

Université de Toulouse

MASTER 2 GEOMATIQUE

« **ScIences Géomatiques en environneMent et Aménagement** » (**SIGMA**)

<http://sigma.univ-toulouse.fr>

RAPPORT DE STAGE

Utilisation d'un modèle numérique de hauteur en stratification des données de l'Inventaire Forestier National

GEORGES Sophie



**Maître de stage : François Morneau
Tuteur-enseignant : David Sheeren**

Septembre 2017

Remerciements

Tout d'abord, un grand merci à mon maître de stage, François Morneau, et à l'IGN pour avoir proposé ce sujet de stage et pour m'avoir accueillie à Nogent et m'avoir fait confiance. Merci doublement à François pour sa patience et ses explications, et pour sa gentillesse. Je remercie également mes professeurs de la formation SIGMA pour m'avoir permis d'acquérir la formation grâce à laquelle j'ai pu effectuer ce stage.

Et puis merci tout simplement à toute l'équipe du château pour leur accueil toujours chaleureux, pour les moments festifs ensemble et une ambiance géniale. Merci aussi aux gens du club d'aïkido, IGN ou pas, pour leur accueil tout aussi agréable et dans la bonne humeur. Je souhaite de plus remercier Jérôme, Benoît et Marine pour le rattrapage algo (et, de manière générale, tout ceux qui m'ont aidée ou soutenue moralement), Pascal, Pierre-Emmanuel et Sandrine pour le vélo, et Christine pour la présentation de soutenance.

Merci à tous les apporteurs de gourmandises en salle café, raconteurs de blagues... Et à tous ceux que j'oublie.

Encore une fois, merci à tous pour ces six mois ensemble et, qui sait, à une prochaine fois ?

Table des matières

Résumé.....	3
Abstract.....	3
I. L'inventaire des ressources forestières en France.....	5
1. Enjeux.....	5
2. Bref historique de l'IFN et de ses méthodes.....	5
3. La place actuelle de l'IFN au sein de l'IGN.....	6
II. La méthode d'inventaire de l'IFN depuis 2004.....	8
1. Définitions.....	8
2. L'inventaire en continu par fenêtre glissante.....	9
3. Les observations de terrain.....	10
4. Les découpages régionaux : SER et GRECO.....	12
5. La stratification des données.....	13
5.1. Principe de la stratification.....	13
5.2. Post-stratification.....	14
5.3. Estimateurs de la variance.....	15
5.4. Stratifications appliquées à l'IFN et outil de calcul en ligne OCRE.....	15
5.5. Problématique : intérêt de stratifications basées sur un modèle de hauteur.....	17
III. Utilisation de modèles de hauteur en stratification : données et méthodes.....	17
1. Tentatives précédentes.....	17
2. Données utilisées dans le cadre du présent travail.....	18
3. Préparation et exploration des données.....	20
3.1. Calcul du MNH.....	20
3.1.1. Détermination des tuiles MNT manquantes.....	20
3.1.2. Création des tuiles MNT manquantes.....	21
3.1.3. Mise en correspondance MNE-MNT et calcul du MNH.....	22
3.2. Opérations statistiques sur le MNH et hauteur dominante.....	24
3.2.1. Statistiques descriptives sur le MNH.....	24
3.2.2. Du MNH au modèle de hauteur dominante.....	24
3.2.3. Hauteur dominante par classe.....	25
3.2.4. Exploration de la BD Forêt et choix des regroupements.....	26
4. Base de données pour la stratification.....	27
4.1. Détermination des surfaces.....	27
4.2. Entrée des données dans la base pour stratification.....	28
5. Mise en place des stratifications de test.....	28
6. Méthode de comparaison entre stratifications.....	32
IV. Résultats : comparaison entre les stratifications créées et la référence.....	33
1. Comparaisons systématiques sans ventilation.....	33
1.1. Facteurs étudiés.....	33
1.2. Effet de Z_ECO.....	35
1.3. Effet des associations REGRT x PI et HDOM_CL x PI en phase 2.....	35
1.4. Comparaison entre REGRT et HDOM_CL en phase 2.....	37
1.5. Comparaison entre phases 1.....	37
1.6. Bilan des comparaisons systématiques.....	38
2. Comparaisons avec ventilation.....	39
2.1. Par type de propriété forestière.....	39
2.2. Par essence ou groupe d'essences.....	40
2.2.1 Par essence principale.....	40

2.2.2. Par cinq classes de composition de peuplement.....	43
2.2.3. Par opposition feuillus / conifères.....	44
2.3. Par classes d'exploitabilité.....	45
2.4. Par classe de diamètre.....	46
V. Discussion.....	47
1. Limites des données disponibles et de leur traitement.....	47
2. Stratifications utilisées.....	48
3. Interprétation des résultats.....	49
VI. Conclusion générale.....	49
Bibliographie.....	51
Annexes.....	I
1. Calcul des composantes de la variance.....	I
2. Cartes.....	II
Classes de hauteur définies en Alsace à partir du MNH forêt complet.....	II
Essences et regroupements d'essences définis en Alsace à partir de BD Forêt V2.....	III
3. Rapports Rmarkdown.....	IV
Processus de création des tuiles MNH.....	IV
Statistiques sur les MNH et constitution du modèle par classes de hauteur.....	XX
Comparaisons entre stratifications.....	XXVI
4. Diagramme de Gantt.....	XXXIV

Résumé

L'Inventaire Forestier National fournit des données de références concernant la forêt française depuis 1960. La méthode actuellement utilisée repose sur des points d'échantillonnage tirés au sort sur tout le territoire et photo-interprétés, puis pour certains visités, et sur un traitement des données par post-stratification. Celle-ci a jusqu'à présent toujours utilisé les mêmes informations, mais pourrait exploiter d'autres données disponibles, en particulier des modèles de hauteur. À partir de modèles numériques de terrain et d'élévation disponibles pour la région Alsace, un modèle numérique de « hauteur dominante » en huit classes de hauteur a été élaboré pour les forêts alsaciennes. La méthode mise en place vise à une automatisation maximale de ce travail grâce au logiciel R.

Ce modèle, ainsi qu'une cartographie simplifiée des essences, ont servi de base à la mise en place de nouvelles stratifications des données sur la période 2010-2014. Celles-ci ont été comparées à la stratification de référence de l'inventaire sans ventilation des données ; puis les stratifications les plus intéressantes, qui se sont avérées être les plus simples, ont été comparées en ventilant les données par différents critères. La stratification par essences regroupées a amélioré les estimations pour les classes de peuplement, l'opposition feuillus / conifères et des essences fréquentes. La stratification par classes de hauteur est plus intéressante pour les classes de diamètre des arbres et la difficulté d'exploitation. Elles n'améliorent par contre pas les résultats par type de propriété forestière ou pour d'autres essences.

Abstract

The French National Forest Inventory has been providing reference data on French forests since 1960. The current method is based on sampling points, randomly selected throughout the territory and subjected to photo-interpretation, then visited for some of them, plus data processing using post-stratification. The latter technique has always used the same type of information so far, although other are available that could be exploited, particularly height estimates. From the digital terrain and elevation models available on the Alsace region, a model of “dominant tree height” using eight height classes was produced for Alsatian forests. The method developed aimed at maximal automation of this process through use of the R software.

This model, as well as a simplified tree species map, were used as bases for new data stratifications for the years 2010-2014. These stratifications were compared to the inventory's reference stratification with no data breakdown; then the best ones, which turned out to be the simplest, were compared to the reference with different breakdown criteria. The stratification using species types improved estimations when breaking data by forest composition, when opposing deciduous trees to conifers, and with the most common tree species. The stratification using height classes gave the best results when breaking data by trees diameter or by easiness of exploitation. On the other hand, neither improved estimations when breaking data by forest-ownership types or on some species.

I. L'inventaire des ressources forestières en France

1. Enjeux

Les volumes de bois sur pied en Europe n'ont cessé d'augmenter ces dernières décennies. La surface de la forêt française a ainsi progressé au point que, de 8 millions d'hectares en 1850, elle est passée à 16 millions en 2014, phénomène encore en cours dans le sud du Massif Central et en Bretagne. De 1,5 milliards de m³ en 1973, les volumes sont passés à plus de 2,5 milliards actuellement ; environ 40 millions de m³ supplémentaires par an ces dix dernières années, contre 30 dans les années 1990 et 20 dans les années 1960-1970. Et ce, malgré trois tempêtes récentes aux dégâts sans précédent (Lothar et Martin en 1999, Klaus en 2009). Par ailleurs, les récoltes ont progressé, puis stagné à partir de 2000 [3].

