « 1 » est placée pour la catégorie à laquelle appartient l'espèce et « 0 » pour l'autre.

Normalisation des données et discrétisation. Nous avons normalisé les traits numériques à l'aide de la fonction de mise à l'échelle Min-Max (*Min-Max Scaling*) dans un intervalle fixe [0,1]. Les données étant exprimées en plusieurs unités de mesure différentes et variables en magnitudes, ceci a été effectué afin que toutes les caractéristiques contribuent de manière égale lors du regroupement.

Le but d'avoir un tel intervalle restreint est de réduire l'espace de variation des valeurs d'une des dimensions de la représentation des données et par conséquent obtenir des écarts-types plus petits, ce qui peut réduire l'effet des données aber-

rantes contrastant grandement avec les valeurs « normalement » mesurées.

La mise à l'échelle est effectuée selon l'équation suivante :

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.3}$$

Où X est la valeur de l'attribut, X_{min} représente sa valeur minimale et X_{max} sa valeur maximale. X_{norm} est le résultat de la mise à l'échelle de X.

D'autres algorithmes ne prennent en entrée que des données catégoriques. Pour cela nous avons donc discrétisé nos variables. La discrétisation consiste à transformer une variable quantitative en qualitative discrète, en découpant ses valeurs en classes (intervalles). Pour conserver au mieux l'information et rendre compte de la répartition des variables numériques, nous avons discrétisé nos données en trois classes d'égale amplitude : « small », « medium » et « large ». Cependant, il est à noter que l'inconvénient majeur de la discrétisation est qu'elle engendre une perte d'information et également une diminution de la capacité d'analyse et de traitement des données.

Réduction de la dimension. Nous avons réduit la dimension de l'ensemble de données contenant les valeurs numériques en un espace de représentation à deux dimensions en utilisant l'analyse en composantes principales (ACP) (Wold *et al.*, 1987). L'objectif de la réduction de la dimension est de transformer les données originales dans un espace de dimension plus réduite tel que la nouvelle représentation des données soit une combinaison linéaire des données originales. L'ACP a pour objectif de passer d'un espace à D dimensions vers un espace à K dimensions, de sorte que $K < D$ tout en conservant autant d'informations que possible. Cela nous permet par exemple de pouvoir représenter les données sous forme de graphique en déformant le moins possible la configuration initiale mais en pouvant visualiser la structure globale.

En parallèle nous avons réduit la dimension de l'ensemble de données contenant les valeurs numériques en un espace de représentation à deux dimensions à l'aide de l'algorithme T-SNE (*t-distributed stochastic neighbor embedding*) (Maaten et Hinton, 2008). T-SNE est une approche non-linéaire permettant la représentation d'un ensemble de points d'un espace à grande dimension dans un plus petit espace. L'objectif de T-SNE est de trouver une modélisation en faible dimension tout en respectant les proximités entre les points dans l'espace d'origine.

3.2.2 Algorithmes de regroupement utilisés

Une fois la préparation des données effectuée, nous avons procédé à l'implémentation des techniques de regroupement décrites ci-dessous.

K-moyennes. K-moyennes (*k-means*) (Steinhaus, 1956) est l'un des algorithmes d'apprentissage non supervisé le plus couramment utilisé pour la partition d'un ensemble de données en un ensemble de k groupes homogènes. Il s'agit d'une méthode de regroupement basée sur des points qui commence par les centres de groupes initialement placés à des positions arbitraires (centroïde) et se poursuit en déplaçant à chaque étape les centres des groupes afin de minimiser l'erreur de regroupement. Sa popularité s'explique par sa simplicité, sa facilité de mise en œuvre et son efficacité (Jain, 2010). Il est cependant nécessaire de définir k à l'avance. Dans notre cas, k est fixé à 10 car il s'agit du nombre de groupes d'espèces identifiés par Cameron et Paquette (Cameron et Paquette, 2016) lors de l'analyse du résultat du groupement hiérarchique.

La similitude entre une paire d'objets est définie par leur distance. La distance euclidienne est souvent utilisée. La partition divise les données en k groupes de sorte que chaque groupe contienne au moins un élément. Ainsi, selon Hartigan et Wong (Hartigan et Wong, 1979), le but de l'algorithme k-moyennes est de

diviser m objets en dimension n en k groupes ($k \leq n$) de telle sorte que la somme des carrés de distance entre chaque objet et le centroïde d'un même groupe soit minimisée.

Ainsi, étant donné un ensemble de points $X = \{x_i \ , \ i = 1, ..., n\}$, qu'on cherche à regrouper en k groupes $c_t \ , \ t = 1, ..., k$ en minimisant la distance entre les points à l'intérieur de chaque groupe :

$$J = \sum_{t=1}^{k} \sum_{i=1}^{n} ||x_i^{(t)} - \mu_t||^2 \tag{3.4}$$

Où μ_t est le centroïde d'un groupe c_t et J est la fonction objectif à minimiser.

Plus la valeur de J est faible, meilleure sera la qualité du partitionnement obtenu.

Regroupement agglomératif. Le regroupement agglomératif (Sharma *et al.*, 2019) est la méthode de regroupement hiérarchique la plus utilisée. Cette méthode fonctionne de manière ascendante. À l'initialisation, chaque objet est considéré comme un groupe à un seul élément. À chaque étape de l'algorithme, les deux groupes les plus similaires sont combinés dans un nouveau groupe plus grand. Cette procédure est itérée jusqu'à ce que tous les points soient membres d'un seul grand groupe. Pour cette expérimentation, nous avons utilisé une distance euclidienne ainsi que la méthode de Ward comme critère d'agrégation.

Mélange gaussien. Le mélange gaussien *Gaussian Mixture* (Banfield et Raftery, 1993; Ouyang *et al.*, 2004) est une méthode de regroupement probabiliste supposant que les groupes présents dans le jeu de données proviennent de différentes distributions gaussiennes. Autrement dit, on tente de modéliser l'ensemble de données sous la forme d'un mélange de plusieurs distributions gaussiennes. Cet algorithme est bien adapté lorsque le nombre de groupes et leur distribution sont connus.

Partitionnement spectral. Le partitionnement spectral *Spectral Clustering* (Von Lux-

burg, 2007) est un type de partitionnement où les points de données sont reconstruits sous forme de graphe (sous forme de matrice d'adjacence). Dans un tel graphe, chaque noeud correspond à une donnée (ou observation) et chaque arête qui relie deux observations est pondérée par la similarité entre ces observations. Les noeuds sont ensuite projetés dans un espace de faible dimension pouvant être facilement séparé pour former des groupes. Cette méthode montre généralement des résultats meilleurs que k-moyennes (Chang *et al.*, 2010).

Ainsi, étant donné un ensemble de points $X = \{x_i\ ,\ i = 1, ..., n\}$. On mesure leur similarité $s_{ij} = s(x_i, x_j)$ avec une fonction de similarité (dans notre cas la distance euclidienne) et on construit la matrice de similarité $S = (s_{ij})_{i,j=1,..,n}$ (Von Luxburg, 2007).

La méthode de classification spectrale procède en plusieurs étapes :

1. À partir de S est obtenu un graphe de similarités.
2. À l'aide du graphe, une représentation vectorielle des données est obtenue.
3. Un regroupement k-moyennes est enfin effectué pour obtenir les k groupes disjoints.

DBScan. DBScan (*density-based spatial clustering of applications with noise*) (Ester *et al.*, 1996) est un algorithme basé sur la notion de densité d'éléments. Il regroupe les points de données « densément groupés » dans un même groupe. Il peut identifier des groupes dans de grands ensembles de données spatiales en examinant la densité locale des points de données.

Cet algorithme ne nécessite pas de configurer un nombre de groupes en entrée et peut identifier les valeurs aberrantes en les éliminant du processus de partitionnement. En effet, DBScan n'utilise que deux paramètres : epsilon ϵ et minPoints. ϵ est le rayon du cercle à créer autour de chaque point de données pour vérifier la

densité et minPoints est le nombre minimum de points de données requis à l'intérieur de ce cercle pour que ce point de données soit classé comme point central. Si la valeur de ϵ choisie est trop petite, une grande partie des données ne sera pas regroupée. En revanche, si la valeur choisie est trop élevée, les groupes fusionneront et la majorité des observations se trouveront dans le même groupe. ϵ doit être choisi en fonction des distances observées entre les données. En général, de petites valeurs de ϵ sont préférables. La valeur minimale des minPoints doit être 3, mais plus l'ensemble de données est grand, plus la valeur minPoints à choisir est grande. DBScan est capable de trouver assez bien des groupes de tailles et de formes arbitraires.

K-modes. K-modes (Huang, 1998) est un algorithme conçu pour le regroupement des données catégoriques. Son fonctionnement est similaire à k-moyennes mais utilise les valeurs ayant la plus grande fréquence dans une distribution pour former des groupes. Au lieu de calculer des distances, il calcule les dissimilarités entre les observations et utilise une méthode basée sur la fréquence pour mettre à jour les modes.

Ainsi, selon Lakshmi et al (Lakshmi *et al.*, 2017), les principales caractéristiques de k-modes sont (1) simple et facile à mettre en œuvre et (2) gère efficacement les grands ensembles de données. Cependant, les principaux inconvénients sont (1) la nécessité de définir le nombre de groupes à l'avance, (2) ne fonctionne que sur les données catégoriques et (3) produit des solutions optimales locales.

K-prototypes. L'une des limites de k-moyennes est qu'il ne convient pas aux variables catégoriques, alors que k-modes ne convient que pour les données catégoriques. Face à ce problème, Huang (Huang, 1998) a proposé un algorithme appelé K-prototypes pour le regroupement d'attributs mixtes (numériques et catégoriques). Son algorithme est une forme d'amélioration combinant les algorithmes

de regroupement K-moyennes et K-modes.

À la différence des méthodes évoquées précédemment, k-prototypes introduit un poids γ, utilisé pour éviter de favoriser un type d'attribut par rapport à un autre. Huang (Huang, 1997) suggère de prendre en compte l'écart-type moyen des attributs numériques pour spécifier γ. À noter que la connaissance du domaine et des données est un point essentiel dans cette configuration. Ainsi, si les attributs numériques sont plus importants pour le regroupement, un petit γ devrait être privilégié, tandis que si les données catégoriques sont plus importantes, un γ plus grand est requis.

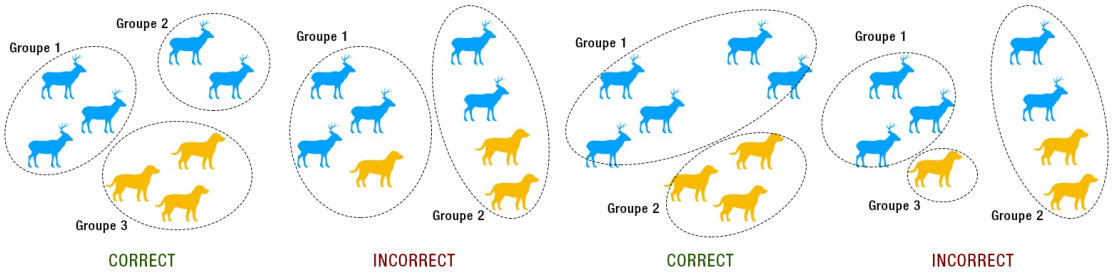
3.2.3 Évaluation

Afin de valider la similarité entre le regroupement effectué par l'équipe des biologistes et notre travail, nous avons opté pour la « V-Mesure » (Rosenberg et Hirschberg, 2007) qui est une métrique basée sur l'entropie, utilisée généralement pour comparer la concordance des résultats de classification entre deux techniques de regroupement sur le même jeu de données. C'est la moyenne harmonique entre l'homogénéité et la complétude :

$$\text{V-mesure} = 2 \times \frac{\text{homogénéité} \times \text{complétude}}{\text{homogénéité} + \text{complétude}} \tag{3.5}$$

L'homogénéité mesure la façon dont l'algorithme de regroupement a assigné au même groupe les éléments de la même classe. Ainsi, un groupe est dit pur, avec un score d'homogénéité égal à 1, si tous ses membres appartiennent à la même catégorie. La complétude vérifie si l'algorithme de regroupement a affecté tous les éléments appartenant à une même classe réelle au même groupe prédit.

Selon Rosenberg et Hirschberg (Rosenberg et Hirschberg, 2007), la V-mesure présente plusieurs avantages. En effet, elle permet d'évaluer une solution de regroupe-



(a) Illustration de l'homogénéité (b) Illustration de la complétude

Figure 3.3: Illustration de l'homogénéité et de la complétude (Clu, 2018)

ment indépendamment de l'algorithme utilisé, de la taille de l'ensemble de données ou encore du nombre de classes et de groupes. Aussi, en évaluant deux critères (homogénéité et complétude), la V-mesure offre un score plus complet en comparaison à d'autres techniques d'évaluation de regroupement se basant sur un seul critère telles que la pureté et l'entropie (Zhao et Karypis, 2001), qui ne mesurent pas la complétude. Enfin, la V-mesure peut être analysée qualitativement en terme d'homogénéité et de complétude, constituant ainsi un outil de diagnostic pour déterminer les erreurs commises lors du regroupement.

Dans cette partie, nous avons pu introduire la méthodologie mise en œuvre dans le but d'effectuer le regroupement d'espèces d'arbres selon leur traits fonctionnels. Dans la partie suivante, nous présenterons les méthodes utilisées pour la prédiction de la croissance du diamètre à hauteur de poitrine des arbres urbains.

3.3 Prédiction des diamètres à hauteur de poitrine

Des modèles fiables de croissance des arbres urbains au fil du temps sont utiles pour (1) sélectionner les espèces appropriées pour les sites de plantation disponibles, (2) anticiper les futurs coûts d'entretien et de retrait des arbres et (3) quantifier les avantages procurés par les arbres (Berland, 2020). Le domaine de la foresterie

urbaine montre un fort intérêt pour la compréhension de la croissance des arbres dans le temps pour plusieurs raisons. Premièrement, une estimation précise de la taille des jeunes arbres plantés permet aux gestionnaires des forêts urbaines de planifier les futurs coûts associés à leur élagage et à leur entretien (McPherson *et al.*, 2016). Deuxièmement, du point de vue de la gestion et de la maintenance des arbres, il est important de sélectionner des arbres adaptés aux contraintes d'espace d'un site de plantation et d'éviter que ce dernier n'endommage les infrastructures environnantes (Vogt *et al.*, 2015). Finalement, le diamètre à hauteur de poitrine est utile pour le calcul de nombreux services écosystémiques. Ainsi, la prédiction du DHP permet de mettre en évidence comment un arbre existant ou une nouvelle plantation devrait se comporter (en termes de services écosystémiques) à travers le temps en fournissant des estimations des avantages pour l'année en cours et dans le futur. Ces estimations permettent ainsi de motiver le choix d'une espèce comparée à une autre ou encore de justifier le budget requis pour la gestion d'une forêt urbaine auprès des décideurs (McPherson et Peper, 2012).

Dans cette logique, plusieurs études des relations de croissance proposent des modèles allométriques permettant entre autre de prédire la hauteur des arbres en fonction du DHP ou encore de prédire la masse foliaire en fonction de la dimension des couronnes d'arbres en milieu urbain à travers plusieurs villes du monde (Monteiro *et al.*, 2016; Troxel *et al.*, 2013; Song *et al.*, 2020). La croissance est en effet différente d'une ville à une autre en raison de divers paramètres tels que le climat principalement, les stress induits par l'activité anthropique ou encore la qualité de la maintenance. McPherson et Peper (McPherson et Peper, 2012) ont démontré que les différences dans l'environnement biophysique et la gestion des arbres peuvent entraîner des différences dans la croissance de ces derniers, même pour les mêmes espèces qui poussent dans deux villes de la même région, ce qui

peut conduire à une mauvaise évaluation des avantages fournis par les arbres. Divers modèles statistiques ont été utilisés pour l'estimation des taux de croissance des arbres, notamment la régression linéaire (Semenzato *et al.*, 2011), polynomiale et logarithmique (Troxel *et al.*, 2013). L'étude de McPherson, van Doorn et Peper (McPherson *et al.*, 2016) offre une collection complète de modèles allométriques en termes d'espèces et de zones géographiques couvertes. Les auteurs ont ainsi proposé des modèles de prédiction de croissance pour les espèces les plus communes pour 16 zones climatiques à travers les États-Unis.

L'objectif principal de notre étude est de développer des modèles de prédiction des DHP pour les espèces d'arbres urbains communes à Montréal, où de tels modèles n'ont pas été créés auparavant. Pour un maximum d'espèces de notre ensemble de données, l'objectif est d'ajuster un modèle de croissance permettant de prédire le DHP à l'année $n + x$ à partir du DHP à l'année n. De plus, à notre connaissance, aucune étude n'a mis en œuvre des modèles basés sur l'intelligence artificielle pour la prédiction des DHP. Nous émettons l'hypothèse que ce type d'approche basé sur des algorithmes d'apprentissage automatique permettrait d'améliorer la précision de la prédiction par rapport à l'utilisation de modèles de régression conventionnels en capturant des relations complexes et non linéaires dans les données (Chen et Asch, 2017). Le modèle ainsi développé sera intégré à SylvCiT et permettra de créer des courbes de croissance pour chaque espèce, de calculer les services écosystémiques en fonction de la croissance des arbres et de prédire les arbres arrivant à maturité dans les années à venir.