On constate également de fortes évolutions dans la structure et la composition des forêts : une prédominance accrue des feuillus et un déplacement du volume depuis les régions de production historiques, Nord-Est et Sud-Ouest (fortement touchées par les tempêtes), vers les régions du centre. Les petites forêts privées morcelées ont connu un fort accroissement, souvent sur d'anciennes parcelles agricoles : la moitié de leur surface actuelle n'était pas forestière il y a 150 ans [3]. Dans les années à venir, les conséquences du changement climatique (sécheresses, incendies, nouveaux ravageurs...) contribueront aussi à en changer rapidement l'abondance et la composition.

Parallèlement, la forêt durablement gérée est souvent citée parmi les moyens de lutte contre ce même changement climatique de par sa capacité à stocker le carbone et à produire des matériaux au cycle de vie moins émetteur de CO₂ (bois énergie par rapport aux carburants fossiles par exemple). La France a ainsi, dans le cadre du protocole de Kyoto, pris des engagements pour une gestion durable des forêts, suivant les critères et indicateurs pour l'Europe (adoptés à Lisbonne en 1998 et améliorés à Vienne en 2003). Ces indicateurs existent en France depuis 1995 et sont revus tous les cinq ans. Ils permettent l'évaluation des ressources forestières mondiale demandée par la FAO (Food and Agriculture Organisation).

Le Grenelle de l'Environnement et les Assises de la forêt en 2007 ont donc fixé l'objectif de récolter davantage de bois. Depuis la loi d'avenir sur l'agriculture et la forêt en 2014, cet objectif est mis en œuvre dans le cadre du plan national forêt bois, qui fixe des lignes directrices pour une durée de 10 ans et décrit les orientations en matières de sylviculture et de bois. De telles décisions supposent notamment de connaître les volumes avec précision pour s'assurer que les prélèvements ainsi encouragés sont durables. Une connaissance fine de la ressource bois disponible est donc plus nécessaire que jamais.

Depuis plus de cinquante ans, c'est l'Inventaire Forestier National (IFN), désormais rattaché à l'Institut National de l'Information Géographique et Forestière (IGN), qui fournit les données les plus détaillées sur les forêts de France métropolitaine. Les méthodes d'inventaire évoluent en permanence afin de s'adapter à une demande d'informations toujours plus détaillées et accessibles.

2. Bref historique de l'IFN et de ses méthodes

L'Inventaire Forestier National français a été mis en place à partir de 1960, suite à une ordonnance de 1958 du général de Gaulle, et est devenu, en 1994, un établissement public à caractère administratif, sous tutelle du ministère chargé des forêts. Sa mission est d'« assurer l'inventaire permanent des ressources forestières nationales, indépendamment de toute question de propriété »,

ce qui est alors fait département par département. Ses moyens ayant été mis en place très progressivement, le premier cycle d'inventaire s'est étalé sur les décennies 1960 et 1970, la Gironde ayant été le premier département inventorié (1960-1961). Cet inventaire systématique n'a en revanche pas été étendu en dehors de la métropole jusqu'à présent [3].

La méthode alors utilisée (actuellement dite « ancienne méthode »), qui comportait dès le début une phase de photo-interprétation et une phase de mesures et vérifications de terrain, est restée largement la même pour les deux cycles suivants (années 1980 et 1990). Le protocole pour les mesures de terrain, avec des placettes concentriques autour du point d'inventaire, a été stabilisé vers 1963-1964 [3]. À partir de 1994 sont ajoutées des observations écologiques et floristiques (exposition, pente, relevé floristique, description du sol, etc.). Depuis 2010, l'IFN observe aussi les prélèvements en forêt, ce qui permet de calculer un bilan complet des flux de bois (par essence, diamètre, propriété, accessibilité, etc.). [7]

L'ancienne méthode donnait des résultats précis sur le moment, mais la durée importante qui séparait deux passages dans le même département rendait difficile ou impossible la mesure des évolutions. De plus, les résultats au niveau d'une région n'avaient réellement de sens que s'il y avait eu peu de temps ou d'évolutions au cours des années nécessaires pour couvrir tous les départements concernés. Des biais dans les résultats sont donc apparus dès le deuxième cycle d'inventaire. Mais ce sont les tempêtes de 1999 qui ont imposé le passage à un système plus adapté pour gérer de telles crises.

En 2004, l'IFN a donc changé de méthode pour adopter un système d'inventaire en continu (la « nouvelle méthode »), qui était en train de devenir la norme internationale. Chaque année, le tirage des points d'inventaire concerne tout le territoire, suivant une grille qui permet de tirer un point pour 10 km². Ces échantillons annuels sont complémentaires, si bien qu'au bout de dix ans un point a été tiré par maille de 1 km². Ce système a montré son intérêt suite à la tempête Klaus en 2009 : comme quatre campagnes nationales nouvelle méthode avaient déjà eu lieu, un nombre suffisant de points vieux de seulement un à quatre ans était disponible partout. Les dégâts dans le massif des Landes de Gascogne ont ainsi pu être estimés en quelques semaines [3].

Le 1^{er} janvier 2012, l'IFN a fusionné avec l'IGN pour former l'Institut National de l'Information Géographique et Forestière. Ce nouvel établissement, dont le sigle reste IGN, se trouve désormais sous une double tutelle : ministère de la Transition écologique et solidaire (ancien ministère de l'Écologie et du Développement durable, auquel était rattaché l'ancien IGN) et ministère de l'Agriculture et de l'Alimentation (qui comprend la forêt). La mission de validation des méthodes de l'Inventaire a été transférée du Comité Scientifique et Technique de l'IFN à une section spécialisée de celui de l'IGN, mise en place fin 2015. Depuis 2012, un Comité de la filière Forêt-Bois a également été institué pour permettre aux représentants des organismes utilisateurs de dialoguer avec l'IGN. Le nouvel établissement a conservé les missions des deux entités précédentes. L'inventaire forestier national est inclus dans le périmètre de la démarche qualité de l'IGN et dispose de la certification ISO 9001 [5].

3. La place actuelle de l'IFN au sein de l'IGN

Le service de l'inventaire forestier et environnemental de l'IGN (SIFE), qui a repris les missions de mise en place de l'inventaire forestier et de publication des résultats, est situé à Nogent-sur-Vernisson (Loiret). Il est chargé de la définition des méthodes, du suivi général des travaux d'inventaire, du traitement et de l'analyse des données et de leur diffusion. Nogent abrite également

un service de la direction informatique (le DIFE) dédié aux développements pour le SIFE. Les opérations de terrain dans le cadre de l'inventaire sont réalisées par les divisions « inventaire forestier » des différentes directions interrégionales. Celles-ci sont situées à Bordeaux, Caen, Lyon, Aix et Nancy (Figure 1). Le sigle « IFN » désigne encore l'enquête elle-même.

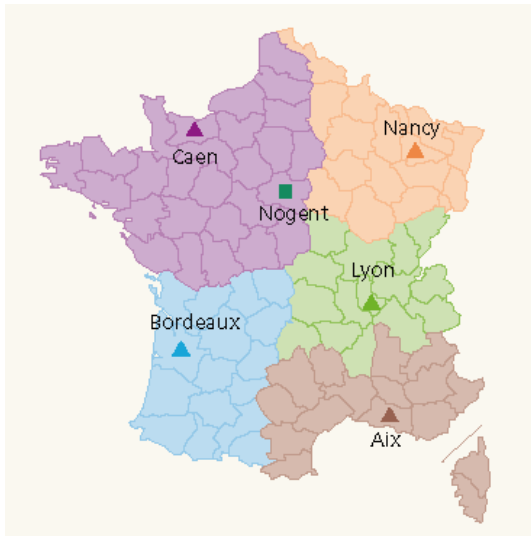


Figure 1 : Directions interrégionales de l'IGN et le siège du SIFE à Nogent-sur-Vernisson. Les divisions varient légèrement en fonction des campagnes d'inventaire et de la charge de travail de chaque direction interrégionale.

Au sein de l'IGN, le SIFE est rattaché à la direction de la production des référentiels. Un laboratoire de recherche dédié à l'inventaire, le LIF, a de plus été créé à Nancy (Figure 2).