3.3.1 Pré-traitements des données

Dans le cadre de nos expérimentations, nous avons utilisé l'ensemble de données des arbres publics de Montréal (décrit en 3.1.1) ainsi que l'historique des dia-

mètres à hauteur de poitrine (décrit en 3.1.2). Pour rappel le répertoire des arbres publics contient 22 champs d'information pour chacun des arbres recensés (voir tableau 3.1). Cependant, la revue de la littérature relative à la prédiction de croissance des arbres nous a permis de sélectionner les variables les plus pertinentes pour réaliser nos expérimentations. Ainsi, parmi les 22 variables, nous en avons conservé huit dans un premier temps. Ces variables sont les suivantes : le numéro d'emplacement, le côté de la rue, le sigle de l'arbre, le nom de l'essence en latin, la mesure du DHP, la date de plantation, la date de la dernière mesure du DHP, la latitude et longitude.

Le principal défi avec cet ensemble de données est la présence importante de données manquantes notamment concernant la date de plantation ainsi que la mesure du DHP des arbres essentiels à notre étude. De plus, les données contiennent beaucoup d'erreurs de transcriptions et de valeurs aberrantes. En effet, sur les 318 309 arbres recensés, 671 observations ne contiennent pas de mesure de DHP ni de date de relevé, et plus de la moitié des observations (167 235) ne contiennent pas de date de plantation. Dans un but d'analyse de la composition de la forêt urbaine montréalaise, nous avons combiné l'ensemble des arbres publics avec une base de données de correspondance des espèces avec leur genre et leur famille. Ainsi, l'ensemble de données brut est constitué de 588 espèces appartenant à 57 genres et 28 familles. Le tableau 3.3 offre un aperçu des 10 espèces les plus présentes dans l'ensemble de données avec les DHP minimum et maximum pour chaque espèce.

Afin de compléter notre ensemble de données, qui fourni certaines informations nécessaires à notre étude, telles que les coordonnées géographiques des arbres et le côté de la rue, nous l'avons combiné avec les données historiques des DHP. Pour ce faire, nous avons combiné chaque entrée en fonction du triplet : côté de la rue, numéro d'emplacement et nom de l'essence. Le jeu de données généré est constitué de 634 957 observations, avec plusieurs observations pour un même arbre.

Nom Latin	**Nombre**	**%**	**Min DHP**	**Max DHP**	**Écart type**
Acer saccharium	33179	10,47	0	323	24,06
Acer platanoides	31539	9,95	1	591	18,83
Fraxinus pennsylvanica	23122	7,29	0	127	14,38
Tilia cordata	15795	4,98	0	121.3	17,48
Celtis occidentalis	9454	2,98	0	170	15,57
Gleditsia triacanthos 'Skyline'	9378	2,95	1	70	10,44
Gleditsia triacanthos var. inermis	7643	2,41	1	89	16,07
Gymnocladus dioicus	6789	2,14	2	73	7,59
Picea pungens	6761	2,13	0	150	10,56
Ulmus pumila	5974	1,88	0	120	20,36

Tableau 3.3: Caractéristiques des 10 espèces d'arbres les plus présentes à Montréal

Le tableau 3.3 montre certaines valeurs aberrantes relatives aux mesures de DHP, tel qu'un DHP maximum de 591 pour l'Acer platanoide. Une valeur aberrante dans une distribution est une valeur s'écartant fortement des autres observations et qui serait anormalement faible ou élevée, contrastant grandement avec les valeurs « normalement » mesurées. Une valeur aberrante peut être due à la variabilité relative au phénomène observé ou bien à des erreurs de mesures. L'utilisation d'une série de données avec des valeurs aberrantes peut conduire à des analyses et résultats biaisés. Il convient donc d'utiliser des méthodes robustes afin de parer à ce problème. Dans le but d'identifier les valeurs aberrantes dans nos données, nous nous sommes basés sur la mesure de l'écart interquartile (en anglais, *interquartile range* ou IQR). Cette mesure, considérée comme un estimateur robuste en statistique, permet de mesurer la dispersion d'une distribution en faisant la différence entre le troisième (Q3) et le premier quartile (Q1).

$$IQR = Q_3 - Q_1$$

Q3, aussi appelé quartile supérieur, représente la valeur pour laquelle 75% des observations lui sont inférieures, tandis que le quartile inférieur Q1 est la valeur pour laquelle 75% des observations lui sont supérieures.

Une fois l'IQR obtenu, nous calculons la borne supérieure, pour laquelle toute valeur supérieure est considérée comme aberrante, selon l'équation suivante :

$$\text{borne}_{sup} = Q3 + k \times IQR$$

La borne inférieure, pour laquelle toute valeur inférieure est considérée comme aberrante, est calculée selon l'équation suivante :

$$\text{borne}_{inf} = Q1 - k \times IQR$$

k étant une constante positive que nous avons fixé à 3 après l'analyse qualitative effectuée avec l'équipe de foresterie urbaine collaborant au projet. La borne inférieure a été estimée à 50mm car c'est le diamètre à hauteur de poitrine minimal pour la plantation des arbres par le service de plantation de la ville de Montréal (d'après les données obtenues auprès de la pépinière municipale de la ville de Montréal).

Les dates de plantation ainsi que les dates de relevés contenaient aussi des valeurs aberrantes. Suite à l'analyse du jeu de données, nous avons défini la borne inférieure à l'année 1960 car elle présente plusieurs enregistrements plausibles et 2020 pour la borne supérieure. Nous avons aussi procédé à la suppression des valeurs manquantes. Puis, nous avons ajouté une variable à notre jeu de données : le nombre de jours depuis la date de plantation, calculé en fonction de la date de plantation et de la date de mesure du DHP, dans l'objectif d'aider les algorithmes entraînés à capturer la tendance de croissance.

Une autre variable présente dans l'ensemble de données représente les coordonnées géospatiales des arbres. Ainsi, l'une des approches pour gérer ces dernières

consiste à les traiter les unes par rapport aux autres. Généralement, ces données sont regroupées à l'aide de techniques telles que K-moyennes et DBScan afin d'être converties en variables catégoriques. Le choix du nombre de groupes se fait généralement de manière empirique en fonction du problème traité. Dans notre cas, il est souhaitable de prévoir de petites zones à l'échelle d'un quartier ou d'un arrondissement afin de mettre en évidence les possibles différences de croissances entre les arbres. Dans cet objectif, nous utilisons la bibliothèque python *pygeohash* (PyGeohash, 2021). Geohash est un système de géocodage basé sur une fonction de hachage permettant de subdiviser la surface terrestre selon une grille hiérarchique composée de cellules rectangulaires. Ainsi, Geohash offre un moyen pratique d'exprimer un emplacement (coordonnées géospatiales) à l'aide d'une courte chaine alphanumérique de taille variable selon le niveau de précision souhaité.

Le tableau 3.4 offre un aperçu des tailles de cellules en fonction de la précision choisie. Notre étude portant sur les arbres de Montréal, la superficie moyenne des arrondissements étant d'environ 20 km^2 nous avons choisi une précision de 5 caractères correspondant à des cellules de $4,89\ km$ de coté.

Aussi, nous avons supprimé :

— Les espèces d'arbres dont les écarts-types du DHP étaient à 0 indiquant que toutes les valeurs de données sont exactement les mêmes,

— Les espèces d'arbres présentant moins de cinquante observations car les spécialistes ont jugé ce nombre insuffisant pour être pris en compte. Ainsi, considérer ces espèces pourrait biaiser nos modèles.

Nous avons utilisé l'encodage à chaud pour les attributs catégoriques (décrit dans la partie 3.2.2). Nous avons choisi cet encodage, car il n'existe pas de relation d'ordre ou de différences particulières entre les observations de l'attribut. Ainsi, cet encodage a été appliqué notamment pour le type d'arbre inventorié et le sigle identifiant l'espèce d'arbre.

Précision	**Largeur de cellule**	**Hauteur de cellule**
1	5000 km	5000 km
2	1250 km	625 km
3	156 km	156 km
4	39,1 km	19,5 km
5	4,89 km	4,89 km
6	1,22 km	0,61 km
7	153 m	153 m
8	38,2 m	19,1 m
9	4,77 m	4,77 m
10	1,19 m	0,596 m
11	149 mm	149 mm
12	37,2 mm	18,6 mm

Tableau 3.4: Table de précision de Geohash. Source : (Xu *et al.*, 2017)

Étant donné que les données sont exprimées en plusieurs unités de mesure différentes, nous avons normalisé les traits numériques à l'aide de la fonction de *Min-Max* dans un intervalle fixe [0,1] comme décrit dans la partie 3.2.2.

3.3.2 Algorithmes de prédiction utilisés

Régression linéaire. La régression linéaire (Freedman *et al.*, 1981) est l'une des techniques fondamentales de statistique et d'apprentissage automatique. Une régression permet de déterminer si et comment un phénomène influence l'autre ou comment plusieurs variables sont liées. On considère généralement un phénomène d'intérêt et un certain nombre d'observations. Chaque observation a deux ou plusieurs caractéristiques. En partant de l'hypothèse que (au moins) l'une des

caractéristiques dépend des autres, la régression linéaire permet d'établir une relation entre elles. Ainsi, un modèle de régression linéaire multiple d'une variable dépendante y sur l'ensemble des variables indépendantes $x = (x_1, ..., x_n)$ est de la forme suivante :

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{3.6}$$

où :

— $\beta_0, \beta_1, ..., \beta_n$ sont les paramètres à estimer (coefficient de régression)

— ϵ est le terme d'erreur aléatoire du modèle

La valeur estimée ou prédite, $\hat{y}_i$ pour chaque observation $i = 1, ..., n$ doit être aussi proche que possible de la valeur réelle correspondante. La différence $y_i - \hat{y}_i$ est appelée résidu. La régression consiste donc à estimer les meilleurs poids permettant de minimiser les résidus.

Pour estimer β_0 et β_1, on peut utiliser la méthode des moindres carrés minimisant la somme des carrés des distances entre y_i et $f(x_i)$, f étant la fonction qui décrit « le mieux » les données, selon l'équation suivante :

$$\Delta = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{3.7}$$

On parle ainsi de régression pour exprimer la diminution de la somme des écarts.

Forêt d'arbres décisionnels. Une forêt d'arbres décisionnels (*Random Forest* ou *RF*)(Breiman, 2001) est un algorithme d'apprentissage par ensembles constitué de plusieurs arbres de décisions. Ces derniers sont des modèles hiérarchiques dont la classification se base sur une succession de règles de décision simples inférées des données d'entraînement et des différents attributs (Breiman *et al.*, 1984).

Ainsi, les RF sont des méthodes d'apprentissage d'ensembles pour la classification, la régression et d'autres tâches. Chaque arbre de la forêt aléatoire est entraîné sur un sous-ensemble aléatoire de données selon le principe du *bagging* (Brei-

man, 1996), avec un sous ensemble aléatoire des attributs. Le bagging consiste à sous-échantillonner (avec remise) l'ensemble d'entraînement et de faire générer à l'algorithme voulu un modèle pour chaque sous-échantillon. La prédiction finale du modèle est donnée via l'agrégation par la moyenne pour une régression ou par vote majoritaire pour une tâche de classification.

Par rapport à un arbre de décision, les forêts aléatoires sont robustes aux valeurs aberrantes, ont une plus grande précision et gèrent bien les grands ensembles de données (Ho, 1995). Bien que les forêts aléatoires se soient avérées être l'un des algorithmes d'apprentissage automatique montrant des résultats satisfaisants, leurs applications dans le domaine de la foresterie urbaine sont limitées.

Par ailleurs, les hyperparamètres des RF sont soit utilisés pour augmenter la puissance prédictive du modèle, soit pour rendre le modèle plus rapide. Dans le cadre de notre étude, nous expérimentons plusieurs modèles de RF en tentant d'optimiser les paramètres suivants :

- Le nombre d'arbres que l'algorithme construit avant de prendre le vote maximum ou de prendre les moyennes des prédictions. En général, un plus grand nombre d'arbres augmente les performances et rend les prédictions plus stables mais ralentit également le calcul,
- Le nombre maximum d'attributs que le RF considère pour diviser un noeud,
- Le nombre minimum de feuilles requises pour fractionner un noeud interne.

Réseaux de neurones. Les réseaux de neurones artificiels sont des modèles d'apprentissage automatique. L'unité de base d'un réseau neuronal est le neurone. Ce dernier est caractérisé par ses poids, un biais et sa fonction d'activation visant à lui conférer un comportement non linéaire (Goodfellow *et al.*, 2016a). Le réseau de neurones artificiel le plus simple est le perceptron (Rosenblatt, 1958). La figure 3.4 permet d'illustrer par l'exemple le fonctionnement d'un perceptron.

Ainsi, on dispose d'une série d'entrées X (dans notre exemple : 3, 10 et 1), chacune associée à un poids w (dans notre exemple $0,2$, $0,01$ et $0,4$). Le neurone va alors calculer la somme S des entrées, pondérée par leur poids respectif avec l'addition d'un biais b (dans l'exemple 0.1) selon l'équation :

$$S = \sum_{i=1}^{n} w_i x_i + b$$

Les poids w vont contrôler le signal. Autrement dit, le poids va pondérer l'influence d'une entrée sur la sortie. Alors que le biais b est une constante qui va garantir que même lorsque toutes les entrées sont des zéros, il y aura toujours une activation dans le neurone. Le résultat de cette somme est ensuite passé à travers une fonction de seuil qui renvoie 0 ou 1 en fonction de la valeur de S et fournit la sortie y.

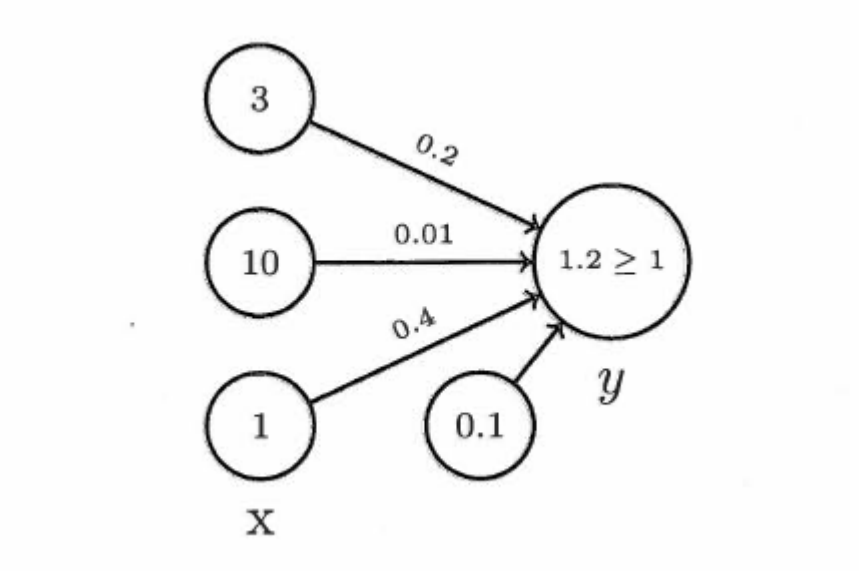


Figure 3.4: Exemple d'un perceptron.

L'entraînement du perceptron consiste à déterminer de manière itérative les valeurs optimales des poids en fonction des données d'entraînement afin de généraliser le modèle. Pour chaque observation, la classe y' prédite est comparée à la classe réelle y, si $y' \neq y$ alors les poids sont recalculés. Ainsi, quand $y' = y$ pour

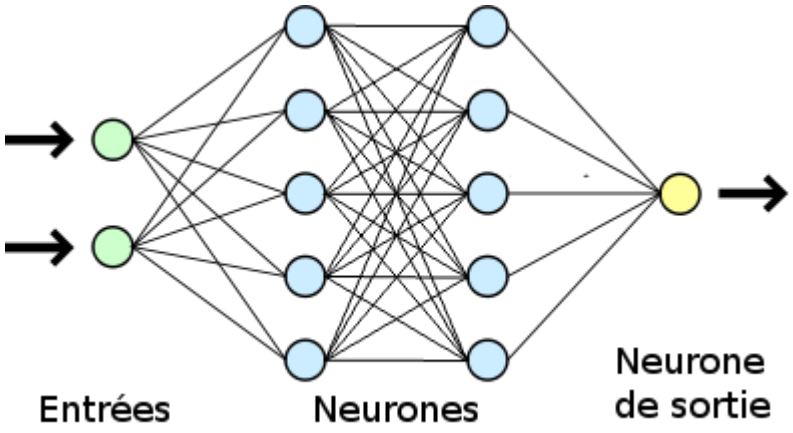


Figure 3.5: Un exemple de réseau de neurones

toutes les observations disponibles dans les données d'entraînement, les poids w sont alors optimaux et ne subissent plus de modifications à chaque itération, on dit alors que l'algorithme a convergé.

Un réseau de neurones est composé de plusieurs perceptrons, donc de plusieurs neurones et de plusieurs couches de neurones. Ces modèles à plusieurs couches sont appelés, entre autres, perceptrons multicouches ou réseau de neurones à propagation avant. À la différence du perceptron simple, le perceptron multicouche (figure 3.5) dispose d'une ou plusieurs couches dites « cachées » entre la couche en entrée et la couche en sortie. Pour la couche d'entrée, le nombre de neurones est déterminé par le nombre de variables d'entrée et chaque neurone représente une variable d'entrée.