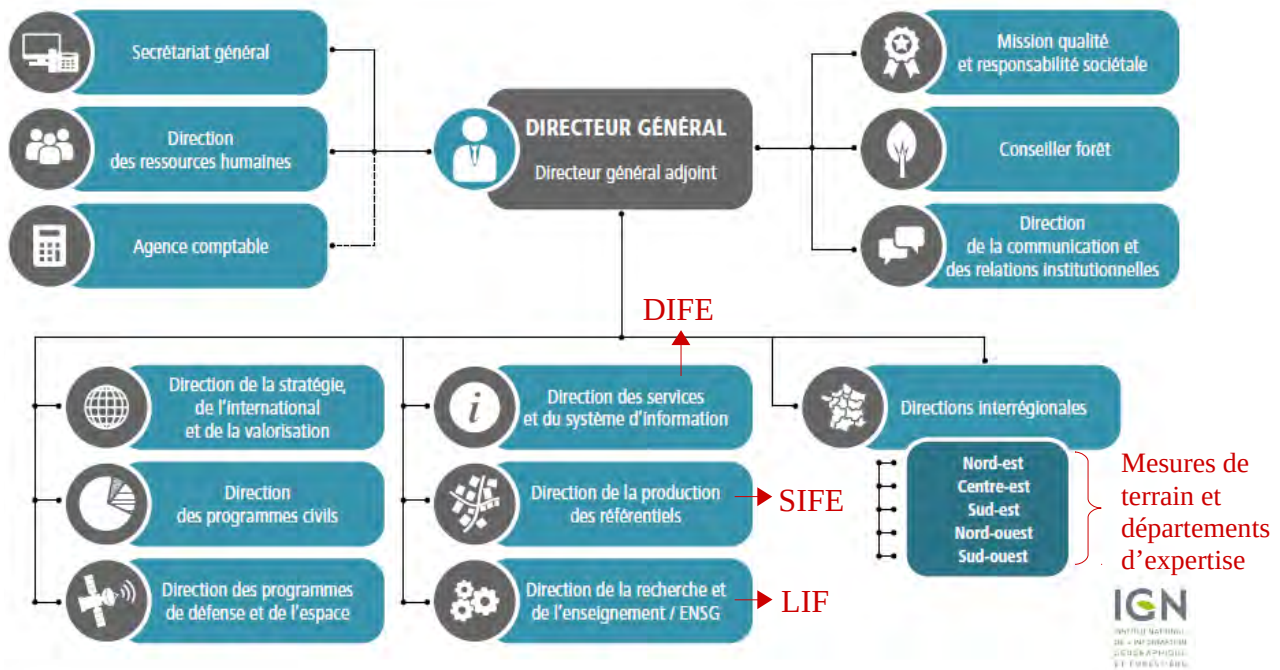


Figure 2 : Organigramme général de l'IGN. En rouge, les principaux services intervenant dans le cadre de l'inventaire.

L'inventaire publie ses résultats sous la forme :

- d'un mémento qui reprend l'ensemble des principaux chiffres ;

- d'un outil dynamique sur internet (cf. plus bas) qui présente un ensemble de tableaux standards ainsi que des tableaux réalisables à la carte ;
- de tableaux standard sous la forme de fichiers .pdf ;
- périodiquement par des publications comprenant à la fois des données et des commentaires.

Les résultats sont donnés au niveau national, par région administrative, et par région écologique (voir II 4. pour la définition de celles-ci). Les données brutes sont également mises à disposition gratuitement en ligne. Pour des raisons de confidentialité, les coordonnées des points d'inventaire ne sont fournies qu'au kilomètre près. Les propriétaires des sites enquêtés sur le terrain peuvent seuls demander les résultats concernant le ou les points concernés [5].

L'inventaire propose aussi sur convention de répondre à des demandes spécifiques grâce à un protocole enrichi. Par exemple :

- qualité du douglas en Bourgogne selon un nombre de catégories de qualité supérieures au standard de l'inventaire ;
- quantification des gélivures du chêne ;
- hauteur d'élagage sur le douglas ;
- appréciation des dégâts causés par l'exploitation.

Il réalise aussi depuis une vingtaine d'années des « études de ressource » pour des commanditaires variés et avec des partenaires de la filière. En général, il s'agit d'estimer les volumes de bois d'œuvre, d'industrie et énergie disponibles à moyen ou long terme dans une région, pour évaluer la capacité de production pour une essence ou une zone donnée, l'impact d'une politique, ou encore les possibilités de développement d'une industrie de transformation. Les estimations des prélèvements par l'IFN ont notamment servi à l'étude nationale « Disponibilités forestières pour l'énergie et les matériaux à l'horizon 2035 ». L'IFN a aussi mis au point des techniques pour détecter les changements d'occupation du sol dans les Départements d'Outre-Mer, une information qui permet à la France de tenir ses engagements en termes de fourniture d'informations pour le protocole de Kyoto. [7]

L'IGN est également membre de l'association ENFIN (European National Forest Inventory Network), qui regroupe les organismes responsables de la réalisation des inventaires forestiers nationaux en Europe. Cette association, qui cherche à harmoniser les données produites par les différents pays, permet notamment de travailler à la coordination et à l'harmonisation de ces inventaires, et a lancé des projets de recherche transnationaux correspondants : les actions COST E43 et USEWOOD, et le projet Horizon 2020 DIABOLO [5]. L'IGN a aussi établi des échanges bilatéraux avec différents pays. Enfin, les données écologiques de l'inventaire forestier national ont été mobilisées pour de très nombreux articles scientifiques [7].

II. La méthode d'inventaire de l'IFN depuis 2004

1. Définitions

Hors surfaces, les estimations publiées par l'IFN concernent uniquement les superficies affectées en forêt dite « de production ». L'inventaire suit la définition internationale de la forêt, issue de la FAO : la forêt est définie comme un territoire occupant une superficie d'au moins 50 ares, avec des arbres capables d'atteindre une hauteur supérieure à cinq mètres à maturité in situ, un couvert arboré de plus de 10 % et une largeur d'au moins 20 mètres. Une forêt « de production » est une forêt disponible pour la production de bois, c'est-à-dire où l'exploitation du bois est possible (sans

considération de rentabilité économique) et compatible avec d'éventuelles autres fonctions. Les peupleraies (taux de couvert des peupliers cultivés supérieur à 75 %) sont classées parmi les forêts de production [7]. Elles font de plus l'objet de mesures spécifiques.

Le volume de bois sur pied au sens de l'IFN correspond au volume de la tige arrêté à la découpe bois fort (7 cm de diamètre). Cela signifie que le volume des branches et de la cime n'est pas inclus dans les estimations. De plus, seuls sont inventoriés les arbres d'au moins 7,5 cm de diamètre à la hauteur de référence (Figure 3) [7].

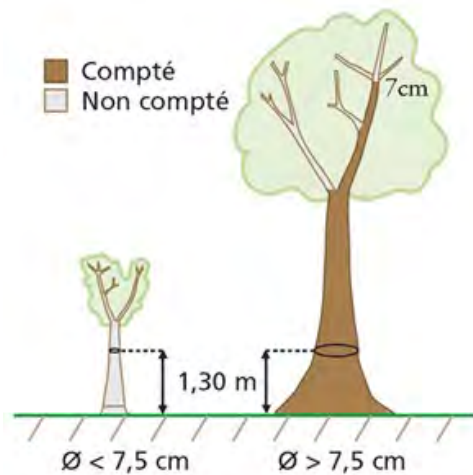


Figure 3 : Notion d'arbre recensable et volume de bois comptabilisé dans le cadre des inventaires.

2. L'inventaire en continu par fenêtre glissante

Le dispositif de l'inventaire est basé depuis ses débuts sur deux phases statistiques : une première phase avec un échantillon systématique et photo-interprété (sur orthophotos IGN en couleurs naturelles ou infrarouge), puis une deuxième phase de visites de terrain sur un sous-échantillon des points de la première phase, selon un plan défini en fonction des résultats de la deuxième phase. Un contrôle qualité est réalisé aux deux étapes : 4 % des points sont photo-interprétés une seconde fois, et 4 % des points de terrain sont vérifiés par deux équipes de deux contrôleurs nationaux [5].

Actuellement, environ 80 000 points sont utilisés en première phase, la BD Ortho® infrarouge de l'IGN servant de base à la photo-interprétation [3]. L'observation porte sur la couverture et l'utilisation du sol au niveau d'une placette de 25 m de rayon centrée sur le point d'inventaire (en cas d'hétérogénéité, seule la partie où est tombé le point est décrite) [5]. Lors de la stratification des données par la suite, ces points photo-interprétés constituent l'échantillon dit de phase 1.

Depuis l'adoption de l'inventaire en continu, ces points sont définis de manière à être représentatifs de l'ensemble du territoire métropolitain, et sont répartis de manière à ce que les échantillons de cinq années successives se complètent. Les résultats publiés concernant une année donnée (par exemple 2014) sont des estimations qui utilisent les résultats de campagnes correspondant à une fenêtre de cinq années centrée sur celle-ci (dans ce cas 2012-2016). Une campagne annuelle dure de novembre année $n-1$ à octobre année n pour la phase terrain, les résultats étant exploités et publiés l'année $n+1$. Ainsi, les résultats complets pour 2014 (utilisant les données de 2012-2016) seront publiés en 2017. Ce « lissage » est sans biais pour l'année centrale de la fenêtre, pour toute statistique à évolution continue et approximativement linéaire sur la largeur de la fenêtre [5].

Pour assurer une bonne répartition des points d'inventaire, le territoire (toutes occupations du sol confondues) est maillé avec une grille carrée de 1 km de côté. D'une année sur l'autre, les points sont tirés dans des mailles de 1 km² différentes, de manière à avoir un point pour 10 km². Au bout de 10 campagnes, toutes les mailles de 1 km² ont été échantillonnées une fois et une seule. De plus, chaque point est revisité cinq ans après (le point de l'année en cours est alors dans la maille adjacente) [3] (Figure 4).

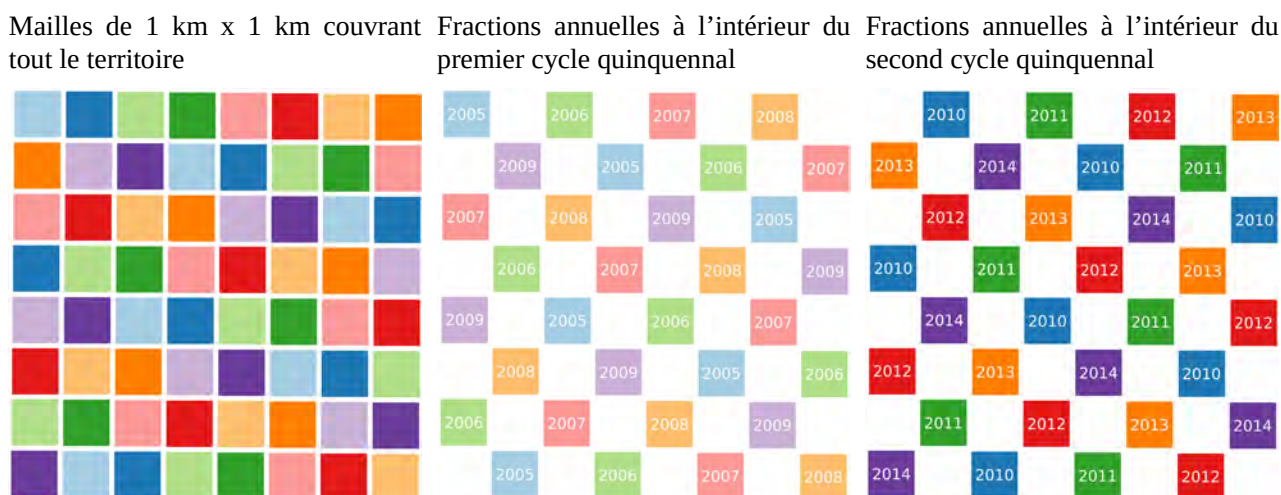


Figure 4 : Maillage complet du territoire par deux cycles quinquennaux (une couleur par fraction annuelle).

Retourner sur les points permet de relever rapidement des évolutions. La campagne 2010 est la première où le retour sur les points précédents a pu être testé, et a fourni des estimations du volume de bois prélevé parfaitement cohérentes avec les chiffres issus de méthodes différentes.

En cas de ruptures fortes sur la période, comme suite à la tempête Klaus, le lissage des résultats par fenêtre de cinq ans n'est plus possible. Il faut alors actualiser les informations post-changement, ce qui est rendu possible par le retour sur les points d'inventaire [3]. La télédétection a également permis de cartographier en quelques semaines les dégâts de tempête sur le massif landais en 2000 et en 2009 [7].

3. Les observations de terrain

Autour de chaque point d'inventaire est définie une placette de 25 m de rayon sur laquelle est déterminée l'occupation du sol, complétée par un transect aléatoire de 1 km le long duquel sont notés les haies et cordons boisés qui le coupent. Des placettes concentriques plus petites sont de plus définies autour du point pour inventorier différentes catégories d'arbres (6 m de rayon : tous les arbres recensables ; 9 m de rayon : arbres d'au moins 22,5 cm de diamètre ; 15 m de rayon : arbres d'au moins 37,5 cm de diamètre) ou pour certaines autres observations (les relevés floristiques ne sont effectués que sur la placette de 15 m de rayon). Certaines mesures comme les observations pédologiques sont seulement effectuées sur le point lui-même ou le plus près possible (Figure 5).

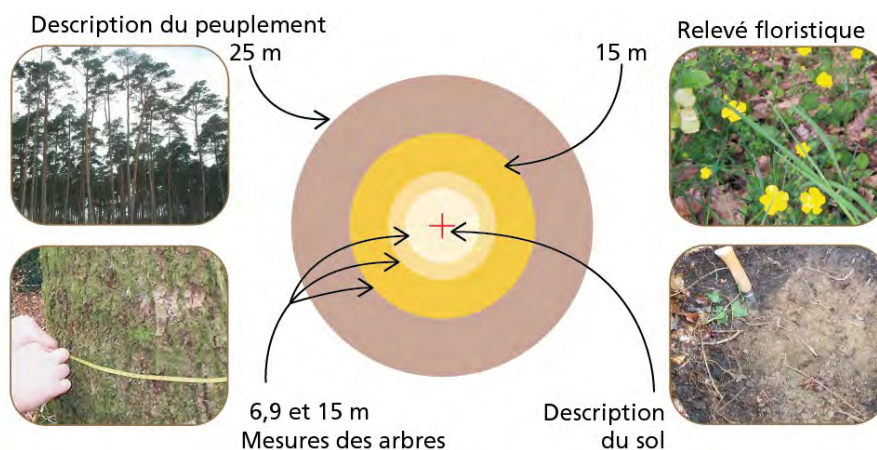


Figure 5 : Principales observations de terrain.

Le résultat de la photo-interprétation (phase 1) est considéré comme suffisant lorsque l'occupation du sol ne correspond pas à une zone potentiellement boisée (eau, zone urbaine, champ...). Tous les points correspondants sont alors conservés tels quels dans l'échantillon de phase 2. Ailleurs, le taux d'échantillonnage pour la deuxième phase varie en fonction des résultats de la première phase : un point de la première phase sur deux en forêt, peupleraie ou bosquet, un sur quatre pour les landes et un sur huit pour les formations linéaires. Ce taux est encore divisé par deux dans les forêts très homogènes ou de peu d'intérêt productif (garrigues et maquis) et les zones de montagne (Figure 6).

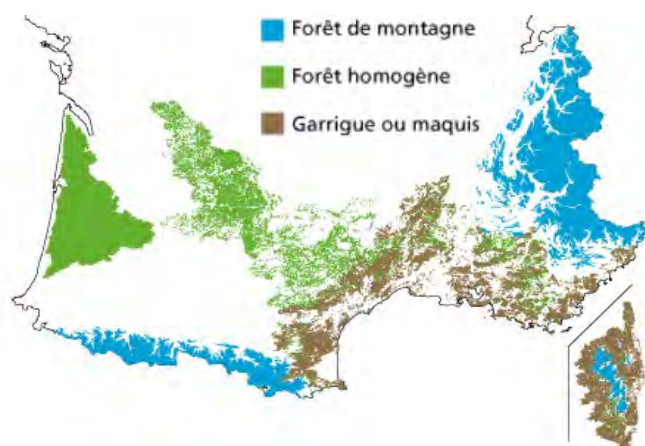
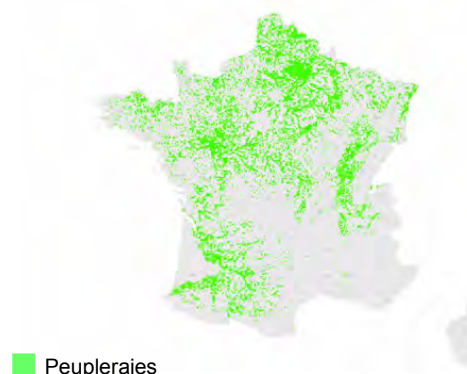


Figure 6 : Zonages utilisés pour adapter la densité d'échantillonnage.