La couche d'entrée ne traite pas le signal mais transmet seulement le signal de manière non linéaire à la couche cachée. Les neurones composant une couche cachée calculent simplement la somme pondérée des entrées et des poids, ajoutent le biais et exécutent une fonction d'activation. La couche de sortie est responsable de la production du résultat final. Le nombre de neurones d'une couche de sortie est défini selon le problème traité. Pour ces modèles, les poids sont recalculés

à l'aide d'une technique de rétro-propagation de l'erreur, basée sur la règle de dérivation en chaîne, à chaque itération lors de l'apprentissage (Rumelhart *et al.*, 1986). Cette méthode calcule le gradient de la fonction de perte par rapport aux poids du réseau selon les données. Les poids sont alors ajustés dans le sens opposé au gradient. La fonction de perte quantifie l'écart entre les prévisions du modèle et les observations réelles du jeu de donnée utilisé pendant l'apprentissage. En répétant ce processus, on obtient des poids optimaux qui minimisent la fonction de perte et permettent au réseau de neurones de faire de meilleures prédictions.

Ces modèles multicouches sont des modèles d'apprentissage profond capables d'apprendre des représentations de données complexes (Goodfellow *et al.*, 2016a; LeCun *et al.*, 2015).

Pour l'architecture de notre réseau, nous nous sommes inspirés des travaux de Vinicius et al (Vinícius Oliveira Castro *et al.*, 2013) qui visaient à modéliser la croissance et le rendement des peuplements d'eucalyptus situés dans le nord du Brésil, au niveau de l'arbre individuel, en utilisant des réseaux de neurones artificiels. Suite à leur expérimentation, un réseau de neurones à une seule couche cachée a montré les meilleurs résultats. C'est donc l'architecture que nous avons prise pour point départ. Le choix de la fonction d'activation n'étant pas mentionné dans leur étude, nous en avons testé plusieurs, notamment la tangente hyperbolique ($tanh$), la sigmoïde et l'unité de rectification linéaire ($ReLU$) (Nair et Hinton, 2010; Sharma, 2017)

Nous faisons aussi l'hypothèse que plus de couches cachées pourraient capturer plus facilement la complexité du phénomène étudié. De ce fait, nous avons testé un réseau à 1, 2, 3 et 4 couches cachées en faisant varier la fonction d'activation pour chaque réseau.

Un hyperparamètre important à fixer pour tout modèle basé sur l'apprentissage

profond est le taux d'apprentissage (*learning rate* : λ). Ce dernier est un scalaire strictement positif qui permet de contrôler la vitesse d'apprentissage du modèle. Il permet de contrôler la mise à jour des poids du modèle lors de la rétropropagation de l'erreur.

Le taux d'apprentissage λ peut être choisi par essais et erreurs, mais il est généralement préférable de le choisir en surveillant les courbes d'apprentissage qui tracent la fonction objectif en fonction du temps. La principale question est de savoir comment définir λ. Si λ est trop grand, la courbe d'apprentissage montrera de fortes oscillations. Si le taux d'apprentissage est trop faible, l'apprentissage se déroulera lentement, ou s'arrêtera complètement.

En règle générale, le taux d'apprentissage initial optimal, en termes de temps total d'apprentissage et de valeur du coût final, est supérieur au taux d'apprentissage qui donne les meilleures performances après les 100 premières itérations environ (Goodfellow *et al.*, 2016b). Par conséquent, il est généralement préférable de surveiller les (100) premières itérations et d'utiliser un taux d'apprentissage supérieur au taux d'apprentissage le plus performant, sans qu'il soit trop élevé au risque de provoquer une grande instabilité (Goodfellow *et al.*, 2016b). Nous avons fixé un taux d'apprentissage à 0,001.

L'idée centrale de l'apprentissage automatique est de produire des modèles qui sauront faire des prédictions correctes non seulement sur les données d'entraînement, mais également sur de nouvelles données. Cependant, les réseaux de neurones comme les autres méthodes d'estimation non linéaires peuvent souffrir d'un sous-apprentissage (*underfitting*) ou d'un sur-apprentissage (*overfitting*). La recherche d'un coût minimal peut conduire à calculer un modèle qui s'ajuste trop bien aux données (qui apprend par coeur l'ensemble de données) mais répond mal au problème que l'on souhaite résoudre (manque de capacité de généralisation).

Le sous-apprentissage fait référence à un modèle qui ne peut ni modéliser les données d'entraînement ni généraliser à de nouvelles données. Le modèle n'a « pas suffisamment appris » des données, entraînant une faible généralisation et des prédictions peu fiables. Cela se produit par exemple lorsque l'on tente de représenter un phénomène non-linéaire par un modèle linéaire.

L'apprentissage profond fait appel à des modèles complexes tout en les contraignant par différentes techniques afin de mitiger le sur-apprentissage. On nomme *technique de régularisation* toute modification à un algorithme d'apprentissage visant à en réduire son erreur de généralisation, mais non son erreur d'apprentissage (Goodfellow *et al.*, 2016b).

Lors de l'apprentissage d'un modèle, la régularisation permet d'imposer une contrainte pour favoriser certaines solutions plus souhaitables. Il existe de nombreuses formes de régularisation, qui dépendent de l'objectif recherché et des hypothèses fixées sur le problème.

Pour éviter le surapprentissage de notre modèle, nous utilisons une couche de décrochage (*dropout*) afin d'ignorer certaines unités (neurones) durant la phase d'entraînement d'un ensemble de neurones choisis aléatoirement. Nous avons fixé le taux de décrochage à 0,5, car étant la valeur donnant généralement les meilleurs résultats selon la littérature (Srivastava *et al.*, 2014). Ainsi, dans notre cas 50% des neurones seront supprimées à chaque phase d'entraînement. De ce fait, chaque itération a un ensemble différent d'unités et donc un résultat en sortie différent. Le réseau devient moins sensible aux poids spécifiques des neurones en forçant l'apprentissage d'une représentation distribuée dans l'ensemble du réseau. Il en résulte un réseau qui est capable d'une meilleure généralisation et est moins susceptible de souffrir de sur-apprentissage.

Nous entraînons nos modèles pendant un maximum de 1 000 époques, c'est-à-dire

que l'algorithme verra l'ensemble de données dans son intégralité 1 000 fois. Un problème inhérent à l'entraînement des réseaux de neurones réside dans le choix du nombre d'époques d'entraînement à utiliser. Trop d'époques peuvent conduire à un sur-apprentissage de l'ensemble de données d'entraînement, tandis que trop peu peuvent entraîner un sous-apprentissage. Pour éviter cela, nous utilisons l'arrêt prématuré (*early stopping*) qui est une technique de régularisation qui consiste à stopper l'entraînement dès que les performances du modèle cessent de s'améliorer sur un ensemble de données de validation.

En pratique, la descente de gradient dont nous avons parlé précédemment peut rencontrer certains problèmes pendant l'entraînement qui ralentissent le processus d'apprentissage ou, dans le pire des cas, empêchent l'algorithme de trouver une solution convenable.

Il est crucial que notre modèle d'apprentissage profond s'entraîne dans un temps acceptable sans pénaliser la précision du modèle. Dans cette optique, plusieurs algorithmes ou stratégies d'optimisation sont chargés d'accélérer la convergence et de fournir la solution la plus optimale : SGD (Pedregosa, 2018), Momentum de Nesterov (Sutskever *et al.*, 2013), AdaGrad (Duchi *et al.*, 2011), RMSProp (Tieleman et Hinton, 2012) et Adam (Kingma et Ba, 2014). Dans le cadre de nos expériences, nous avons opté pour l'utilisation d'Adam, qui est la méthode d'optimisation dominante dans la recherche pour sa vitesse de convergence et sa robustesse.

Pour évaluer les performances de nos modèles et leur capacité de généralisation sur l'intégralité des données disponibles, nous utilisons une validation croisée qui consiste à diviser l'ensemble des données en K parties (*folds*) de tailles égales. Tous les échantillons sont utilisés pour l'entraînement du modèle, sauf un destiné à tester les performances de l'algorithme. Cette étape est répétée k fois. Les

performances du modèle finales sont la moyenne des différents scores obtenus à chaque itération (Berrar, 2019). Cependant, il arrive que certaines espèces soient sous représentées ou absentes dans l'un des sous-échantillons de tests. Afin de parer à ce problème, nous utilisons une validation croisée stratifiée. La stratification permet de s'assurer que la répartition des espèces est la même dans tous les sous ensembles d'apprentissage et de validation utilisés.

3.3.3 Évaluation

L'objectif de nos travaux est de prédire le DHP d'une espèce d'arbre pour une année n. Dans le but d'évaluer la performance de nos modèles, nous nous sommes principalement concentrés sur l'évaluation de l'erreur absolue moyenne (*Mean Absolute Error* ou *MAE*), l'erreur quadratique moyenne (*MSE* pour *Mean Square Error*) et la racine carrée de l'erreur quadratique moyenne (*Root Mean Square Error* ou *RMSE*) qui sont des métriques standards dans la littérature pour l'évaluation de régressions. Ces métriques permettent de mesurer la distance entre une valeur prédite $\hat{y}$ et la valeur réelle y (vérité terrain). Ainsi, la MSE permet de mesurer la moyenne des carrés des erreurs. Autrement dit, la différence quadratique moyenne entre les valeurs prédites et les valeurs réelles. Cette mesure pénalise les erreurs. En effet, en raison de la mise au carré, un poids plus important est attribué aux erreurs. Pour une série de n observations y_i , $i = 1, ..., n$, la MSE est définie par :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

La MAE mesure la différence moyenne absolue entre les valeurs prédites et réelles, définie tel que :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

La $RMSE$ est la racine carrée de la MSE et est définie comme suit :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

La MSE et la $RMSE$ pénalisent fortement les erreurs commises par le modèle évalué. Elles sont particulièrement utiles pour détecter si des valeurs aberrantes perturbent les prédictions. En outre, la MAE et la $RMSE$ expriment l'erreur de prédiction du modèle dans la même unité que la variable prédite. Ainsi, plus la valeur est basse, meilleure est la performance du modèle.

CHAPITRE IV

SYLVCIT

Nous allons présenter dans ce chapitre les travaux et la méthodologie mise en œuvre pour l'implémentation de SylvCiT. Nous allons donc commencer par décrire les calcul des services écosystémiques et les indices écologiques à la partie 4.1. Ensuite, nous traiterons de la méthode de calcul des améliorations à la partie 4.2. Puis, nous aborderons la démarche suivie pour la recommandation d'essence d'arbres à la partie 4.3. Enfin, nous conclurons ce chapitre avec la partie 4.4 détaillant l'implémentation et les technologies mises en œuvre pour l'élaboration de notre logiciel SylvCiT.

4.1 Services écosystémiques et indices biologiques

4.1.1 Stockage de carbone et valeur monétaire

La biomasse est toute matière organique dérivée directement ou indirectement du processus de photosynthèse par les plantes et les algues, à l'exception des matières fossilisées (da Silva *et al.*, 2018). La biomasse offre plusieurs possibilité d'usage mais actuellement l'accent est mis sur son utilisation comme combustible dans les secteurs de la production d'énergie (Arteaga-Pérez *et al.*, 2015). L'estimation de la densité de biomasse est ainsi requise par exemple pour déterminer les bilans nationaux canadiens de carbone forestier (Lambert *et al.*, 2005).

Selon Ketterings *et al.* (Ketterings *et al.*, 2001), la méthode idéale pour mesurer la biomasse consiste à échantillonner de manière destructive et à peser physiquement des arbres entiers mais cette méthode prend du temps et est peu pratique pour acquérir des données sur un certain nombre d'espèces en milieu urbain.

Au début des années 1980, le service Canadien des forêts dans le cadre du programme de recherche appelé ÉNergie de la FORêt (ENFOR) a recueilli des données essentielles à l'estimation de la biomasse. Ainsi, Lambert *et al.* (Lambert *et al.*, 2005) ont développé un système d'équations permettant de séparer la biomasse en compartiments (feuillage, bois, branches, écorce) à partir de mesures de DHP uniquement, ou à partir de mesures de DHP et de hauteur. Ces formules permettent d'estimer la biomasse aérienne pour 33 essences forestières, par groupe de feuillus et de conifères. Ainsi, dans le cadre de notre étude et avec l'expertise de l'équipe de foresterie urbaine collaborant au projet, nous avons choisi de baser nos calculs sur ces équations :

$$\begin{aligned}
y_{wood} &= \beta_{wood1} D^{\beta_{wood2}} + e_{wood} \\
y_{bark} &= \beta_{bark1} D^{\beta_{bark2}} + e_{bark} \\
y_{foliage} &= \beta_{foliage1} D^{\beta_{foliage2}} + e_{foliage} \\
y_{branches} &= \beta_{branches1} D^{\beta_{branches2}} + e_{branches} \\
y_{total} &= \widehat{y}_{wood} + \widehat{y}_{bark} + \widehat{y}_{foliage} + \widehat{y}_{branches} + e_{total}
\end{aligned} \tag{4.1}$$

Où y_i est le compartiment de biomasse sèche i d'un arbre vivant, β_{jk} sont des paramètres de modèle avec des estimations de coefficient où j est le bois, l'écorce, le feuillage, la couronne ou les branches et $k \in \{1, 2\}$, D est le DHP de l'arbre (cm) et $\widehat{y}_i$ est la prédiction de y_i, e est le terme d'erreur.

Les valeurs de β_{jk} sont tirées directement de l'article susmentionné. Il est à no-

ter que certaines espèces sont absentes de l'étude. Pour ces essences, les valeurs utilisées correspondront aux valeurs incluses à la toute fin du tableau 3 fourni dans l'étude de (Lambert *et al.*, 2005), sous « conifère » et « feuillus » ou sous « toutes » si le type d'arbre n'est pas défini.

Pour les besoins de nos travaux, nous avons choisi d'ignorer les termes d'erreur (agissant comme si chaque terme d'erreur était égal à zéro). Nous fournissons simplement des valeurs estimées, sans spécifier la marge d'erreur pour chaque valeur unique. Cela permet de simplifier les calculs et de les rendre plus rapides.

Une fois la biomasse aérienne estimée, nous devons calculer la biomasse racinaire. Pour cela, nous nous appuyons sur les travaux de (Li *et al.*, 2003) qui suppose que la biomasse racinaire totale (variable dépendante) peut être estimée à partir de la biomasse totale aérienne (variable indépendante). Ainsi, nous utilisons les deux équations suivantes :

$$\begin{aligned} RB_s &= 0,222 \times AB_s \\ RB_h &= 1,576 \times AB_h^{0,615} \end{aligned} \tag{4.2}$$

Où RB et AB sont respectivement la biomasse racinaire et aérienne et les indices s et h désignent le type d'essence respectivement conifère (*Softwood*) et feuillu (*Hardwood*).

La biomasse aérienne ainsi que la biomasse racinaire sont additionnées afin de calculer la biomasse totale. Cependant, selon (Nowak *et al.*, 2008), les arbres cultivés en plein air et entretenus ont tendance à avoir moins de biomasse aérienne que ce que prédisent les équations de biomasse dérivées des forêts pour les arbres du même DHP (Nowak, 1994). Pour tenir compte de cette différence, les résultats de la biomasse des arbres urbains sont multipliés par un facteur de 0,8. Ainsi, la

biomasse totale pour un arbre est obtenue par :

$$\begin{aligned} biomasse_{total} &= y_{total} + RB \\ biomasse_{urbain} &= 0,8 \times biomasse_{total} \end{aligned} \tag{4.3}$$

Concernant le calcul du stockage de carbone, la biomasse totale est simplement multipliée par 0,5 (Nowak et Crane, 2002).

Enfin, pour estimer la valeur monétaire associée au stockage du carbone par les arbres urbains, les valeurs de carbone ont été multipliées par 164$ (CAD) par tonne de carbone (Environnement et Changement climatique Canada, 2016).

4.1.2 Richesse et diversité

La richesse fait référence au nombre d'espèces dans un échantillon, une communauté ou une zone donnée. Quant à la diversité (Nolan et Callahan, 2006), elle est basée sur l'**indice de Shannon**. Elle fait référence au nombre d'espèces effectives et tient compte de l'abondance de chaque espèce ainsi que du nombre d'espèces. Un inventaire parfaitement homogène (avec le même nombre d'arbres pour chaque espèce) a un nombre d'espèces égal à sa richesse en espèces. Un nombre plus petit représente une répartition moins uniforme des espèces au sein de l'inventaire.

Les tableaux 4.1 et 4.2 illustrent deux exemples d'inventaires. Le nombre effectif d'espèces est de 5 pour l'inventaire équilibré, et $\sim 4,131$ pour l'inventaire inégal.

Cette approche est aussi utilisée pour le calcul de la diversité fonctionnelle en prenant en compte le nombre de groupes fonctionnels ainsi que leur abondance relative. Ainsi, le tableau 4.3 permet d'illustrer les différents niveaux de diversité fonctionnelle selon les indices de diversité calculés. Le cas idéal est un indice de

Espèces	Nb arbres
Acer platanoides	2
Acer saccharinum	2
Fraxinus pennsylvanica	2
Populus canescens	2
Quercus macrocarpa	2

Tableau 4.1: Inventaire équilibré

Espèces	Nb arbres
Acer platanoides	4
Acer saccharinum	3
Fraxinus pennsylvanica	1
Populus canescens	1
Quercus macrocarpa	1

Tableau 4.2: Inventaire inégal

diversité fonctionnelle de 9 ayant pour signification la présence de tous les groupes fonctionnels.

Indice de diversié fonctionnelle	Niveau de diversité
1 à 2,5	Très faible
2,5 à 5	Faible
5 à 7,5	Intermédiaire
7,5 à 9	Élevée

Tableau 4.3: Niveaux de diversité fonctionnelle selon les indices de diversité

4.1.3 Règle 10-20-30 (Santamour)

La règle de Santamour également appelée règle 10-20-30 est une ligne directrice pour réduire le risque de perte catastrophique d'arbres. La règle suggère qu'une population d'arbres urbains ne devrait pas inclure plus de 10 % d'une espèce, 20 % d'un genre ou 30 % d'une famille.

4.2 Calcul des améliorations

Afin de calculer l'amélioration apportée par l'ajout d'un arbre d'une espèce spécifique à l'inventaire, nous utilisons la formule suivante :

$$\frac{n_1}{n_0} - 1 = m \tag{4.4}$$

où

n_0 est le nombre d'espèce effectif avant l'ajout d'un nouvel arbre
n_1 est le nombre d'espèce effectif après l'ajout d'un nouvel arbre
m est le taux d'amélioration.

Il est à noter que nous devons calculer le nombre d'espèces effectif deux fois : une fois avant d'ajouter l'arbre et une autre fois après l'ajout de l'arbre. Afin d'optimiser le temps de calcul, nous utilisons une approximation lorsque nous travaillons avec un grand nombre d'espèces.

Par ailleurs, cette même méthode est utilisée pour le calcul des améliorations des services écosystémiques ou encore pour le calcul de l'amélioration d'autres indices biologiques survenant lors de l'adjonction d'un nouvel arbre.

4.3 Algorithme de recommandation

Nous émettons l'hypothèse qu'une interface visuelle offrant aux personnes utilisatrices la possibilité de contrôler le processus de recommandation et soutenant cette contrôlabilité avec une bonne visualisation du processus permettrait d'améliorer la qualité des recommandations du point de vue utilisateur notamment en offrant un moyen d'affiner les recommandations et de rendre le processus transparent. Par conséquent, nous proposons une intégration étroite des techniques de visualisation

et de recommandation pour permettre aux personnes utilisatrices d'interagir avec le système de recommandation à chaque étape. Ainsi, nous leur offrons la possibilité de pondérer les données d'entrée et les critères de recommandation en fonction de leurs objectifs.

Dans le cas de *SylvCiT*, l'outil logiciel que nous proposons, l'objectif principal est d'optimiser la résilience et d'augmenter le rendement en terme de services écosystémiques de la forêt urbaine analysée en augmentant notamment la diversité fonctionnelle. Les objectifs de la personne utilisatrice peuvent être une entrée du système et peuvent être définis comme la priorisation d'un service écosystémiques par rapport à un autre (par exemple : augmentation du stockage de carbone). En ces termes, l'objectif du système est d'identifier les essences d'arbres les plus adéquates, de les combiner avec le contexte de la forêt urbaine et de recommander les essences les plus favorables d'être plantées.

Ainsi, le contexte de la forêt urbaine comprend les informations suivantes :

- La localisation géographique des arbres analysés,
- Les essences d'arbres composant la sélection,
- La présence des arbres dans une zone d'îlot de chaleur.

Dans notre contexte, les objectifs de la personne utilisatrice sont utilisés pour affiner les résultats des recommandations et les classer.

Ainsi, les recommandations de SylvCiT :

- Sont abordées dans le bon contexte géographique et climatologique,
- Sont personnalisées, car elles ciblent les objectifs définis par la personne utilisatrice,
- Servent un objectif spécifique, qui est de diversifier au niveau fonctionnel la forêt urbaine.

En fonction des arbres sélectionnés par la personne utilisatrice et après qu'elle ait

choisi l'importance de chaque indice à améliorer, la base de données des espèces d'arbres candidates est réduite aux seules espèces d'arbres qui remplissent la règle de Santamour (Santamour Jr, 2004). En effet, un inventaire qui répond à la règle de Santamour est considéré comme plus diversifié qu'un inventaire qui n'y répond pas.

Ensuite, nous calculons les améliorations qu'un nouvel arbre entraînerait dans une zone donnée. Puis, nous pondérons l'amélioration apportée à chaque indice par la valeur fournie par la personne utilisatrice, additionnons les valeurs pondérées et calculons finalement un score pour chaque espèce.

Enfin, une liste d'espèces triées par score est recommandée. Cette liste doit être utilisée pour prendre des décisions sur les espèces les plus appropriées pour une zone compte tenu des caractéristiques d'arbres souhaitées. Le calcul du score se fait selon la formule suivante :

$$Score = \sum_{i=1}^{n} w_i x_i \tag{4.5}$$

Où w_i est le poids qui représente l'importance accordée par la personne utilisatrice à un indice spécifique et x_i est la valeur d'amélioration sur un indice réalisée en ajoutant un arbre à l'inventaire.

Évaluation. L'évaluation d'un système de recommandation est cruciale afin de mesurer la qualité et la pertinence des recommandations pour la personne utilisatrice. Cependant, le but d'un système de recommandation n'est pas le même pour tout le monde. En effet, il existe une multitude de systèmes de recommandation auxquels nous sommes exposés au quotidien (Amazon, Facebook ou encore Netflix). Pour chaque système de recommandation existe une ou plusieurs métriques permettant d'estimer la performance d'un algorithme selon des critères statistiques ou métiers (taux de clics, proportion ou nombre de recommandations acceptées, ...). En conséquence, le choix d'une ou plusieurs métriques d'évaluation

adaptées aux spécifications du système de recommandation est crucial. Pour cela, trois approches d'évaluation peuvent être envisagées : hors-ligne (*offline*), étude utilisateurs (*user study*), et en ligne (*online*) (Beel et Langer, 2015).

Les évaluations hors ligne sont la méthode d'évaluation la plus courante pour les systèmes de recommandation (Beel *et al.*, 2013). L'évaluation hors ligne permet de tester l'efficacité des algorithmes du système de recommandation sur un ensemble de données. Ainsi, généralement, pour effectuer une évaluation hors ligne, un algorithme de recommandation est testé sur un ensemble de données duquel certaines informations ont été supprimées. L'algorithme est ensuite analysé sur sa capacité à recommander les informations manquantes, transformant ainsi le problème en un problème de classification ou de prédiction. Ces évaluations peuvent se faire sur des ensembles de données réelles ou synthétiques. L'avantage d'un tel ensemble de données synthétiques ou *simulées* est la possibilité de simuler des scénarios probables et de soumettre le système à d'autres types de tests. Cependant, les données simulées peuvent présenter des biais et ne pas être exactement représentatives de données réelles. Les évaluations hors-ligne ont l'avantage de fournir des résultats très rapidement et permettent de tester plusieurs algorithmes entre eux avant de tester en ligne le plus pertinent, ou d'itérer sur les résultats hors-ligne pour proposer un algorithme pertinent dès le premier test réel. Les résultats de ces prédictions sont analysés en fonction d'une ou plusieurs métriques. Ces métriques proviennent majoritairement du domaine de la recherche d'information. La précision et le rappel (Buckland et Gey, 1994) font partie de ces métriques.

La *précision* est le rapport entre le nombre de bonnes recommandations et le nombre total de recommandations soumises à la personne utilisatrice.

$$\text{Précision} = \frac{\text{Nb. de bonnes recommandations}}{\text{Nb. total de recommandations}} \tag{4.6}$$

Le *rappel* est le rapport entre le nombre de bonnes recommandations et le nombre

total de bonnes recommandations existant dans l'ensemble de données.

$$\text{Rappel} = \frac{\text{Nb. de bonnes recommandations}}{\text{Nb. total de bonnes recommandations}} \tag{4.7}$$

Ces deux métriques sont utilisées pour calculer la F-mesure qui est la moyenne harmonique de la précision et du rappel. La F-mesure permet ainsi d'évaluer conjointement la précision et le rappel d'un système.

$$\text{F-mesure} = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{4.8}$$

Jusqu'à récemment, le domaine des systèmes de recommandation était principalement axé sur le test et l'amélioration de la précision des algorithmes de prédiction (McNee *et al.*, 2006; Ziegler *et al.*, 2005). Cependant, l'industrie et la communauté de recherche conviennent que le but des systèmes de recommandation est d'aider les personnes utilisatrices à prendre de meilleures décisions, et qu'une grande précision en soi ne garantit pas cet objectif (McNee *et al.*, 2006; Murray et Häubl, 2008; Pu *et al.*, 2011). En effet, ce type d'évaluation ne répond pas à la question de savoir si les personnes utilisatrices sont satisfaites des recommandations proposées par le système.

De nombreux autres facteurs, tels que la composition de l'ensemble de recommandations et des considérations personnelles (par exemple, la connaissance du domaine (Knijnenburg et Willemsen, 2010)) peuvent également affecter l'expérience utilisateur avec les systèmes de recommandation. Ainsi, l'étude des personnes utilisatrices à travers le recueil de leurs points de vue/satisfaction permet de collecter la perception, les interactions et plus généralement l'expérience utilisateur d'un système de recommandation.

Une méthode simple généralement utilisée, mais coûteuse en terme de temps, est le recrutement d'un ensemble de personnes utilisatrices afin d'observer leurs interactions, dans un environnement contrôlé, et d'enregistrer les temps nécessaires

pour effectuer une tâche ou encore la qualité des résultats de cette tâche. De plus, généralement, des questionnaires sont soumis aux personnes utilisatrices afin de recueillir des informations telles que l'appréciation de l'interface utilisateur ou encore la pertinence des recommandations. Par exemple, ResQue (*Recommender systems' Quality of user experience*) (Pu *et al.*, 2011) est un questionnaire composé de 60 questions permettant d'évaluer plusieurs facteurs influençant l'expérience de la personne utilisatrice avec un système de recommandation tels que, entre autres, la qualité des recommandations, leur utilité, la qualité des interfaces et des interactions.

L'évaluation en ligne permet aussi de recueillir certaines informations relatives à l'interaction des personnes utilisatrices avec le système de recommandation. Ce type d'évaluation est conduit sur des personnes utilisatrices réelles utilisant le système dans des conditions réelles sur une longue période sans être conscientes qu'elles appartiennent à une population de test. Cela permet de collecter les usages et habitudes des personnes utilisatrices, leur rétention et de mettre en évidence certains problèmes n'ayant pas été soulevés lors de l'étude utilisateurs. Les critères d'évaluation doivent être définis en amont, les plus populaires étant le taux de clics ou encore le pourcentage de recommandations validées. Aussi, ce type d'évaluation permet de mener des tests A/B afin de comparer plusieurs algorithmes ou interfaces utilisateur sur des populations différentes pour évaluer l'impact de ces modifications sur l'expérience utilisateur (Kohavi et Longbotham, 2017). En règle générale, ces systèmes redirigent un petit pourcentage du trafic vers chaque moteur de recommandation différent et enregistrent les interactions des personnes utilisatrices avec les différents systèmes.

Étant donné l'absence de données avec lesquelles nous pourrions comparer nos recommandations, nous opterons pour l'évaluation en ligne ainsi que l'étude utilisateur pour l'évaluation de notre système de recommandation. Cette évaluation sera

effectuée dans le cadre du Programme pilote *Visage Municipal* (FRQNT2020) en étroite collaboration avec trois municipalités du Québec, soit les villes de Boucherville, Saint-Lambert et Varennes. En plus de nous fournir des données qui seront utilisées dans le développement de l'outil, les spécialistes en foresterie urbaine de chacune de ces villes seront directement impliqué.e.s dans toutes les phases de développement des différents modules et notamment dans l'évaluation de la qualité des recommandations. Cela permettra ainsi d'adapter l'outil et les recommandations aux besoins des spécialistes urbains. Cependant, la collaboration avec ces villes débutant seulement en mai 2021, nous ne sommes pas en mesure de fournir une évaluation des recommandations. Nous proposerons cependant dans le chapitre suivant une étude de cas permettant de mettre en évidence les améliorations fournies par l'ajout des espèces d'arbres que nous recommandons.

4.4 SylvCiT : Implémentation et technologies

Un algorithme de recommandation efficace est essentiel pour inspirer confiance en un système de recommandation. Cependant, pour qu'un système de recommandation soit efficace, les interfaces ainsi que les technologies mises en œuvres jouent parfois un rôle au moins aussi important que la qualité des recommandations (Swearingen et Sinha, 2001).

Notre approche se base sur trois étapes :

1. Inventaire et évaluation de la composition de la forêt urbaine en termes de services écosystémiques délivrés (parties 3.1.1 et 4.1).
2. Recommandation de nouvelles espèces d'arbres à planter en fonction du besoin spécifié par la personne utilisatrice en termes de nombre, mais aussi de service à prioriser (partie 4.3).
3. Simulation des services délivrés par l'ensemble d'arbres analysé avec l'ajout

des arbres recommandés afin d'évaluer les apports aux services écosystémiques délivrés (partie 4.2).

SylvCiT a été implémenté à l'aide d'une variété de cadriciels et bibliothèques les plus adaptés à chacune des tâches nécessaires.

Les interfaces utilisateur ont été conçues à l'aide de l'éditeur de maquettes Balsamiq (Bal, 2020) puis améliorées au fur et à mesure de notre collaboration avec l'équipe de foresterie urbaine.

En raison du volume de données potentiel que nous devrons gérer à l'avenir, notamment dans le cadre du Programme pilote Visage Municipal, nous devions trouver un moyen efficace d'accéder aux inventaires d'arbres des différentes villes. Nous avons donc choisi d'indexer les données. L'indexation permet la recherche et l'extraction de n'importe quelle information de façon efficace et rapide.

Pour procéder à l'indexation, nous avons choisi d'utiliser Apache Solr (Solr, 2021), une plateforme de recherche en code source libre basée sur le moteur d'indexation Apache Lucene (Lucene, 2021). Solr est écrit en Java et fonctionne comme un serveur de recherche de texte intégral. En tant que moteur de recherche, il comprend la recherche en texte intégral, l'indexation en temps quasi réel, le regroupement dynamique, l'intégration de bases de données, la gestion de mutliples formats de documents (par exemple, csv, json) et la recherche géospatiale. Ses autres atouts sont le passage à l'échelle, la tolérance aux pannes, l'indexation distribuée, la réplication et la répartition de charge, le basculement et la restauration automatisée. Cette conception confère à notre système des performances satisfaisantes, la possibilité de créer des requêtes diverses et complexes pour récupérer des données et un traitement de requête à grande vitesse.

Nous avons également utilisé les outils géospatiaux fournis par Solr. En effet, Solr dispose de trois outils intégrés, à savoir un filtre géospatial, une boîte de

délimitation géospatiale et une fonction de calcul de distance géospatiale. Avec l'aide du géo-filtre, on peut récupérer tous les *documents* (dans notre contexte, des arbres) pertinents à distance d'un point donné, par ex. récupérer tous les documents dans un rayon de 10 km à partir d'une latitude / longitude donnée. La zone de délimitation est utile pour filtrer les résultats dans une zone spécifiée, par exemple dans un polygone sélectionné par une personne utilisatrice. La fonction de distance géospatiale est utile pour trier les résultats en fonction de la distance des documents par rapport aux points de requête.

Selon Taschuk et Wilson (Taschuk et Wilson, 2017), les logiciels produits pour la recherche, publiés ou non, souffrent d'un certain nombre de problèmes courants qui rendent difficile, voire impossible, leur exécution en dehors de l'institution d'origine. Pour pallier à ce problème, lors de la phase de développement, nous nous sommes basés sur les dix règles proposées par ces deux chercheurs, afin d'assurer la reproductibilité des expériences. Dans cette logique et afin d'assurer la portabilité, autrement dit, la capacité à être exécuté sous n'importe quel environnement, nous nous reposons sur Docker (Merkel, 2014). Docker est une technologie en code source libre de virtualisation de conteneurs similaire à une machine virtuelle plus légère et plus performante (Anderson, 2015). Cette technologie permet de résoudre certains problèmes identifiés par Boetigger (Boettiger, 2015) tels que :

— « L'enfer des dépendances » (*Dependency Hell*) : Le problème de dépendance survient lorsque plusieurs briques logicielles ont des dépendances à des librairies partagées, mais avec des versions différentes et incompatibles. En fournissant une image binaire dans laquelle tout le matériel a déjà été installé, configuré et testé, Docker règle ce problème,

— Le problème de la documentation imprécise, notamment à travers les *Dockerfiles* et *Docker-compose* qui sont des scripts simples, écrits généralement en *YAML*, décrivant exactement comment construire l'image et or-

chestrer les conteneurs,

— « L'érosion logicielle » (*code-rot*) notamment liée aux changements continus introduits dans le système ou son environnement le rendant non-maintenable. Docker fournit des utilitaires offrant un moyen robuste d'enregistrer et d'exécuter les versions exactes d'un logiciel,

— Les barrières à l'adoption et la réutilisation, d'une part, en fournissant la possibilité de réutiliser des images et conteneurs et d'autre part à travers le système de gestion de version mis en place similaire à Git (Loeliger et McCullough, 2012).

L'application dorsale (*backend*) est basée sur le langage de script en code source libre Python et le cadriciel web basé sur Python Django (Django, 2021). L'application frontale (*frontend*) a été développée avec le cadriciel ReactJS(React, 2020), populaire pour la construction d'interfaces utilisateur sur le Web.

Enfin, le code source est accessible au public et disponible pour faciliter la reproduction des résultats. Le système est gratuit, ouvert et publié sous la licence GPL-v3. Il est disponible dans le dépôt suivant :
`https://gitlab.ikb.info.uqam.ca/ikb-lab/data-science/sylvcit`.