Dans les zones de montagne, un point supplémentaire est défini à partir des coordonnées du point principal, ramenant la densité d'échantillonnage à l'équivalent d'un point par maille (mais réduisant les déplacements). En zone populicole (forte probabilité de peupleraies, donc fort enjeu productif, figure 6), on fait de même et on rajoute deux autres points (Figure 7) [5].

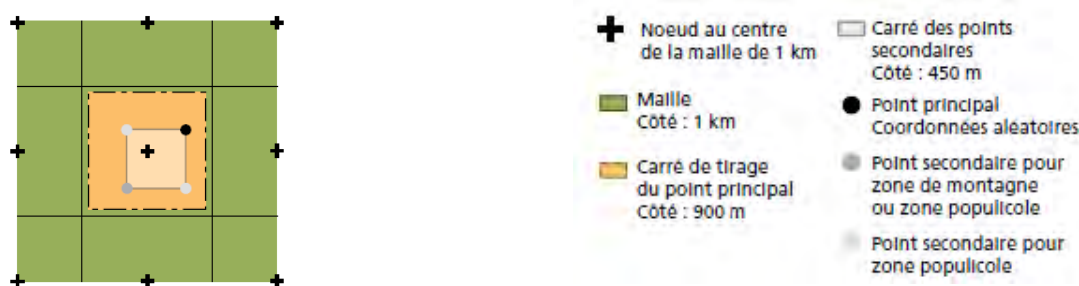


Figure 7 : Tirage des points supplémentaires en zones de montagne ou populicole.

Au final, environ 8 000 points sont inventoriés sur le terrain chaque campagne, dont 7 000 en forêt disponible pour la production, peupleraies comprises. Des mesures de terrain supplémentaires se sont progressivement rajoutées pour une meilleure appréhension des écosystèmes, dont le volume de bois mort sur pied et au sol depuis 2008 [3].

4. Les découpages régionaux : SER et GRECO

L'IFN a découpé le territoire métropolitain suivant des critères écologiques et forestiers. Historiquement, il s'agissait de 309 régions forestières, considérées suffisamment homogènes sur le plan pédologique et climatique pour que les forêts naturelles y soient de type comparable. Du fait de leur nombre, le catalogues des stations et le guide de choix des essences manquaient encore pour plus d'un tiers d'entre elles en 2000. Depuis 2011, elles ont donc été remplacées par un découpage plus simple, utilisant 91 sylvoécorégions (SER) : celles-ci ont été obtenues pour l'essentiel par regroupement des anciennes régions, mais pour certaines les limites climatiques ou pédologiques ont été mises à jour [1] [3].

Les SER sont de plus regroupées en douze grandes régions écologiques (GRECO) : une qui regroupe les cinq SER correspondant aux alluvions récentes des grands bassins fluviaux, et onze qui rassemblent les autres SER en grandes zones géographiques (Grand Ouest cristallin et océanique, Centre Nord semi-océanique, Grand Est semi-continental, Vosges, Jura, Sud-ouest océanique, Massif central, Alpes, Pyrénées, Méditerranée, Corse). Ces découpages simplifiés sont plus faciles à mettre en œuvre pour la publication de données. Les nouvelles divisions sont aussi plus adaptées pour publier des données à l'échelle nationale, ou pour des thématiques comme le suivi du changement climatique (Figure 8) [1] [3].

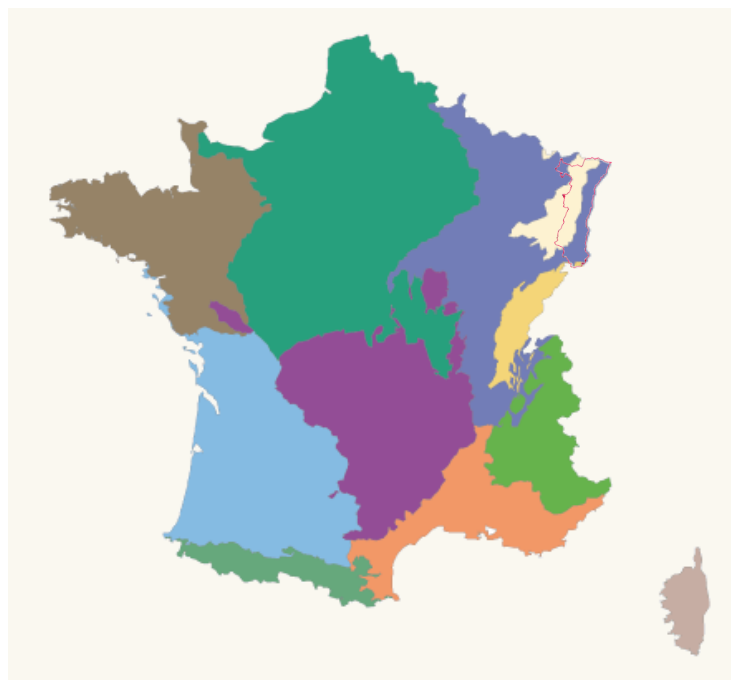


Figure 8 : Les onze GRECO non alluviales. En rouge, l'emprise de la région Alsace.

5. La stratification des données

5.1. Principe de la stratification

Dans la méthode d'échantillonnage de l'IFN décrite précédemment, on a vu que les points étaient tirés au sort selon une méthodologie qui assure chaque année une couverture représentative de l'ensemble du territoire. On obtient ainsi des estimations des grandeurs que l'on cherche à connaître (surface, volume...) sans avoir à effectuer de mesures exhaustives. Cependant, on peut s'approcher encore davantage des valeurs « réelles » grâce aux méthodes de stratification, qui permettent de réduire la variance sur les grandeurs étudiées.

Dans le cas de données géographiques, stratifier correspond à une segmentation du territoire d'intérêt telle qu'elle le couvre entièrement sans recouvrement entre les segments [1]. Les strates doivent de plus comporter au moins un point d'échantillonnage ; de plus, tout point doit appartenir à une strate et une seule (de surface non nulle). Si les surfaces des strates sont non connues mais estimées, il s'agit d'un plan à plusieurs phases (éventuellement elles-mêmes stratifiées) [3].

L'inventaire utilise normalement deux phases de stratification (Figure 9). On a donc au final un échantillon à deux phases à probabilité inégale (à cause du taux d'échantillonnage variable) mais à répartition systématique.

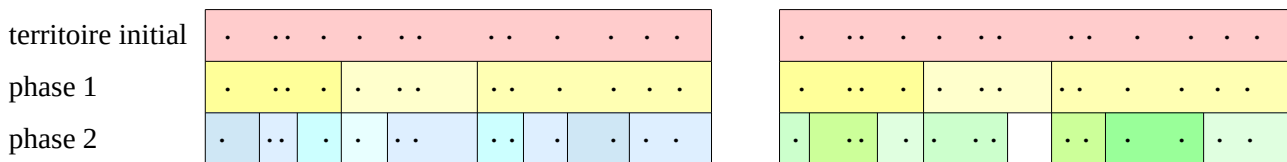


Figure 9 : Exemples de stratifications à deux phases avec les mêmes données initiales et la même phase 1, mais une phase 2 différente. Celle de droite n'est pas valide en l'état (une strate vide).

Pour illustrer comment une stratification permet de réduire la variance sur la surface forestière calculée, prenons la stratification la plus simple possible : par photo-interprétation, on classe des points répartis de manière homogène en « forêt » ou « non-forêt ». La BD Forêt fournit également une information « forêt » ou « non-forêt ». On obtient les surfaces suivantes pour les strates :

$$S_{f_pi} = S_{f_carto} \times P(pi_{f_f}) + S_{nf_carto} \times P(pi_{f_nf}) \quad \text{et de même}$$

$$S_{nf_pi} = S_{f_carto} \times P(pi_{nf_f}) + S_{nf_carto} \times P(pi_{nf_nf})$$

où S_{f_carto} et S_{nf_carto} sont respectivement les surfaces en forêt et hors forêt, d'après la BD Forêt. Comme ces surfaces sont directement tirées de données dont on fait l'hypothèse qu'elles sont connues sans erreur, la variance sur leur valeur est nulle.