Afin de permettre aux personnes utilisatrices de manipuler notre outil et de visualiser par elles-même les arbres publics de Montréal, d'effectuer des analyses et de recevoir des recommandations, nous avons mis à disposition SylvCiT en accès libre à l'adresse suivante : `https://sylvcit.uqam.ca/`

CHAPITRE V

RÉSULTATS ET DISCUSSIONS

Nous décrivons et commentons dans ce chapitre les résultats obtenus suite à nos expériences. Nous présenterons pour chaque partie les résultats obtenus, les discuterons et apporterons des éléments de réflexion concernant les perspectives futures pour chaque partie. Ainsi, les parties 5.1 et 5.2 présentent respectivement les résultats relatifs au regroupement fonctionnel et la prédiction des diamètres à hauteur de poitrine, tandis que la partie 5.3 conclut ce chapitre par la présentation d'un cas d'usage.

5.1 Regroupement fonctionnel

5.1.1 Résultats

	K-moyennes	**Reg. agglomératif**	**Mélange gaussien**	**Part. spectral**	**DBScan**	**K-modes**	**K-prototypes**
V-mesure	0,709	0,772	0,692	0,748	0,557	0,560	0,690

Tableau 5.1: Résultats du regroupement fonctionnel

Le tableau 5.1 présente les résultats des regroupements de l'ensemble de données composé de 11 caractéristiques (7 numériques et 4 catégoriques) de 271 espèces. Ce regroupement est effectué en 10 groupes en utilisant plusieurs algorithmes de regroupement. La métrique utilisée est la V-mesure décrite à la partie 3.2. Cette

métrique permet de valider la qualité des groupes formés. Ainsi, un résultat de 1 signifie que l'homogénéité ainsi que la complétude des groupes sont maximisées. À l'inverse, si le regroupement ne satisfait aucune des deux conditions, la V-mesure sera égale à zéro. Ainsi, nous comparons le résultat en sortie de chaque algorithme de regroupement avec le résultat du regroupement hiérarchique initial proposé par Cameron et Paquette (Cameron et Paquette, 2016). Nous avons effectué plusieurs expériences et nous présentons ici les meilleurs résultats pour chaque technique utilisée.

En évaluant les résultats de ces expériences, la V-mesure révèle des résultats différents en fonction de l'algorithme utilisé.

Le regroupement agglomératif avec une distance euclidienne ainsi que la méthode de Ward comme critère d'agrégation a obtenu les meilleurs résultats pour l'ensemble des traits fonctionnels composés seulement des traits numériques.

Nous avons également testé un partitionnement spectral qui effectue le regroupement sur un espace de faible dimension après la reconstruction de l'ensemble sous forme de graphe. En effet, la topologie de l'ensemble de données, qui montre un chevauchement entre les points, nous a motivés à utiliser cet algorithme. Ainsi, le partitionnement spectral a montré les meilleurs résultats sur l'ensemble de données composé seulement des traits numériques.

La performance de K-moyennes est moindre en comparaison avec les deux algorithmes précédents en raison de son fonctionnement et de l'utilisation de la distance euclidienne. En effet, cherchant à minimiser les distances par rapport à un centroïde comme centre de gravité d'un groupe, cet algorithme aura pour effet d'identifier des groupes de formes sphériques.

Nous avons supposé que notre ensemble de données suivait une distribution gaus-

sienne et avons donc essayé de regrouper les variables numériques avec un modèle gaussien. Cela a donné un résultat légèrement inférieur à ceux obtenus avec les approches précédentes.

Nous avons testé l'algorithme DBScan sur l'ensemble de données contenant seulement les variables numériques, pour lequel nous avons défini la valeur d'epsilon de manière empirique afin de nous rapprocher au mieux du nombre de groupes voulu (10). Les résultats sont inférieurs à ceux obtenus par les algorithmes précédents. Cela pourrait être dû au fait que DBScan éprouve des difficultés à trouver des groupes de densités différentes.

Nous avons tenté d'utiliser K-modes qui est un algorithme destiné au regroupement des données catégoriques. Pour cela, nous avons utilisé l'ensemble de données contenant les valeurs numériques discrétisées ainsi que les données catégoriques. Les résultats sont très inférieurs comparés aux autres algorithmes, ce qui pourrait être dû à une perte d'information survenue lors de la discrétisation.

Enfin, nous avons tenté d'effectuer le regroupement avec l'algorithme K-prototypes, algorithme populaire pour les ensembles de données composés de variables mixtes (numérique et catégorique). Cela a montré un résultat peu satisfaisant, l'algorithme n'arrivant pas à converger vers une solution optimale en 1 000 itérations.

Tous les algorithmes que nous avons expérimenté et dont les résultats sont décrits ci-dessus sont détaillés dans la partie 3.2.2.

En conclusion, malgré son manque de robustesse, sa sensibilité aux bruits et aux valeurs aberrantes (Xu et Wunsch, 2005), la technique de regroupement hiérarchique mise en œuvre, initialement inspirée des travaux de l'approche fonctionnelle de Cameron et Paquette (Cameron et Paquette, 2016), reste préférable selon l'analyse de la composition des groupes formés. En effet, cette technique per-

met de former des groupes interprétables composés d'espèces proches en terme de caractéristiques et de traits fonctionnels. L'interprétation des groupes fonctionnels est illustrée par le tableau 5.2 extrait de la méthodologie de Cameron et Paquette (Cameron et Paquette, 2016).

Groupe	**Type fonctionnel**	**Espèces représentatives**
1A	Conifères généralement tolérants à l'ombre, mais pas à la sécheresse ou l'inondation.	Les épinettes, sapins et thuya, et le pin blanc.
1B	Conifères héliophiles, tolérants à la sécheresse (pins).	Les pins, mélèzes, genévriers, et ginkgo
2A	Arbres tolérants à l'ombre à feuilles larges et minces, croissance moyenne.	La plupart des érables, les tilleuls et quelques autres petits arbres.
2B	Ressemblent à 2A sauf pour les semences très lourdes et dispersées par la gravité.	Les marronniers
2C	Grands arbres tolérants à l'inondation.	La plupart des ormes, les frênes, micocoulier, érables rouges, argentés, et negundo
3A	Petits arbres tolérants à la sécheresse, bois lourd, feuilles épaisses, croissance faible.	Rosacées (sorbier, poirier, aubépine et amélanchier), et les lilas.
3B	Groupe « moyen ». Intolérant à l'inondation.	Grandes Rosacées (cerisier, pommier), Catalpa, Maackia, autres espèces diverses.
4A	Grands arbres à semences et bois lourds. Plusieurs tolérants à la sécheresse.	Les chênes, noyers, et caryers.

4B	Grande tolérance à la sécheresse, mais pas à l'ombre ou à l'inondation. Semences lourdes, feuilles riches.	Les légumineuses (févier, chicot, robinier, gainier)
5	Espèces pionnières à très petites semences. Croissance rapide, tolérantes à l'inondation, bois léger.	Tous les peupliers, saules, aulnes et bouleaux (sauf jaune)

Tableau 5.2: Interprétation des groupes fonctionnels

5.1.2 Discussions

Le regroupement de données est une méthode fondamentale et très populaire d'analyse des données. Cependant, sa nature subjective signifie que différents algorithmes de regroupement ou une préparation des données différente peuvent produire des résultats très variés et parfois contradictoires. En effet, c'est ce que nous avons pu observer à travers les résultats de nos expériences.

Aucune méthode mise en œuvre n'a pu atteindre les résultats obtenus via la méthode initiale de regroupement hiérarchique, que nous avons reproduite avec succès. Ces méthodes dépendent fortement de l'algorithme de regroupement utilisé, de la fonction de calcul de distance et du type de données. Plusieurs autres méthodes de regroupement pourraient être utilisées, mais l'avis d'un.e expert.e est nécessaire afin de valider la qualité des groupes formés.

Nos travaux de recherche peuvent être encore améliorés notamment en augmentant le nombre d'espèces analysées. Cependant, le manque de données et la variabilité des traits fonctionnels dépendant des conditions climatiques entre autres, rendent difficile l'élaboration d'un modèle de regroupement unique. Ce dernier devra être construit pour les espèces de chaque région climatique différente.

5.2 Prédiction des diamètres à hauteur de poitrine

5.2.1 Résultats

	MAE (cm)	**MSE** (cm^2)	**RMSE** (cm)
LR	5,294	60,202	7,759
RF_100_minsamples2	**3,695**	**38,977**	**6,243**
RF_1000_minsamples3	3,662	36,813	6,067
RF_1600_minsamples3	3,662	36,808	6,067
NN_lr0.001_0,5_1hiden_layer_sigmoid	**4,366**	**48,632**	**6,974**
NN_lr0.001_0,5_2hiden_layer_sigmoid	4,552	52,610	7,253
NN_lr0.001_0,5_3hiden_layer_sigmoid	4,746	52,779	7,265
NN_lr0.001_0,5_4hiden_layer_sigmoid	5,098	54,950	7,412
NN_lr0.001_0,5_1hiden_layer_relu	**4,532**	**50,483**	**7,110**
NN_lr0.001_0,5_2hiden_layer_relu	5,199	55,495	7,450
NN_lr0.001_0,5_3hiden_layer_relu	5,008	51,914	7,210
NN_lr0.001_0,5_4hiden_layer_relu	5,400	60,490	7,780
NN_lr0.001_0,5_1hiden_layer_tanh	**4,354**	**49,413**	**7,029**
NN_lr0.001_0,5_2hiden_layer_tanh	4,556	51,911	7,205
NN_lr0.001_0,5_3hiden_layer_tanh	4,629	54,678	7,394
NN_lr0.001_0,5_4hiden_layer_tanh	5,071	59,801	7,733

Tableau 5.3: Résultats des prédictions des DHP

Nous avons récapitulé les résultats de nos expériences dans le tableau 5.3. Ces résultats ont été obtenus sur le jeu de données de l'ensemble des arbres publics de Montréal avec une validation croisée $k = 5$, pour tester un modèle de prédiction de croissance des arbres urbains de la ville de Montréal. Pour rappel, nous avons testé une régression linéaire (LR), trois modèles basés sur les forêts d'arbres décisionnels (RF) et des réseaux de neurones (NN) (algorithmes décrits en 3.3.2)

Dans l'ensemble, les résultats du tableau 5.3 montrent que le RF obtient de meilleures prédictions que les autres algorithmes. On peut observer que les valeurs de MAE, MSE et RMSE sont inférieures à celles des autres modèles, ce qui se traduit par une meilleure performance de prédiction. Cependant, une analyse plus approfondie est encore nécessaire pour valider si cet avantage est suffisamment stable. Lors de nos expériences, nous avons constaté que l'augmentation du nombre d'arbres du RF au-dessus de 100 améliore de très peu la précision de l'estimation (de l'ordre de 0,033 cm) et augmente le temps nécessaire à la convergence de l'algorithme. Les bonnes performances du RF s'expliquent par :

- sa capacité à gérer les relations non linéaires entre le DHP et les variables explicatives fournies en entrée,
- sa robustesse aux valeurs aberrantes,
- sa capacité à fournir des prédictions robustes pour des jeux de données débalancés (Biau et Scornet, 2016; Liu *et al.*, 2013).

Les résultats des modèles à base de réseaux de neurones ne montrent pas une variabilité importante. Ainsi, en deuxième position en terme de performance, figure le réseau de neurones à une couche cachée avec la fonction d'activation tanh. La faible performance des réseaux de neurones utilisant les fonctions sigmoïde et tanh comparé aux forêts d'arbres décisionnels pourrait être expliquée par le fait que les neurones saturent, c'est-à-dire que ces derniers sont tout le temps activées et deviennent insensibles aux variations du signal en entrée. Il devient alors difficile pour l'algorithme d'apprentissage de continuer à adapter les poids pour améliorer les performances du modèle (Goodfellow *et al.*, 2016a).

La fonction d'activation ReLU a plusieurs avantages tels que la résolution du problème de la disparition du gradient et l'accélération de la convergence du réseau lors de son entraînement (Goodfellow *et al.*, 2016a). Néanmoins, ReLU souffre du problème des « neurones mourants » (*dying neurons*) lors duquel certains neu-

rones ne sont jamais activés si les poids ne sont pas mis à jour correctement. En d'autres termes, si trop d'activations tombent en dessous de zéro, alors la plupart des neurones seront désactivées bloquant de ce fait l'apprentissage. Ceci pourrait expliquer la performance moindre de cette configuration comparée aux autres modèles.

On peut aussi noter que les réseaux de neurones à une seule couche cachée surpassent les autres modèles à architecture plus profonde. Cela devra être analysé plus en détails, car cela indique que l'utilisation de réseaux de neurones à plus d'une couche est superflue pour notre cas d'étude. Selon Heaton (Heaton, 2008) pour de nombreux problèmes pratiques, il n'y a aucune raison d'utiliser plus d'une couche cachée.

Enfin, on remarque que la régression linéaire a obtenu de moins bons résultats que les autres modèles. Cela pourrait être expliqué par le fait que ce modèle n'a pas été capable de capturer la complexité non-linéaire des données en entrée.

5.2.2 Discussions

Les réseaux de neurones et les forêts d'arbres décisionnels sont des algorithmes prometteurs pour modéliser les taux de croissance des arbres urbains. Nos résultats soulignent l'importance de la qualité et de la préparation des données. De plus, les modèles développés ont une puissance prédictive intéressante et permettent de former un modèle unique pour un ensemble d'espèces, contrairement aux approches existantes dans la littérature ayant développé une équation allométrique unique par espèce pour chaque région climatique.

Plusieurs limitations doivent être prises en compte lors de l'interprétation de nos résultats. Premièrement, la croissance a été estimée à l'aide des données fournies par la ville de Montréal, pouvant contenir parfois des erreurs de mesures

et de transcriptions. Deuxièmement, la procédure de nettoyage des données a entraîné une perte d'information importante, réduisant l'ensemble de données initial de moitié en raison des valeurs manquantes pour certaines observations et de la présence de valeurs aberrantes. Néanmoins, ces analyses ont utilisé un ensemble de données relativement volumineux avec de nombreuses observations pour la construction de modèles. Finalement, nous nous sommes appuyés sur un nombre de variables restreint pour l'entraînement de nos modèles, ce qui peut avoir limité la capacité des modèles à prédire la croissance des arbres.

5.2.3 Perspectives futures

Dans les parties précédentes, nous avons présenté les résultats de plusieurs techniques basées sur l'apprentissage automatique pour la prédiction des taux de croissance des arbres urbains de Montréal. Un obstacle à la mise en œuvre pratique de cette méthode est la variabilité des conditions climatiques et d'entretiens des arbres qui ont une influence directe sur leur croissance. Bien que les résultats montrent de bonnes performances, la mise en œuvre du modèle reste limitée à la région climatique de Montréal. La solution la plus souhaitable réside dans le développement d'un modèle unique, apte à prédire les taux de croissances avec une bonne précision dans diverses conditions. Les travaux futurs seront ainsi axés sur la validation de la méthode proposée avec différents ensembles d'arbres au Québec et au Canada. Pour atteindre cet objectif, nous avons commencé la collecte de ces ensembles de données. Cependant, la disponibilité et la qualité des données représentent un risque pour la réalisation du projet. Le processus de croissance implique souvent des processus couplés à des conditions de vie complexes. Nous prévoyons collecter des variables relatives au contexte de vie telles que la qualité du sol, la présence en îlot de chaleur ou encore la présence de l'arbre en parc ou sur rue afin d'affiner la qualité des prédictions.

Comprendre les mécanismes sous-jacents intervenant dans la croissance des arbres en milieu urbain permettra une meilleure prédiction de l'état de santé de nos forêts urbaines dans le futur.

5.3 SylvCiT

5.3.1 Cas d'utilisation

Dans cette partie, nous allons développer un cas d'utilisation plausible de SylvCiT pour la plantation de nouveaux arbres. L'objectif est de démontrer de quoi l'outil est capable, comment il permet d'aider à mieux comprendre la forêt urbaine et à fournir une aide à la décision pour le choix des essences à planter.

Une étude de la forêt urbaine se faisant généralement pour un quartier ou un arrondissement, nous avons sélectionné le quartier de Verdun (Montréal) pour notre scénario. L'étude inclut tous les arbres publics présents sur ce territoire. L'objectif global est d'entreprendre une évaluation réaliste du potentiel de stockage du carbone du quartier de Verdun. Les objectifs spécifiques de l'étude sont les suivants :

- Caractériser un inventaire complet des arbres du quartier avec mesure précise de la richesse en espèces,
- quantifier le stockage brut de carbone de la forêt urbaine et sa valeur monétaire,
- évaluer l'impact de la plantation des essences d'arbres suggérées.