$P(pi_{f_f})$ et $P(pi_{f_nf})$ sont les proportions de points, parmi ceux désignés comme forêt par la photo-interprétation, qui sont respectivement en forêt et hors forêt d'après la BD Forêt : plus $P(pi_{f_f})$ est proche de 1 (si elle vaut 1, la photo-interprétation confirme toujours la BD Forêt), plus la variance sur S_{f_pi} est faible. En pratique, $P(pi_{f_f})$ est plutôt de l'ordre de 0,9, du fait notamment des changements qui ont pu se produire entre la dernière mise à jour de la BD Forêt et la prise de vues ayant servi à la photo-interprétation. Le raisonnement est le même pour S_{nf_pi} avec les points désignés hors forêt par la photo-interprétation. La variance pour S_{f_pi} et S_{nf_pi} dépend au final uniquement de $P(pi_{f_f})$ et $P(pi_{f_nf})$ (Figure 10).

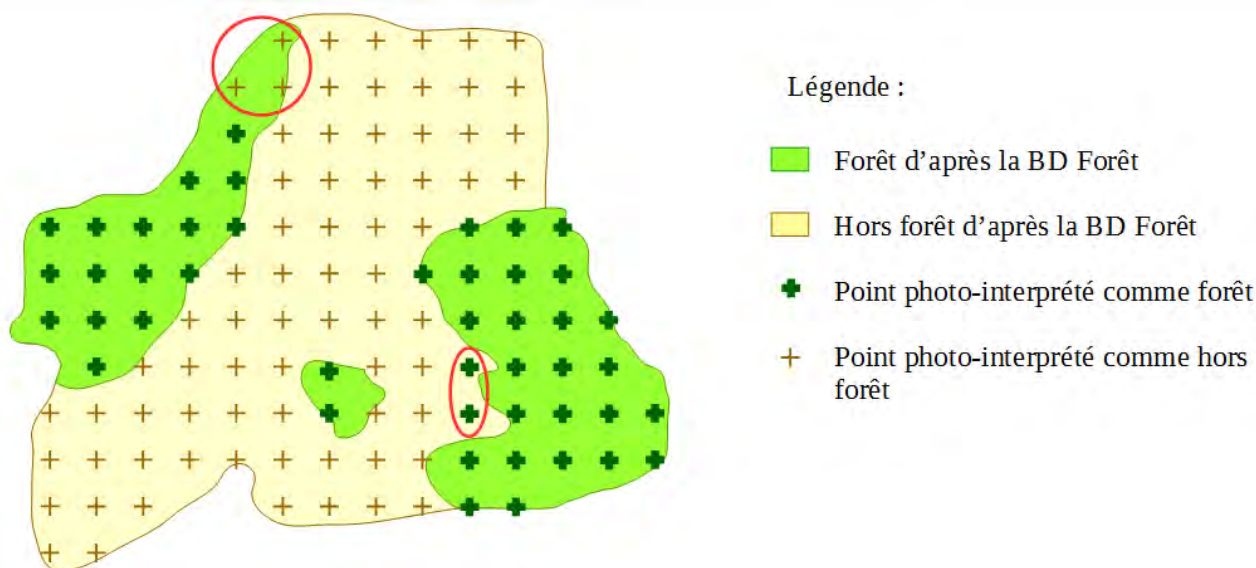


Figure 10 : Exemple de stratification avec deux strates seulement. Ici, $P(\text{pi}_f) = 0,96$ (43 des 45 points reconnus comme en forêt par la photo-interprétation y sont bien d'après la BD Forêt) et $P(\text{pi}_{nf}) = 0,04$ (deux points ont été photo-interprétés comme forêt en désaccord avec la BD Forêt).

La surface de forêt totale se calcule donc comme suit :

$$S_f = S_{f_pi} \times P(f \text{ dans } f_pi) + S_{nf_pi} \times P(f \text{ dans } nf_pi)$$

où $P(f \text{ dans } f_pi)$ est la proportion de la forêt qui se trouve dans S_{f_pi} (proche de 1). Comme on a minimisé la variance sur S_{f_pi} et S_{nf_pi} en faisant appel à des surfaces déjà connues, cette décomposition aboutit à une variance sur S_f plus faible que sans stratification. Les surfaces de strates qui ne sont pas déjà connues sont estimées grâce au nombre de points dans la strate, sachant que la strate n'est pas forcément continue.

Les stratifications utilisées pour l'inventaire utilisent le même principe avec simplement plus de strates que la seule distinction forêt/non-forêt. De plus, alors que dans l'exemple simplifié tous les points ont le même poids, dans la pratique les points de l'inventaire ont des poids variables déduits du plan d'échantillonnage pour tenir compte de la probabilité de tirage (dont on a vu qu'elle est modifiée dans certaines zones). L'effectif équivalent d'équiprobables noté $n' = (\sum w_i)^2 / \sum w_i^2$ (où w_i est le poids du point i) joue alors le rôle de l'effectif de l'échantillon pour des tirages équiprobables [5].

5.2. Post-stratification

La plupart des utilisateurs de stratifications stratifient l'échantillon en amont, grâce aux connaissances dont ils disposent en amont sur la composition de la population (avec, par exemple, un niveau de sondage plus élevé dans des strates que l'on sait hétérogènes) [1]. L'ancienne méthode d'inventaire en faisait usage, mais cela signifie que l'on contraint le plan d'échantillonnage en amont. Il devient alors impossible d'optimiser la stratification par rapport à une variable particulière a posteriori : si par exemple un plan d'échantillonnage construit pour estimer les surfaces s'avère inadapté à l'estimation des volumes (précision insuffisante sur ceux-ci), il n'est plus possible de le modifier.

Il a donc été décidé, pour la nouvelle méthode, d'effectuer une **post**-stratification sur les données déjà disponibles. Outre les possibilités d'optimisation, la post-stratification présente dans le cas de l'inventaire moins de contraintes organisationnelles. De plus, il devient possible d'exploiter en stratification des données non encore disponibles lors de la campagne concernée ; par exemple, la cartographie des types forestiers mise à jour (BD Forêt V2), qui n'a encore été réalisée que pour 77 départements. Une fois disponible, elle peut être utilisée pour stratifier des données récoltées avant sa mise en place.

5.3. Estimateurs de la variance

Les résultats que l'on cherche à obtenir au final seront ventilés par sous-domaines ; une essence par exemple est un sous-domaine. L'échantillon est supposé complet : pour tout point de l'échantillon, on connaît sa relation à tout sous-domaine, c'est-à-dire s'il lui appartient ou pas (tous les points d'un sous-domaine ne correspondent pas forcément à une surface contiguë). On connaît aussi la valeur en ce point de la grandeur que l'on veut estimer, s'il ne s'agit pas d'une surface (volume de bois par exemple). Ceci est indispensable pour garantir l'additivité des résultats (tout = somme des parties) pour des ventilations quelconques [5].

La variance d'une grandeur estimée au niveau d'un sous-domaine peut être décomposée en deux ou trois composantes additives (voir en annexe pour leur calcul) :

- pour l'estimateur S_f de la surface du sous-domaine f :
$$V(S_f) = V_{SD}(S_f) + V_{ST}(S_f)$$
- et pour l'estimateur $T_f(y)$ du total de y (grandeur autre que la surface) pour f :
$$V(T_f(y)) = V_{SD}(T_f(y)) + V_{ST}(T_f(y)) + V_{HT}(T_f(y))$$
 - V_{SD} est la variance « de sous-domaine » introduite par l'incertitude sur la surface du sous-domaine.
 - V_{ST} est la variance liée à la stratification, introduite par les erreurs d'échantillonnage sur les surfaces de strates, lorsque celles-ci sont estimées à partir du nombre de points (elle est nulle si les surfaces sont connues).
 - V_{HT} est la variance « d'hétérogénéité » de la grandeur considérée, due à sa variabilité intra-strate. Elle n'est pas définie si la grandeur est la surface du sous-domaine, car celle-ci n'a qu'une seule valeur, contrairement aux autres grandeurs qui ont une valeur par point [5].

5.4. Stratifications appliquées à l'IFN et outil de calcul en ligne OCRE

Dans le cas de l'IFN, la diffusion des résultats auprès des utilisateurs, telle que requise par le contrat État-IGN, nécessite de pouvoir ventiler les résultats selon certains critères géographiques. Ainsi, les découpages administratifs sont un critère de stratification particulièrement pertinent lorsque les données sont présentées par région, échelle de restitution des résultats auprès des collectivités publiques, ainsi que pour les résultats publiés sous forme papier depuis l'édition 2006. Le régime de propriété de la forêt (domaniale, autre forêt publique ou forêt privée, issu des données de l'Office National des Forêts) est aussi une clé fréquente de répartition. L'IFN dispose de plus de ses propres découpages du territoire, SER et GRECO (cf. supra), qui sont plus pertinents sur le plan écologique, et il utilise aussi les types forestiers définis par la FAO.

Les stratifications qui servent normalement à générer les résultats de l'inventaire utilisent donc des strates croisant les limites de départements et les types de propriété, ce qui permet d'avoir des

valeurs exactes de surface par département (et donc aussi pour les régions). Les strates ainsi définies ont en effet une surface connue cartographiquement, donc considérée comme sans erreur, et en particulier sans erreur d'échantillonnage.

Les phases de stratification correspondent aux deux étapes d'inventaire : l'échantillon utilisé pour construire une phase 1 comporte tous les points tirés pour photo-interprétation l'année considérée ; en phase 2, on garde les points hors forêt identifiés avec certitude par la photo-interprétation (champs...) et ceux pour lesquels on dispose d'une vérité de terrain. En phase 2, un nombre minimal de points par strate est indiqué (en pratique 10).

L'interface en ligne OCRE permet de générer les résultats d'inventaire à la volée en se basant sur les stratifications disponibles. Il en existe deux versions différentes :

1. La version publique librement accessible (<http://inventaire-forestier.ign.fr/ocre-gp/ocregp/initCP.html>) se base sur les stratifications par défaut des cinq dernières années. L'utilisateur peut choisir un découpage administratif (interrégions, régions) ou écologique (GRECO, SER), ou encore une ventilation ou non des résultats par certaines variables. L'ancienne méthode d'inventaire est de plus disponible pour les données historiques. Les résultats sont fournis sous forme de tableaux et histogrammes exportables (Figure 11). Les données brutes sont également accessibles.



Figure 11 : Aperçu de l'interface OCRE dans sa version publique et des résultats fournis.

2. La version de test, accessible uniquement en interne (dont le fonctionnement sera montré par la suite), permet de choisir les stratifications utilisées pour effectuer ces calculs. Il est donc possible de travailler sur des périodes autres que les cinq dernières années, ou d'autres stratifications que celles par défaut. Les variables disponibles sont aussi plus nombreuses. En dehors de la version de test, la stratification utilisée en interne est la même que celle de la version publique.

5.5. Problématique : intérêt de stratifications basées sur un modèle de hauteur

À l'heure actuelle, OCRE n'exploite qu'une seule des nombreuses possibilités de stratification, y compris dans sa version interne. Pourtant, on a vu que la technique de post-stratification fait partie intégrante de la nouvelle méthode d'inventaire et qu'elle offre une grande souplesse dans l'exploitation des données, qui permettrait en principe des gains de précision des résultats. Cette sous-utilisation est avant tout la conséquence du peu de données disponibles France entière.

En effet, une stratification donnée ne peut être optimisée que par rapport à une seule variable à la fois (ou au mieux une combinaison linéaire de variables), la surface mais pas le volume sur pied par exemple [1]. De plus, les surfaces correspondant à un découpage non utilisé dans la constitution de la stratification utilisée doivent être estimées, avec une précision dépendant du nombre de points. La stratification actuelle peut ainsi présenter une erreur de l'ordre de 5 à 10 % dans l'estimation de la surface des GRECO. On peut donc supposer que des stratifications plus spécifiquement adaptées à une grandeur ou un critère de ventilation en particulier fourniraient des estimations plus précises.

Ce serait aussi l'occasion d'exploiter des informations peu ou pas utilisées en stratification pour le moment. Or l'IFN, du fait de son appartenance à l'IGN, a accès à de gros volumes de données géographiques variées. Certaines de ces données peuvent apporter de nouvelles informations sur les peuplements, comme la hauteur des arbres, que l'on sait potentiellement corrélées aux résultats d'inventaire. Elles ont donc un potentiel intéressant à explorer pour décrire et stratifier les données de l'inventaire et pourraient permettre d'en affiner les résultats.

L'objectif du présent travail est donc, sur une zone de test qui est la région Alsace, de :

- mettre en forme les données de hauteur disponibles (modèles numériques de terrain, modèle numérique d'élévation) pour les rendre exploitables en stratification
- mettre en place de nouvelles stratifications, construites notamment avec ces nouvelles données, mais aussi avec d'autres informations potentiellement intéressantes
- vérifier si ces nouvelles stratifications offrent un gain de précision des résultats par rapport à la méthode habituelle de l'inventaire.

III. Utilisation de modèles de hauteur en stratification : données et méthodes

1. Tentatives précédentes

Plusieurs projets ont déjà été menés par l'IFN à titre expérimental pour exploiter des modèles numériques de hauteur des arbres (MNH, aussi appelés dans ce contexte modèles numériques de canopée, MNC) en tant que support à la stratification. Jusqu'à présent, ces projets ont été limités par le fait qu'ils requièrent à la fois un MNE (modèle numérique d'élévation ou MNS, modèle numérique de surface) et un MNT (modèle numérique de terrain) sur la région, pour pouvoir soustraire le MNT au MNS et ainsi obtenir un MNH. Qui plus est, ceux-ci doivent être de qualité suffisante. Les données adaptées n'étant pas disponibles partout, les précédentes tentatives portent donc sur des régions précises :

1. Dans le cadre du projet ANR-10-BIOE-008 « Caractérisation de la ressource forestière pour les bioénergies », différentes stratifications ont été testées sur le département de la Nièvre,

dont une qui associait un MNH à une stratification par types de formations végétales, pour estimer leur surface et leur volume. Cependant, le MNH, bien qu'issu de prises de vue aériennes de haute résolution (25 cm), était de mauvaise qualité car les photos avaient été prises en hiver, donc sans feuillage sur les essences caduques. L'utilisation du MNH a dégradé les résultats par rapport aux seuls types de formations végétales (qui eux donnaient des résultats intéressants) [1][2].

2. Ces travaux ont été prolongés par un stage de master international en écologie appliquée. Le même MNH que dans l'étude précédente a été utilisé pour la Nièvre, tandis que des données lidar (light detection and ranging, obtenues avec un laser aéroporté) étaient disponibles sur la zone Orléans-Sologne, fournissant un MNT et un MNE, donc un MNH, de meilleure qualité. Parmi les stratifications testées, deux associent le MNH (reclassé suivant des classes de hauteur de 5 m) et le type de propriétés forestières, l'une y ajoutant le type de végétation et l'autre la composition en essences du peuplement forestier.

La mauvaise qualité des données nivernaises a de nouveau posé problème, avec cependant des résultats moins mauvais pour les peuplements résineux. L'auteure relève également de mauvais résultats en limite de chantiers photographiques (altitudes aberrantes dues à des problèmes dans le traitement des données). Les résultats étaient bien meilleurs avec le MNH lidar, ce que montre le calcul du facteur d'efficacité (RE) des stratifications correspondantes. Celui-ci est un rapport de variances qui permet de vérifier si une stratification offre de meilleurs résultats que la méthode de référence (cf. infra). C'était systématiquement le cas dans ce travail, sauf pour le charme. On touche là à une limite de la méthode, qui est la représentation correcte des essences de sous-étage par un MNH. [4]

2. Données utilisées dans le cadre du présent travail

La couverture du territoire par des MNT et MNH est actuellement suffisamment avancée pour que des régions entières soient disponibles. L'objectif de ce stage est donc d'explorer les possibilités offertes par ces données pour la stratification à une échelle plus importante que précédemment et en automatisant le plus possible les opérations. Avec deux départements seulement mais des forêts diversifiées, la région Alsace présentait un bon compromis en termes de taille et de complexité. Toutes les données nécessaires y étaient de plus disponibles, y compris la cartographie à jour des types forestiers (BD Forêt V2).