La figure 5.1 illustre l'interface d'accueil de SylvCiT, permettant notamment de visualiser la distribution des arbres sur une carte en fonction de leur localisation. Les codes couleurs et les diamètres de chaque point sont définis en fonction respectivement de l'espèce et du DHP. De plus, une liste de la proportion d'arbres

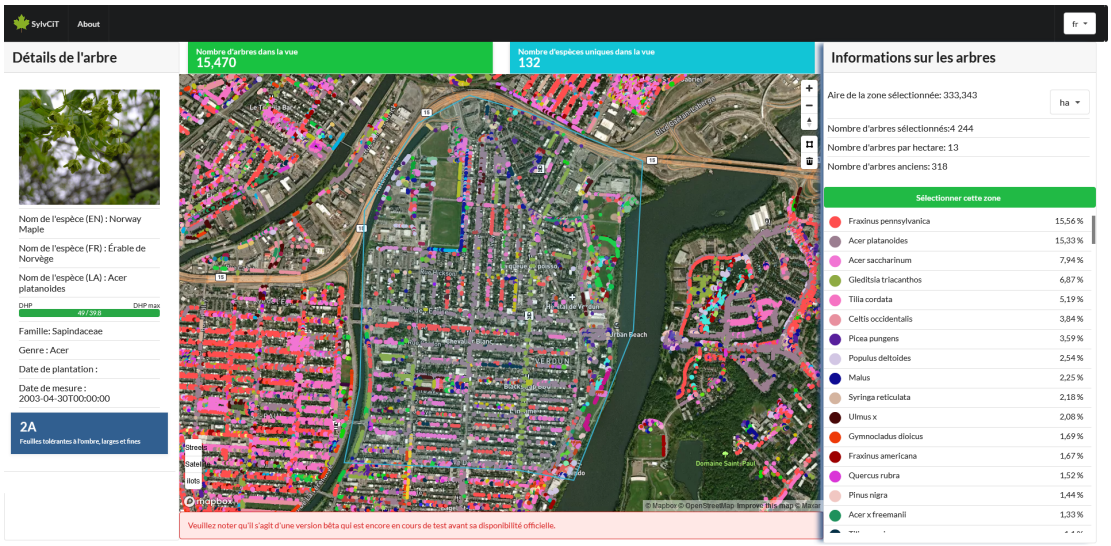


Figure 5.1: Affichage des arbres de Montréal - SylvCiT

par espèces dans la vue actuelle offre un aperçu des espèces les plus présentes.

On peut remarquer que les frênes et les érables (respectivement *Fraxinus* et *Acer*) représentent plus de 30% de la forêt urbaine, ces espèces constituent une part importante du patrimoine arboricole montréalais en général. D'autres statistiques sont aussi mises à disposition telles que le nombre d'arbres dans la vue ainsi que le nombre d'espèces uniques.

Par ailleurs, lors du clic sur un arbre dans la carte, des détails concernant ce dernier sont affichés. Ces détails incluent notamment l'espèce, la famille, le genre, le groupe fonctionnel, mais aussi le DHP actuel ainsi que sa date de plantation et de mesure (si disponibles).

En outre, la personne utilisatrice a la possibilité de changer la couche affichée, par exemple, pour une couche permettant de visualiser les îlots de chaleur dans la zone étudiée.

Enfin, cette interface permet de sélectionner un ensemble d'arbres pour l'analy-

ser. Dans notre cas, nous avons sélectionné une surface d'environ 333 hectares, comportant 4244 arbres.

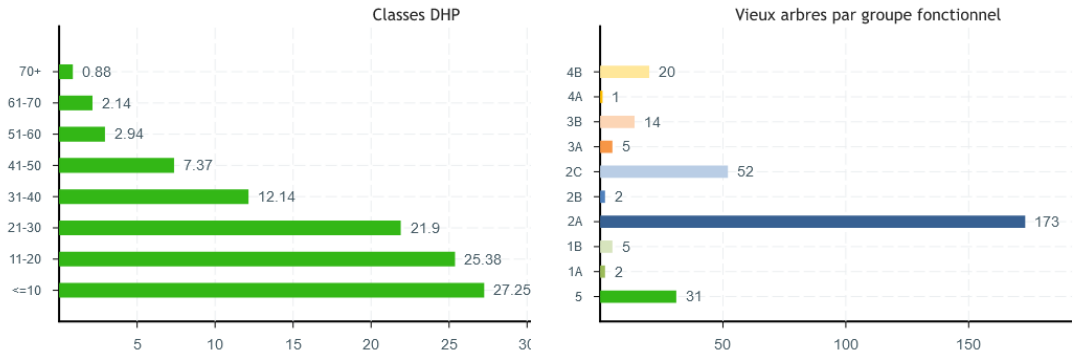


Figure 5.2: Proportion d'espèces par classe de DHP et distribution des vieux arbres par groupe fonctionnel - SylvCiT

Lors de l'analyse, des informations concernant la richesse et la diversité d'espèces ainsi que celles des groupes fonctionnels sont calculées selon la sélection effectuée. La figure 5.2 montre le résultat de l'analyse des données des arbres sélectionnés en fonction du DHP, permettant d'évaluer la répartition des arbres en fonction de leur âge ainsi que de leur groupe fonctionnel. Il en ressort que le patrimoine arboricole de Verdun est plus ou moins jeune avec plus de la moitié des arbres ne dépassant pas un DHP de 20 cm permettant à terme d'assurer la relève des arbres matures. Cependant, il est à noter qu'une attention particulière devra être portée aux jeunes arbres afin d'assurer leur survie. Cette présence importante de jeunes arbres peut être expliquée par le plan arboricole défini par l'arrondissement de Verdun en 2014 afin d'accroître son couvert forestier ainsi que les opérations d'abattage des frênes souffrant ou amenés à souffrir de l'agrile du frêne (Arrondissement de Verdun, 2014).

Dans le graphique présentant la proportion de vieux arbres par groupe fonctionnel, on remarque la présence importante des arbres des groupes 2A et 2C constitués

en majorité d'érables, ormes et frênes, qui représentent les espèces plantées historiquement à Montréal. Ces groupes représentent aussi 32,83% des arbres pour le 2A et 20,55% pour le 2C.

SylvCiT fournit également une estimation du stockage de carbone et de sa valeur monétaire. Pour l'ensemble d'arbre de notre sélection, le stockage de carbone est de 916 462,55 kg pour une valeur de 151 692,8$ CAD. Ces chiffres peuvent être expliqués par la relative jeunesse de la forêt urbaine analysée.

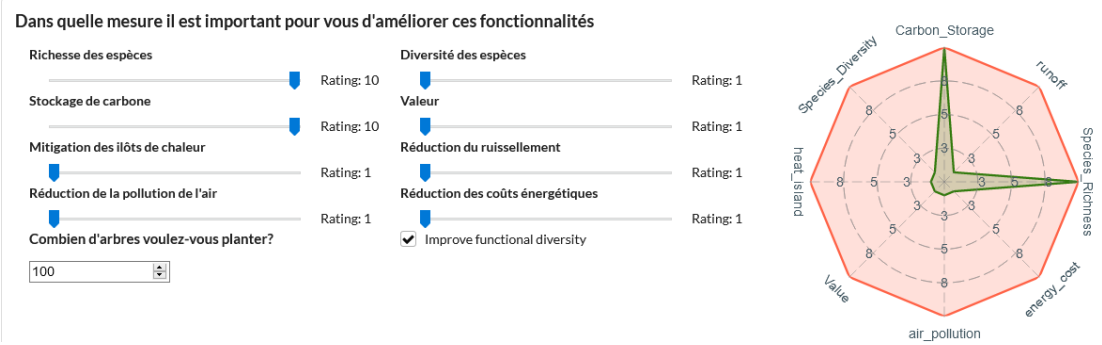
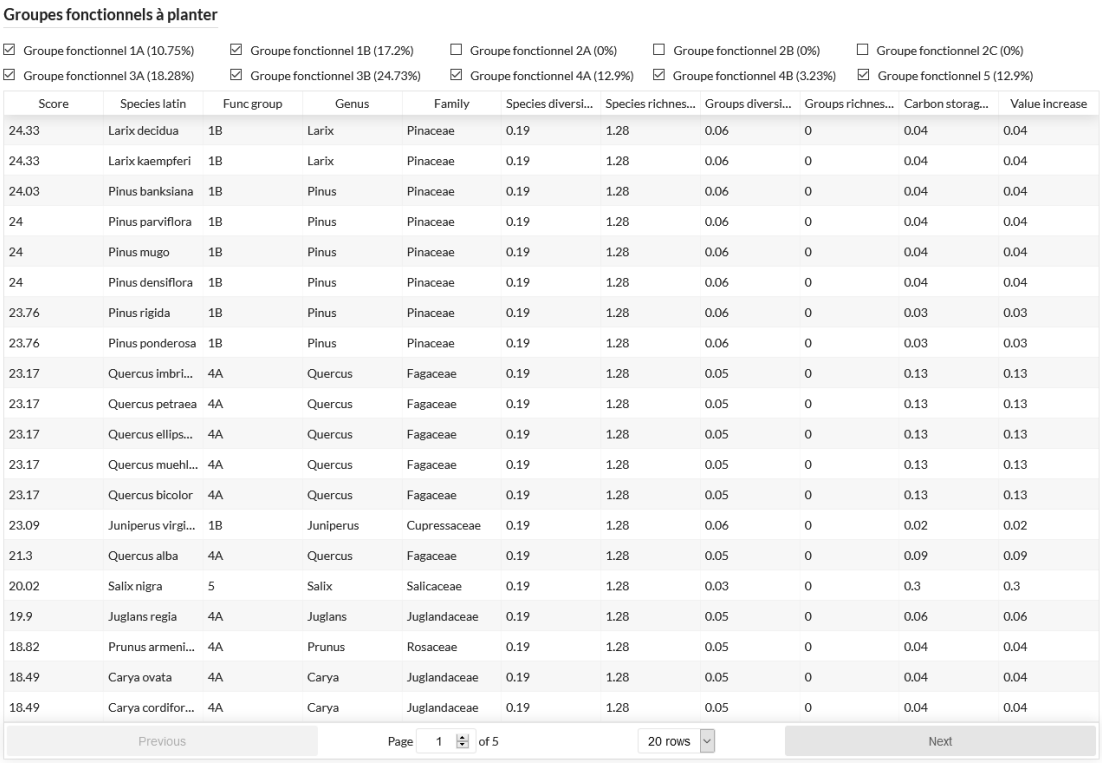


Figure 5.3: Choix du nombre d'arbre et critères de pondération - SylvCiT

La figure 5.3 montre comment la personne utilisatrice peut choisir le nombre d'arbres à planter et pondérer certains critères de recommandation en fonction de ses besoins. Cette pondération sera prise en compte lors de la recommandation des groupes fonctionnels/espèces. Pour notre scénario, nous avons choisi d'améliorer la richesse des espèces ainsi que le potentiel de stockage de carbone. D'autres critères auraient pu être choisis tels que la réduction de la pollution de l'air ou encore du ruissellement des eaux.

Lors de la conception de notre outil, nous avons analysé en détail les systèmes de recommandation interactifs existants et trouvé différentes techniques de visualisation qui ont été utilisées pour prendre en charge le contrôle donné à la personne utilisatrice pour améliorer les recommandations ou pour explorer l'espace de re-

commandation. Parmi ces techniques, nous avons constaté que les curseurs et les graphiques avec des éléments déplaçables sont les plus populaires. Nous avons donc adopté ces éléments qui permettent notamment d'améliorer l'interaction de la personne utilisatrice avec le système et de lui offrir un contrôle total des recommandations selon ses critères.



Groupes fonctionnels à planter

☑ Groupe fonctionnel 1A (10.75%) ☑ Groupe fonctionnel 1B (17.2%) ☐ Groupe fonctionnel 2A (0%) ☐ Groupe fonctionnel 2B (0%) ☐ Groupe fonctionnel 2C (0%)
☑ Groupe fonctionnel 3A (18.28%) ☑ Groupe fonctionnel 3B (24.73%) ☑ Groupe fonctionnel 4A (12.9%) ☑ Groupe fonctionnel 4B (3.23%) ☑ Groupe fonctionnel 5 (12.9%)

Score	Species latin	Func group	Genus	Family	Species diversi...	Species richnes...	Groups diversi...	Groups richnes...	Carbon storag...	Value increase
24.33	Larix decidua	1B	Larix	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
24.33	Larix kaempferi	1B	Larix	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
24.03	Pinus banksiana	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
24	Pinus parviflora	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
24	Pinus mugo	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
24	Pinus densiflora	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.04	0.04
23.76	Pinus rigida	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.03	0.03
23.76	Pinus ponderosa	1B	Pinus	Pinaceae	0.19	1.28	0.06	0	0.03	0.03
23.17	Quercus imbri...	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.13	0.13
23.17	Quercus petraea	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.13	0.13
23.17	Quercus ellips...	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.13	0.13
23.17	Quercus muehl...	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.13	0.13
23.17	Quercus bicolor	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.13	0.13
23.09	Juniperus virgi...	1B	Juniperus	Cupressaceae	0.19	1.28	0.06	0	0.02	0.02
21.3	Quercus alba	4A	Quercus	Fagaceae	0.19	1.28	0.05	0	0.09	0.09
20.02	Salix nigra	5	Salix	Salicaceae	0.19	1.28	0.03	0	0.3	0.3
19.9	Juglans regia	4A	Juglans	Juglandaceae	0.19	1.28	0.05	0	0.06	0.06
18.82	Prunus armeni...	4A	Prunus	Rosaceae	0.19	1.28	0.05	0	0.04	0.04
18.49	Carya ovata	4A	Carya	Juglandaceae	0.19	1.28	0.05	0	0.04	0.04
18.49	Carya cordifor...	4A	Carya	Juglandaceae	0.19	1.28	0.05	0	0.04	0.04

Previous Page 1 of 5 20 rows Next

Figure 5.4: Liste des groupes et espèces recommandés - SylvCiT

La figure 5.4 offre un aperçu des groupes fonctionnels/espèces recommandés en fonction de la sélection d'arbres et des critères choisis à l'étape précédente. La recommandation des arbres est dans un ordre décroissant en fonction du score obtenu. Ce score est calculé en fonction de l'amélioration apportée par l'espèce sur la diversité, la richesse, le stockage de carbone et la valeur monétaire. Chaque critère est pondéré selon les choix effectués dans l'interface précédente (décrit en 4.3).

La personne utilisatrice peut choisir de supprimer ou d'inclure des groupes fonctionnels à partir desquels les espèces seront recommandées. On remarque que les groupes fonctionnels les moins présents dans l'ensemble analysé sont les plus recommandés. Un fait intéressant à noter est que le groupe 2B n'a pas été recommandé, malgré sa présence minime dans l'ensemble analysé. Ceci pourrait s'expliquer par le fait que parmi nos espèces candidates à la recommandation seulement 13 appartiennent au groupe 2B, constitués de marronniers. De plus, nos critères de recommandation étaient basés principalement sur l'amélioration de la richesse des espèces et du stockage de carbone, ce qui a eu pour effet d'éliminer ce groupe des recommandations. Un fait intéressant est que les espèces du groupe 2B (constituant seulement 103 arbres sur les 4244 analysés) ont obtenus les meilleurs scores et représentent 17,2% des espèces recommandées. Le groupe 4A représente 12,9% des espèces recommandées et se place en deuxième position du nombre d'arbres pour ce groupe après le 2B.

Ainsi, les groupes et espèces recommandées par SylvCiT sont des choix potentiels avantageux répondant aux critères définis précédemment.

Ce cas d'étude peut être reproduit, que ce soit pour l'arrondissement de Verdun ou tout autre arrondissement de Montréal, à travers l'outil SylvCiT disponible à l'adresse suivante : `https://sylvcit.uqam.ca/`

5.3.2 Discussion

Notre étude illustre l'importance de l'analyse de l'état actuel de la forêt urbaine pour éclairer les futures décisions de planification et de gestion. Nos recommandations sont effectuées en fonction de l'état actuel de la forêt analysée et s'inscrivent dans une stratégie de gestion globale et intégrée des arbres à l'échelle d'un quartier, arrondissement ou encore d'une ville.

La base de données de SylvCiT contient plus de 500 espèces candidates à la recommandation, indigènes et non-indigènes, sous-espèces, variétés, hybrides et cultivars, sélectionnés en amont par l'équipe de foresterie urbaine collaborant au projet. Par conséquent, SylvCiT considère un large spectre d'espèces d'arbres pour divers contextes urbains.

SylvCiT fournit une analyse ainsi qu'une aide à la décision précieuse pour la sélection des arbres par les décideurs en matière de plantation urbaine, avec comme idée principale d'avoir le bon arbre au bon endroit et que ce dernier soit résilient aux changements globaux notamment à travers la diversification des espèces recommandées.

Une méthode simple d'application de la diversification fonctionnelle sur laquelle repose notre outil consiste à regrouper les espèces d'arbres selon la similitude de leurs traits fonctionnels, créant alors des groupes fonctionnels (Cameron et Paquette, 2016). Par exemple, des espèces tolérantes à la sécheresse, de haute taille et ayant une écorce épaisse seront incluses au sein d'un même groupe, alors que des espèces intolérantes à la sécheresse, de faible taille et ayant une écorce mince formeront un groupe différent. En s'assurant de maintenir au sein de nos forêts urbaines des espèces d'arbres issues de groupes fonctionnels nombreux et éloignés, en proportions plus ou moins égales et avec une certaine redondance, on s'assure donc d'immuniser nos forêts urbaines contre un grand nombre de stress distincts, connus et inconnus, et de minimiser leur exposition au risque. La diversification des espèces plantées est aussi particulièrement importante pour la variété des services écosystémiques produits puisque chaque espèce va fournir des services particuliers.

En outre, contrairement à I-Tree (i-Tree, 2021) (décrit à la partie 2.2), le code source de SylvCiT est accessible par la communauté de recherche et requiert peu

de connaissances du domaine de la foresterie urbaine, l'outil étant pensé pour être utilisable par une large gamme de personnes des plus novices aux plus expertes.

Par ailleurs, lors des recommandations, nous nous abstenons de fournir des suggestions strictes de plantation d'essences. De plus, nous reconnaissons le jugement individuel des différents groupes de personnes utilisatrices. Enfin, la gestion et la planification des forêts urbaines sont soumises à plusieurs contraintes telles que la disponibilité des espèces en pépinière ou encore des considérations esthétiques.

5.3.3 Perspectives futures

Plusieurs pistes de recherche s'offrent à nous. Le domaine de la foresterie urbaine est un domaine en pleine effervescence qui s'intéresse à l'application de techniques informatiques pour résoudre certaines problématiques rencontrées par les villes. Ci-dessous, nous listerons de manière non exhaustive quelques défis que nous devrons surmonter au cours du développement de SylvCiT. En effet, des besoins/défis apparaîtront tout au long du processus de développement.

Tout d'abord, pour le développement futur, il est prévu d'inclure d'autres services écosystémiques tels que la réduction de la température, de la pollution ou encore la mitigation des eaux de ruissellement. Nous prévoyons notamment d'effectuer l'évaluation monétaire des services écosystémiques que nous calculerons.

De plus, l'un des défis les plus importants est de pouvoir implémenter des algorithmes de recommandation basés sur des méthodes d'apprentissage automatique. Nous souhaitons explorer une approche basée sur l'apprentissage par renforcement (Sutton et Barto, 2018) qui semble adapté à la nature du problème de recommandation que nous traitons, car cette approche permet de créer un système qui apprend constamment, qui n'a pas besoin de mises à jour périodiques et qui peut facilement s'adapter aux changements des caractéristiques des ensembles d'arbres

analysés. Cependant, pour l'entraînement de ce type d'algorithmes, il faut : (1) soit des ensembles de données importants de scénarios de recommandation de plantation en fonction d'ensemble d'arbres, (2) soit entraîner ces algorithmes au fur et à mesure de l'utilisation de l'outil afin de profiter de la rétroaction des personnes utilisatrices. C'est cette dernière option que nous privilégions.

L'amélioration de la transparence et de la contrôlabilité de l'outil au cours du processus de recommandation est un objectif important de SylvCiT. En effet, la littérature relative aux systèmes de recommandation conseille la prise en charge de niveaux d'interactions plus avancés tels que des contrôles qui définissent quelles données peuvent être suivies et prises en compte et à quelles fins. Il est aussi important d'adapter les systèmes de recommandation et leurs interfaces utilisateur à différentes caractéristiques personnelles et contextuelles.

Par ailleurs, nous pourrions intégrer beaucoup plus de paramètres lors de la recommandation, notamment les coûts de plantation et d'entretien ainsi que les risques entomologiques et pathologiques. Pour cela, il faudrait inventorier les risques auxquels sont susceptibles d'être soumises les espèces d'arbres dans la zone étudiée pour pouvoir les prendre en compte pour les recommandations de plantations.

Enfin, nous travaillons actuellement au développement d'un modèle de prédiction de l'adaptabilité d'une essence d'arbre en fonction du climat régional. Reconnaître les zones de plantation les plus favorables à long terme permettra alors un aménagement forestier beaucoup plus efficace.

CONCLUSION

Dans ce mémoire, nous nous sommes intéressés à plusieurs problèmes. Ainsi, nous avons reproduit et intégré à notre outil une méthodologie de regroupement fonctionnel permettant d'offrir un outil simple à prendre en main aux personnes utilisatrices et satisfaisant les contraintes auxquelles pourraient être soumises ces dernières, notamment en termes de contexte local ou de disponibilité d'espèces.

En outre, nous avons proposé un modèle permettant de modéliser la croissance des arbres en milieu urbain notamment à l'aide d'algorithmes d'apprentissage automatique. Une estimation plus précise des taux de croissance des espèces représente un progrès critique dans le domaine de la gestion des forêts urbaines pour deux raisons : (1) à l'heure actuelle, aucun référentiel n'existe sur les taux de croissance spécifiques aux espèces d'arbres urbains en Amérique du Nord ; toutes les valeurs disponibles sont basées sur des mesures d'arbres dans des peuplements forestiers dont la dynamique de croissance est très différente de celle des villes, (2) l'estimation des taux de croissance des espèces en milieu urbain est essentielle pour prévoir l'état et la santé de la forêt urbaine dans le futur.

Par ailleurs, nous avons intégré le calcul de services écosystémiques et divers indices biologiques permettant d'offrir une vue rapide et simple d'une forêt urbaine.

Enfin, nous proposons une recommandation d'essences d'arbres adaptées à un contexte urbain. Ces recommandations ont pour objectif de maximiser à la fois la résilience des arbres présents et les bénéfices sociaux, économiques et environnementaux qu'ils procurent à la population.

Ainsi, nous avons intégré nos différentes expérimentations dans un outil permettant de visualiser la distribution géographique des arbres, de quantifier les services rendus par chaque espèce et enfin de générer des recommandations de plantation des essences d'arbres adéquates.

Ce type de système permet d'avoir une meilleure compréhension de l'environnement qui nous entoure et a un grand potentiel pour augmenter la résilience de nos forêts urbaines. Les aménagistes et gestionnaires urbains pourront ainsi évaluer rapidement en quelques clics la santé de leur forêt urbaine et estimer en même temps sa diversité fonctionnelle, en espèces, en genres et en familles sans pour autant avoir recours à de longues analyses ou de nombreux calculs d'indices. De plus, le module de recommandation est le premier du genre visant à guider des acteurs municipaux dans l'établissement de leur plan de plantation tout en minimisant les risques associés aux changements globaux. Aussi, le projet s'inscrit pleinement dans le cadre de la « Stratégie canadienne sur la forêt urbaine (2019-2024) » (ArbresCanada, 2019) notamment le volet « techniques et technologies de planification et de gestion des forêts urbaines » définissant les objectifs pour assurer un approvisionnement durable des bienfaits écologiques, économiques et sociaux offerts par les forêts urbaines canadiennes.

Globalement, ce projet souligne les défis que nous pourrions rencontrer ainsi que les opportunités qui pourraient se présenter dans le domaine de la foresterie urbaine et de l'intelligence artificielle, notamment l'apprentissage automatique et l'apprentissage profond.

Les travaux futurs consisteront à intégrer le calcul de nouveaux services écosystémiques et leur valeur monétaire telle que la quantité de polluants filtrés dans l'air, le volume de ruissellement évité et la valeur ornementale. Aussi, plus le système est utilisé, plus nous aurons de retours d'expérience nous permettant de construire

un jeu de données composé de cas sur lesquels nous entraînerons des algorithmes d'apprentissage pour affiner les recommandations. Enfin, des modèles de distributions d'espèces basés sur l'apprentissage automatique et permettant de prédire l'adaptabilité des espèces aux changements climatiques futurs seront intégrés à notre système de recommandation afin d'améliorer ce dernier.

Il est à noter que SylvCiT a été reconnu par le Programme des Nations Unies pour l'environnement (PNUE) comme un outil clé pour l'adoption de solutions fondées sur la nature (United Nations Environment Programme, 2020). L'outil est également récipiendaire de la première édition du programme pilote « Visage municipal » du Fonds de recherche du Québec – Nature et technologies (FRQNT) lancé en avril 2020 (Visage municipal, 2021).

RÉFÉRENCES

(2018). « Cluster Analysis : Assignement and update ». https://www.slideshare.net/radiohead0401/cluster-analysis-assignment-update. Consulté le : 2021-04-12.

(2020). Balsamiq : A graphical tool to sketch out user interfaces. https://balsamiq.com/. Visité le 2020-01-17.

Adomavicius, G. et Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, *17*(6), 734–749.

Adomavicius, G. et Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender systems handbook* 217–253. Springer.

Anderson, C. (2015). Docker [software engineering]. *Ieee Software*, *32*(3), 102–c3.

Arbres Canada (2019). Stratégie canadienne sur la forêt urbaine 2019-2024.

ArbresCanada (2019). *Stratégie canadienne sur la forêt urbaine 2019-2024*. Rapport technique.

Arrondissement de Verdun (2014). *Plan arboricole - Arrondissement de Verdun*. Rapport technique, Montréal.

Arteaga-Pérez, L. E., Segura, C., Espinoza, D., Radovic, L. R. et Jiménez, R. (2015). Torrefaction of pinus radiata and eucalyptus globulus : A combined experimental and modeling approach to process synthesis. *Energy for Sustainable Development*, *29*, 13–23.

Bakalov, F., Meurs, M.-J., König-Ries, B., Sateli, B., Witte, R., Butler, G. et Tsang, A. (2013). An approach to controlling user models and personalization effects in recommender systems. Dans *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*, 49–56. ACM Press. http://dx.doi.org/10.1145/2449396.2449405

Banfield, J. D. et Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 803–821.

Beel, J., Genzmehr, M., Langer, S., Nürnberger, A. et Gipp, B. (2013). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. Dans *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation*, 7–14.

Beel, J. et Langer, S. (2015). A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. Dans *International conference on theory and practice of digital libraries*, 153–168. Springer.

Benedict, M. A., McMahon, E. T. *et al.* (2012). *Green infrastructure : linking landscapes and communities*. Island press.

Berland, A. (2020). Urban tree growth models for two nearby cities show notable differences. *Urban Ecosystems*, *23*, 1253–1261.

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 542–545.

Biau, G. et Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bodnaruk, E. W., Kroll, C. N., Yang, Y., Hirabayashi, S., Nowak, D. J. et Endreny, T. A. (2017). Where to plant urban trees ? A spatially explicit methodology to explore ecosystem service tradeoffs. *Landscape and Urban Planning*, *157*, 457–467.
http://dx.doi.org/10.1016/j.landurbplan.2016.08.016

Boettiger, C. (2015). An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79.

Bouneffouf, D., Bouzeghoub, A. et Ganarski, A. L. (2013). Risk-aware recommender systems. Dans *International Conference on Neural Information Processing*, 57–65. Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (1984). Classification and regression trees. belmont, ca : Wadsworth. *International Group*, *432*, 151–166.

Buccolieri, R., Santiago, J.-L., Rivas, E. et Sanchez, B. (2018). Review on urban tree modelling in cfd simulations : Aerodynamic, deposition and thermal effects. *Urban Forestry & Urban Greening*, *31*, 212–220.

Buckland, M. et Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, *45*(1), 12–19.

Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, *69*(Supplement 32), 175–186.

Cameron, E. et Paquette, A. (2016). *Méthodologie et guide d'utilisation – Formation créditée*. Rapport technique, Université du Québec à Montréal and Centre d'étude de la forêt.

Cariñanos, P., Casares-Porcel, M., Guardia, C. D. d. l., Aira, M. J., Belmonte, J., Boi, M., Elvira-Rendueles, B., Linares, C. D., Fernández-Rodriguez, S., Maya-Manzano, J. M., Pérez-Badía, R., Cruz, D. R.-d. l., Rodríguez-Rajo, F. J., Rojo-Úbeda, J., Romero-Zarco, C., Sánchez-Reyes, E., Sánchez-Sánchez, J., Tormo-Molina, R. et Maray, A. M. V. (2017). Assessing allergenicity in urban parks : A nature-based solution to reduce the impact on public health. *Environmental Research*, *155*, 219 – 227.

Chang, E.-C., Huang, S.-C. et Wu, H.-H. (2010). Using k-means method and spectral clustering technique in an outfitter's value analysis. *Quality & Quantity*, *44*(4), 807–815.

Chen, J. H. et Asch, S. M. (2017). Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine*, *376*(26), 2507.

Conseil Régional Environnement Montréal (2007). Le verdissement montréalais pour lutter contre les îlots de chaleur urbains, le réchauffement climatique et la pollution atmosphérique.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. et Turrin, R. (2011). Comparative evaluation of recommender system quality. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems* 1927–1932.

da Silva, C. M. S., Carneiro, A. d. C. O., Vital, B. R., Figueiró, C. G., de Freitas Fialho, L., de Magalhães, M. A., Carvalho, A. G. et Cândido, W. L. (2018). Biomass torrefaction for energy purposes–definitions and an overview of challenges and opportunities in brazil. *Renewable and Sustainable*

Energy Reviews, *82*, 2426–2432.

Delshammar, T., Östberg, J. et Öxell, C. (2015). Urban trees and ecosystem disservices–a pilot study using complaints records from three swedish cities. *Arboriculture & Urban Forestry*, *41*(4).

DESA, U. (2018). World urbanization prospects 2018. *United Nations Department for Economic and Social Affiars*.

Django (2021). Django : The web framework for perfectionists with deadlines. `https://www.djangoproject.com/`. Visité le : 2020-01-15.

Données Québec (2012). Ilots de chaleur/fraicheur urbains et température de surface 2012. `https://www.donneesquebec.ca/recherche/dataset/ilots-de-chaleur-fraicheur-urbains-et-temperature-de-surface`. Consulté le : 2021-03-19.

Dubes, R. et Jain, A. K. (1976). Clustering techniques : the user's dilemma. *Pattern Recognition*, *8*(4), 247–260.

Duchi, J., Hazan, E. et Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, *12*(Jul), 2121–2159.

Eigenbrod, F., Bell, V. A., Davies, H. N., Heinemeyer, A., Armsworth, P. R. et Gaston, K. J. (2011). The impact of projected increases in urbanization on ecosystem services. *Proceedings of the Royal Society B : Biological Sciences*, *278*(1722), 3201–3208. `http://dx.doi.org/10.1098/rspb.2010.2754`

Eirinaki, M., Gao, J., Varlamis, I. et Tserpes, K. (2018). Recommender systems for large-scale social networks : A review of challenges and solutions.

ENVI-met (2021). ENVI-met. `https://www.envi-met.com/trees-and-vegetation/`. Consulté le on : 2021-03-12.

Environnement et Changement climatique Canada (2016). *Mise à jour technique des estimations du coût social des gaz à effet de serre réalisées par Environnement et Changement climatique Canada*. Rapport technique.

Escobedo, F. J., Kroeger, T. et Wagner, J. E. (2011). Urban forests and pollution mitigation : Analyzing ecosystem services and disservices. *Environmental Pollution*, *159*(8), 2078 – 2087. `http://dx.doi.org/https://doi.org/10.1016/j.envpol.2011.01.010`

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al.* (1996). A density-based

algorithm for discovering clusters in large spatial databases with noise. Dans *Kdd*, volume 96, 226–231.

Freedman, D. A. *et al.* (1981). Bootstrapping regression models. *Annals of Statistics*, *9*(6), 1218–1228.

Ge, M., Delgado-Battenfeld, C. et Jannach, D. (2010). Beyond accuracy : evaluating recommender systems by coverage and serendipity. Dans *Proceedings of the fourth ACM conference on Recommender systems*, 257–260.

Georgi, N. J. et Zafiriadis, K. (2006). The impact of park trees on microclimate in urban areas. *Urban Ecosystems*, *9*(3), 195–209.

Goodfellow, I., Bengio, Y. et Courville, A. (2016a). *Deep learning*. MIT press.

Goodfellow, I., Bengio, Y. et Courville, A. (2016b). *Deep Learning*. MIT Press.

Govender, P. et Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution : a review (1980–2019). *Atmospheric Pollution Research*, *11*(1), 40–56.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

Haase, D., Larondelle, N., Andersson, E., Artmann, M., Borgström, S., Breuste, J., Gomez-Baggethun, E., Gren, Å., Hamstead, Z., Hansen, R. *et al.* (2014). A quantitative review of urban ecosystem service assessments : concepts, models, and implementation. *Ambio*, *43*(4), 413–433.

Han, L., Zhou, W., Li, W. et Li, L. (2014). Impact of urbanization level on urban air quality : A case of fine particles (pm2. 5) in chinese cities. *Environmental Pollution*, *194*, 163–170.

Hardin, P. J. et Jensen, R. R. (2007). The effect of urban leaf area on summertime urban surface kinetic temperatures : a terre haute case study. *Urban forestry & urban greening*, *6*(2), 63–72.

Hartigan, J. A. et Wong, M. A. (1979). Ak-means clustering algorithm. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, *28*(1), 100–108.

He, C., Parra, D. et Verbert, K. (2016). Interactive recommender systems : A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, *56*, 9–27. http://dx.doi.org/10.1016/j.eswa.2016.02.013

Heaton, J. (2008). *Introduction to neural networks with Java*. Heaton Research, Inc.

Herlocker, J. L., Konstan, J. A., Terveen, L. G. et Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53.

Hirons, A. et Sjöman, H. (2019). *Tree species selection for green infrastructure : a guide for specifiers*. Trees & Design Action Group.

Ho, T. K. (1995). Random decision forests. Dans *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Huang, L., Li, J., Zhao, D. et Zhu, J. (2008). A fieldwork study on the diurnal changes of urban microclimate in four types of ground cover and urban heat island of nanjing, china. *Building and environment*, *43*(1), 7–17.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. Dans *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, 21–34. Citeseer.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304.

i-Tree (2021). i-Tree. `https://www.itreetools.org/`. Consulté le on : 2021-04-28.

Isely, E. S., Isely, P., Seedang, S., Mulder, K., Thompson, K. et Steinman, A. D. (2010). Addressing the information gaps associated with valuing green infrastructure in west Michigan : INtegrated Valuation of Ecosystem Services Tool (INVEST). *Journal of Great Lakes Research*, *36*(3), 448 – 457. `http://dx.doi.org/https://doi.org/10.1016/j.jglr.2010.04.003`

Isinkaye, F. O., Folajimi, Y. et Ojokoh, B. A. (2015). Recommendation systems : Principles, methods and evaluation. *Egyptian Informatics Journal*, *16*(3), 261–273.

Jain, A. K. (2010). Data clustering : 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651–666.

Jannach, D., Zanker, M., Ge, M. et Gröning, M. (2012). Recommender systems in computer science and information systems–a landscape of research. Dans *International conference on electronic commerce and web technologies*, 76–87. Springer.

Jorgensen, E. (1974). *Towards an urban forestry concept.* Environment Canada.

Karimi, M., Jannach, D. et Jugovac, M. (2018). News recommender systems–survey and roads ahead. *Information Processing & Management*, *54*(6), 1203–1227.

Ketterings, Q. M., Coe, R., van Noordwijk, M., Palm, C. A. *et al.* (2001). Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests. *Forest Ecology and management*, *146*(1-3), 199–209.

Kingma, D. P. et Ba, J. (2014). Adam : A method for stochastic optimization.

Knijnenburg, B. P. et Willemsen, M. C. (2010). The effect of preference elicitation methods on the user experience of a recommender system. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* 3457–3462.

Kohavi, R. et Longbotham, R. (2017). Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, *7*(8), 922–929.

Konstan, J. A. et Riedl, J. (2012). Recommender systems : from algorithms to user experience. *User modeling and user-adapted interaction*, *22*(1), 101–123.

Lakshmi, K., Visalakshi, N. K., Shanthi, S. et Parvathavarthini, S. (2017). Clustering categorical data using k-modes based on cuckoo search optimization algorithm. *ICTACT Journal on Soft Computing*, *8*(1).

Lambert, M., Ung, C. et Raulier, F. (2005). Canadian national tree aboveground biomass equations. *Canadian Journal of Forest Research*, *35*(8), 1996–2018.

Lamothe, F., Roy, M., Racine-Hamel, S.-É., Edger, M.-A., Lefebvre, L., Njozing, B., Perron, S., Pinard, M., Tailhandier, M., Tcholakov, Y. *et al.* (2019). Enquête épidémiologique-vague de chaleur à l'été 2018 à montréal.

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. et Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, *525*(7569), 367–371.

Lemonsu, A., Viguié, V., Daniel, M. et Masson, V. (2015). Vulnerability to

heat waves : Impact of urban expansion scenarios on urban heat island and heat stress in Paris (France). *Urban Climate*, *14*, 586 – 605. http://dx.doi.org/https://doi.org/10.1016/j.uclim.2015.10.007

Li, D., Liao, W., Rigden, A. J., Liu, X., Wang, D., Malyshev, S. et Shevliakova, E. (2019). Urban heat island : Aerodynamics or imperviousness ? *Science Advances*, *5*(4), eaau4299.

Li, Z., Kurz, W. A., Apps, M. J. et Beukema, S. J. (2003). Belowground biomass dynamics in the carbon budget model of the canadian forest sector : recent improvements and implications for the estimation of npp and nep. *Canadian Journal of Forest Research*, *33*(1), 126–136.

Lin, J., Kroll, C. N., Nowak, D. J. et Greenfield, E. J. (2019). A review of urban forest modeling : Implications for management and future research. *Urban Forestry & Urban Greening*, *43*, 126366.

Liu, M., Wang, M., Wang, J. et Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification : Application to the recognition of orange beverage and chinese vinegar. *Sensors and Actuators B : Chemical*, *177*, 970–980.

Livesley, S., Baudinette, B. et Glover, D. (2014). Rainfall interception and stem flow by eucalypt street trees–the impacts of canopy density and bark type. *Urban Forestry & Urban Greening*, *13*(1), 192–197.

Loeliger, J. et McCullough, M. (2012). *Version Control with Git : Powerful tools and techniques for collaborative software development*. « O'Reilly Media, Inc. ».

Lu, J., Wu, D., Mao, M., Wang, W. et Zhang, G. (2015a). Recommender system application developments : a survey. *Decision Support Systems*, *74*, 12–32.

Lu, J., Wu, D., Mao, M., Wang, W. et Zhang, G. (2015b). Recommender system application developments : a survey. *Decision Support Systems*, *74*, 12–32.

Lucene (2021). Apache Lucene. https://lucene.apache.org/. Consulté le : 2021-04-28.

Maaten, L. v. d. et Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mason, N. W., Mouillot, D., Lee, W. G. et Wilson, J. B. (2005). Functional richness, functional evenness and functional divergence : the primary

components of functional diversity. *Oikos*, *111*(1), 112–118.

McDonald, A., Bealey, W., Fowler, D., Dragosits, U., Skiba, U., Smith, R., Donovan, R., Brett, H., Hewitt, C. et Nemitz, E. (2007). Quantifying the effect of urban tree planting on concentrations and depositions of pm10 in two uk conurbations. *Atmospheric Environment*, *41*(38), 8455–8467.

McNee, S. M., Riedl, J. et Konstan, J. A. (2006). Being accurate is not enough : how accuracy metrics have hurt recommender systems. Dans *CHI'06 extended abstracts on Human factors in computing systems*, 1097–1101.

McPherson, E. G., Muchnick, J. *et al.* (2005). Effect of street tree shade on asphalt concrete pavement performance. *Journal of Arboriculture*, *31*(6), 303.

McPherson, E. G. et Peper, P. J. (2012). Urban tree growth modeling. *Journal of Arboriculture & Urban Forestry. 38 (5) : 175-183*, *38*(5), 175–183.

McPherson, E. G., Simpson, J. R., Peper, P. J., Gardner, S. L., Vargas, K. E. et Xiao, Q. (2007). Northeast community tree guide : benefits, costs, and strategic planting. *Gen. Tech. Rep. PSW-GTR-202. Albany, CA : US Department of Agriculture, Forest Service, Pacific Southwest Research Station ; 106 p*, *202*.

McPherson, E. G., van Doorn, N. S. et Peper, P. J. (2016). Urban tree database and allometric equations. *Gen. Tech. Rep. PSW-GTR-253. Albany, CA : US Department of Agriculture, Forest Service, Pacific Southwest Research Station. 86 p.*, *253*.

Melville, P. et Sindhwani, V. (2017). Recommender systems. In *Encyclopedia of Machine Learning and Data Mining*. Routledge.

Merkel, D. (2014). Docker : lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

Mochida, A. et Lun, I. Y. (2008). Prediction of wind environment and thermal comfort at pedestrian level in urban area. *Journal of wind engineering and industrial aerodynamics*, *96*(10-11), 1498–1527.

Monteiro, M. V., Doick, K. J. et Handley, P. (2016). Allometric relationships for urban trees in great britain. *Urban Forestry & Urban Greening*, *19*, 223–236.

Murray, K. et Häubl, G. (2008). Interactive consumer decision aids, w : Handbook of marketing decision models, red. b. wierenga.

Nair, V. et Hinton, G. E. (2010). Rectified linear units improve restricted

boltzmann machines. Dans *Icml*.

Network, C. U. F. (2016). Canadian municipalities with urban forestry management plans, strategies, mandates or consideration for trees. `http://docs.wixstatic.com/ugd/64e90e_fdb8b6ce39f94cccabe290fb1d21f9ee.pdf`. Consulté le : 2020-01-05.

Nolan, K. A. et Callahan, J. E. (2006). Beachcomber biology : The shannon-weiner species diversity index. Dans *Proc. Workshop ABLE*, volume 27, 334–338.

Nowak, D. J. (1994). Atmospheric carbon dioxide reduction by chicago's urban forest. *Chicago's Urban Forest Ecosystem : Results of the Chicago Urban Forest Climate Project*, 83–94.

Nowak, D. J. et Crane, D. E. (2002). Carbon storage and sequestration by urban trees in the usa. *Environmental pollution*, *116*(3), 381–389.

Nowak, D. J., Crane, D. E., Stevens, J. C., Hoehn, R. E., Walton, J. T. et Bond, J. (2008). A ground-based method of assessing urban forest structure and ecosystem services. *Aboriculture & Urban Forestry. 34 (6) : 347-358.*, *34*(6).

Oke, T. R. (1989). The micrometeorology of the urban forest. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *324*(1223), 335–349.

Ouyang, M., Welsh, W. J. et Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, *20*(6), 917–923.

Paquette, A., d'étude de la forêt, C., sur le contrôle de la croissance de l'arbre, C. H.-Q. et du Québec a Montréal, U. (2016). *Guide stratégique pour l'augmentation de la canopée et de la résilience de la foret urbaine de la région métropolitaine de Montréal*. Rapport technique, Jour de la terre.

Parra, D., Brusilovsky, P. et Trattner, C. (2014). See what you want to see : visual user-driven approach for hybrid recommendation. Dans *Proceedings of the 19th international conference on Intelligent User Interfaces*, 235–240.

Parsa, V., Salehi, E., Yavari, A. et van Bodegom, P. (2019). An improved method for assessing mismatches between supply and demand in urban regulating ecosystem services : A case study in Tabriz, Iran. *PLoS ONE*, *14*(8). `http://dx.doi.org/10.1371/journal.pone.0220750`

Pataki, D. E., Carreiro, M. M., Cherrier, J., Grulke, N. E., Jennings, V., Pincetl, S., Pouyat, R. V., Whitlow, T. H. et Zipperer, W. C. (2011).

Coupling biogeochemical cycles in urban environments : ecosystem services, green solutions, and misconceptions. *Frontiers in Ecology and the Environment*, *9*(1), 27–36. http://dx.doi.org/10.1890/090220

Pedregosa, F. (2018). The stochastic gradient method. http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html. Consulté le : 2020-03-15.

Portugal, I., Alencar, P. et Cowan, D. (2018). The use of machine learning algorithms in recommender systems : A systematic review. *Expert Systems with Applications*, *97*, 205–227.

Pu, P., Chen, L. et Hu, R. (2011). A user-centric evaluation framework for recommender systems. Dans *Proceedings of the fifth ACM conference on Recommender systems*, 157–164.

PyGeohash (2021). pygeohash. https://pypi.org/project/pygeohash/. Consulté le : 2021-04-28.

Qu, M., Zhu, H., Liu, J., Liu, G. et Xiong, H. (2014). A cost-effective recommender system for taxi drivers. Dans *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 45–54.

React (2020). React : A javascript library for building user interfaces. https://reactjs.org/. Visité le : 2020-01-15.

Ricci, F., Rokach, L. et Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* 1–35. Springer.

Rosenberg, A. et Hirschberg, J. (2007). V-measure : A conditional entropy-based external cluster evaluation measure. Dans *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 410–420.

Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Ruiz-Benito, P., Gómez-Aparicio, L., Paquette, A., Messier, C., Kattge, J. et Zavala, M. A. (2014). Diversity increases carbon storage and tree productivity in s panish forests. *Global Ecology and Biogeography*, *23*(3), 311–322.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Safoury, L. et Salah, A. (2013). Exploiting user demographic attributes for

solving cold-start problem in recommender system. *Lecture Notes on Software Engineering*, *1*(3), 303–307.

Santamour Jr, F. S. (2004). Trees for urban planting : diversity uniformity, and common sense. *C. Elevitch, The Overstory Book : Cultivating connections with trees*, 396–399.

Santé Canada (2019). Les impacts sur la santé de la pollution de l'air au canada : Estimation de la morbidité et des décès prématurés.

Semenzato, P., Cattaneo, D. et Dainese, M. (2011). Growth prediction for five tree species in an italian urban forest. *Urban Forestry & Urban Greening*, *10*(3), 169–176.

Sharma, M. et Mann, S. (2013). A survey of recommender systems : approaches and limitations. *International Journal of Innovations in Engineering and Technology*, *2*(2), 8–14.

Sharma, S. (2017). Activation functions in neural networks. *towards data science*, *6*.

Sharma, S., Batra, N. *et al.* (2019). Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. Dans *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 568–573. IEEE.

Shi, Y., Larson, M. et Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix : A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, *47*(1), 1–45.

Sjöman, H., Hirons, A. et Bassuk, N. (2018). Improving confidence in tree species selection for challenging urban sites : a role for leaf turgor loss. *Urban Ecosystems*, *21*(6), 1171–1188.

Solr (2021). Apache Solr. `https://solr.apache.org/`. Consulté le : 2021-04-28.

Song, L., Tekin, C. et Van Der Schaar, M. (2014). Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, *9*(3), 433–445.

Song, X. P., Lai, H. R., Wijedasa, L. S., Tan, P. Y., Edwards, P. J. et Richards, D. R. (2020). Height–diameter allometry for the management of city trees in the tropics. *Environmental Research Letters*, *15*(11), 114017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R.

(2014). Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Steck, H., van Zwol, R. et Johnson, C. (2015). Interactive recommender systems : Tutorial. Dans *Proceedings of the 9th ACM Conference on Recommender Systems*, 359–360.

Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, *1*(804), 801.

Sutskever, I., Martens, J., Dahl, G. et Hinton, G. (2013). On the importance of initialization and momentum in deep learning. Dans *International conference on machine learning*, 1139–1147.

Sutton, R. S. et Barto, A. G. (2018). *Reinforcement learning : An introduction*. MIT press.

Swearingen, K. et Sinha, R. (2001). Beyond algorithms : An hci perspective on recommender systems. Dans *ACM SIGIR 2001 workshop on recommender systems*, volume 13, 1–11. Citeseer.

Taschuk, M. et Wilson, G. (2017). *Ten simple rules for making research software more robust*. Public Library of Science.

Tello, M.-L., Tomalak, M., Siwecki, R., Gáper, J., Motta, E. et Mateo-Sagasta, E. (2005). Biotic urban growing conditions—threats, pests and diseases. In *Urban forests and trees* 325–365. Springer.

Tieleman, T. et Hinton, G. (2012). Lecture 6.5-rmsprop : Divide the gradient by a running average of its recent magnitude. *COURSERA : Neural networks for machine learning*, *4*(2), 26–31.

Troxel, B., Piana, M., Ashton, M. S. et Murphy-Dunning, C. (2013). Relationships between bole and crown size for young urban trees in the northeastern usa. *Urban forestry & urban greening*, *12*(2), 144–153.

Turner-Skoff, J. B. et Cavender, N. (2019). The benefits of trees for livable and sustainable communities. *PLANTS, PEOPLE, PLANET*, *1*(4), 323–335. `http://dx.doi.org/https://doi.org/10.1002/ppp3.39`. Récupéré de `https://nph.onlinelibrary.wiley.com/doi/abs/10.1002/ppp3.39`

United Nations Environment Programme (2020). *Strengthening Actions for Nature in North America : Regional Input to UNEA-5*. Rapport technique.

Valdez, A. C., Ziefle, M. et Verbert, K. (2016). HCI for recommender systems : The past, the present and the future. Dans *Proceedings of the 10th*

ACM Conference on Recommender Systems - RecSys '16, 123–126. ACM Press. http://dx.doi.org/10.1145/2959100.2959158

Vergriete, Y. et Labrecque, M. (2007). Rôles des arbres et des plantes grimpantes en milieu urbain : Revue de littérature et tentative d'extrapolation au contexte montréalais. *Rapport d'étape destiné au Conseil Régional de l'Environnement Montréal*.

Ville de Montréal (2021a). Arbres publics sur le territoire de la Ville. https://donnees.montreal.ca/ville-de-montreal/arbres. Consulté le : 2021-03-03.

Ville de Montréal (2021b). Vue sur les indicateurs de performance. http://ville.montreal.qc.ca/vuesurlesindicateurs/index.php?kpi=2598. Consulté le : 2021-04-06.

Villegas, N. M., Sánchez, C., Díaz-Cely, J. et Tamura, G. (2018). Characterizing context-aware recommender systems : A systematic literature review. *Knowledge-Based Systems*, *140*, 173–200.

Vinícius Oliveira Castro, R., Boechat Soares, C. P., Leite, H. G., Lopes de Souza, A., Saraiva Nogueira, G. et Bolzan Martins, F. (2013). Individual growth model for eucalyptus stands in brazil using artificial neural network. *International Scholarly Research Notices*, *2013*.

Visage municipal (2021). Programme visage municipal (mun) – programme pilote. http://www.frqnt.gouv.qc.ca/bourses-et-subventions/consulter-les-programmes-remplir-une-demande/bourse/programme-visage-municipal-mun--programme-pilote-andk0sc31586435077436.

Vogt, J., Hauer, R. J., Fischer, B. C. *et al.* (2015). The costs of maintaining and not maintaining the urban forest : a review of the urban forestry and arboriculture literature. *Arboriculture & Urban Forestry*, *41*(6), 293–323.

Volkovs, M., Yu, G. W. et Poutanen, T. (2017). Dropoutnet : Addressing cold start in recommender systems. Dans *NIPS*, 4957–4966.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416.

Walker, D. (2017). *The planners guide to CommunityViz : The essential tool for a new generation of planning*. Routledge.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Wold, S., Esbensen, K. et Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Xiao, Q. et McPherson, E. G. (2016). Surface water storage capacity of twenty tree species in davis, california. *Journal of Environmental Quality*, *45*(1), 188–198.

Xu, J., Rahmatizadeh, R., Bölöni, L. et Turgut, D. (2017). A sequence learning model with recurrent neural networks for taxi demand prediction. Dans *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, 261–268. IEEE.

Xu, R. et Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645–678.

Zhang, S., Yao, L., Sun, A. et Tay, Y. (2019). Deep learning based recommender system : A survey and new perspectives. *ACM Computing Surveys (CSUR)*, *52*(1), 1–38.

Zhao, Y. et Karypis, G. (2001). Criterion functions for document clustering : Experiments and analysis.

Ziegler, C.-N., McNee, S. M., Konstan, J. A. et Lausen, G. (2005). Improving recommendation lists through topic diversification. Dans *Proceedings of the 14th international conference on World Wide Web*, 22–32.