Les modèles numériques fournis par l'IGN pour toute l'Alsace ont été déposés sur le serveur de l'IFN. Ils comprennent :

- Un MNT complet sur la région, formé de tuiles de 5 x 5 km et de 5 m de résolution, obtenu à partir de la BD Alti® par numérisation manuelle ou vectorisation.
- Un MNT lidar disponible sur une grande partie de la région, mais loin d'être exhaustif. Les données disponibles couvrent notamment les zones inondables comme la plaine du Rhin, mais aussi des zones de forêt dans le Haut-Rhin (département 68) qui est beaucoup mieux couvert que le Bas-Rhin (67). Elles prennent la forme de tuiles de 1 km de côté et 1m de résolution, mais beaucoup d'entre elles sont incomplètes vu l'étroitesse de certaines des zones couvertes. Elles sont accompagnées d'une grille vectorielle qui indique leur position, et de jeux de tuiles complets représentant l'origine des données MNT disponibles (Figure 12), plus la distance d'interpolation dans le cas du lidar.

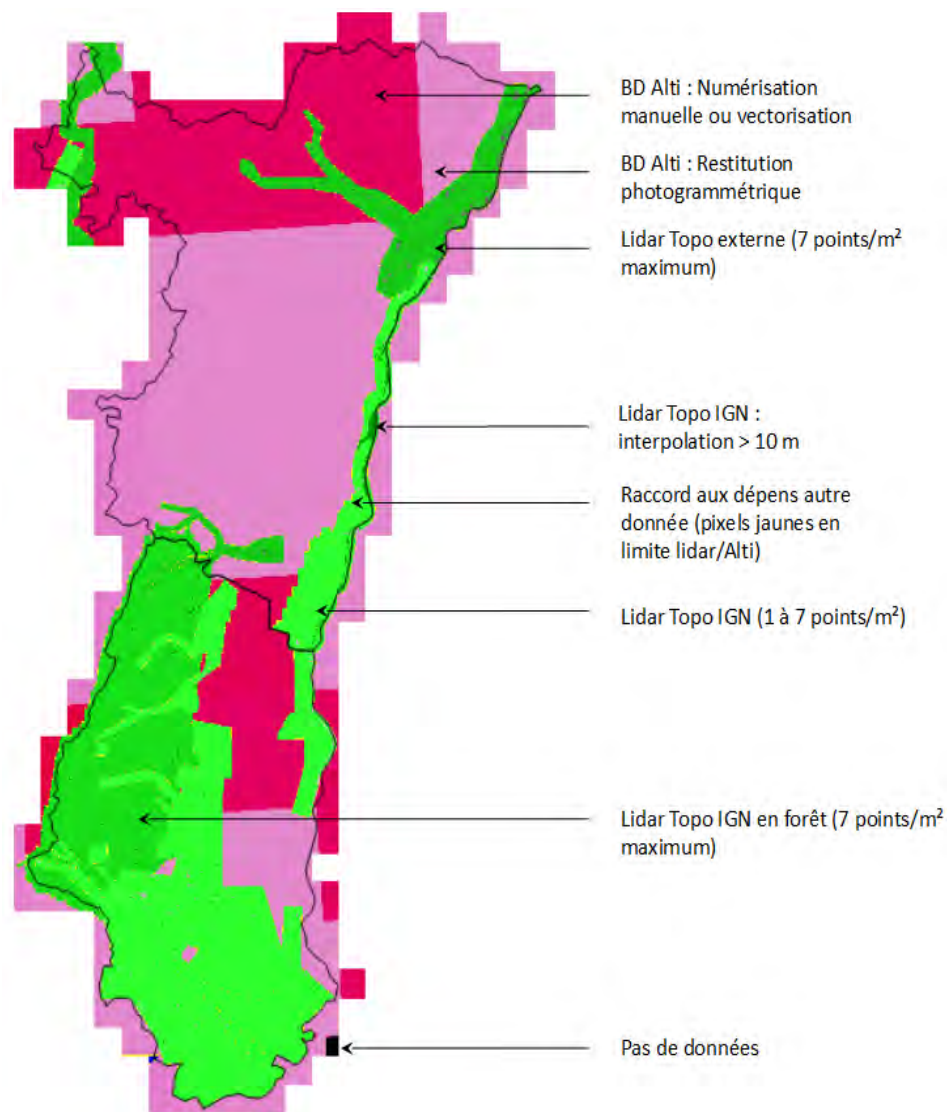


Figure 12 : Origine des meilleures données MNT disponibles pour toute la région Alsace.

- Un MNE d'une résolution de 20 cm, sous forme de tuiles formant une grille de 1 km de côté, disponible sur toute la région. Les tuiles sont un peu plus grandes que 1 km de côté, dépassant de 2 m des cases de la grille (la même que pour les tuiles MNT lidar) dans toutes les directions, si bien qu'elles se chevauchent entre elles. Il est issu d'une reconstitution du relief par photogrammétrie à partir de la BD Ortho®, sa résolution correspondant à celle des orthophotos d'origine.

Toutes ces tuiles comportent dans leur nom les coordonnées de leur sommet nord-ouest pour les MNT, sud-ouest pour le MNE (de la forme . . . _x_y_ . . .), ce qui permet d'identifier chaque fichier.

La base de données de l'inventaire dispose aussi de données vectorielles PostGIS. Celles-ci incluent notamment une couche « communes » et la BD Forêt V2. Comme ce sont des données France entière, plutôt que d'être utilisées directement, elles ont servi à créer des couches vectorielles limitées à l'Alsace : les contours de la région, la BD Forêt V2 restreinte à l'Alsace et une couche tirée de celle-ci correspondant à la présence/absence de forêts (landes arbustives non comprises).

3. Préparation et exploration des données

Les données telles qu'elles sont initialement disponibles sont beaucoup trop lourdes et nombreuses pour qu'il soit possible de les visualiser toutes (12 308 tuiles de 1 km² pour le MNE soit 1 134 Go, 7 Mo par tuile du MNT Alti, les moins lourdes ; près de 2 To pour l'ensemble des données). Par conséquent, leur qualité n'a pas pu être vérifiée au préalable, et les problèmes dus aux données elles-mêmes ont souvent dû être repérés et corrigés au fur et à mesure qu'ils apparaissaient.

La grande majorité des opérations a été effectuée avec le logiciel libre de statistiques R (version 3.3.0, associé à l'environnement RStudio 1.0.143) via des scripts spécifiques. Ceux-ci font notamment appel aux packages **inventR**, créé et utilisé en interne à l'IFN (auteur : François Morneau), et **data.table**, qui permet des gains de temps substantiels lorsqu'on traite de très nombreux fichiers. Ce dernier permet de créer des tableaux de données (« data tables ») autorisant de nombreuses opérations avec une syntaxe simple et des temps de calcul rapides. Tous les tableaux cités par la suite lors des manipulations sous R sont des « data tables ». Les opérations sur les tuiles MNT et MNE elles-mêmes ont été effectuées avec les fonctions du package **raster**.

Les scripts et leur commentaire détaillé sont présentés en annexe du présent document sous forme de rapports RMarkdown. La visualisation des données, notamment pour repérer et comprendre les problèmes, a été effectuée avec le SIG QGIS Desktop version 2.18.4 puis 2.14.14, plus stable.

3.1. Calcul du MNH

La première étape du travail a consisté à obtenir un modèle de hauteur de la forêt alsacienne au complet. À partir des données disponibles sur le serveur, l'objectif était d'avoir un MNH ramené à une résolution de 1 m et uniquement sur les zones de forêt de l'Alsace, les données d'origine débordant toujours un peu des contours de la région et couvrant le terrain quelle que soit son occupation. De plus, comme les données MNT lidar sont de bien meilleure qualité que le MNT Alti, il fallait trouver le moyen de les utiliser en priorité là où elles sont présentes.

3.1.1. Détermination des tuiles MNT manquantes

Comme le MNT Alti couvre toute la région, il sert de base à la zone de travail pour cette étape. On liste donc d'abord les tuiles de 5 km de côté (tuiles Alti) puis, en se basant sur leurs coordonnées, on définit les coordonnées des 25 tuiles de 1 km² correspondant à chacune (Table 1).

Table 1 : début du tableau *tuiles_1m*.

dep	nom	mnt_5m	coord	x_tuile	y_tuile
68	RGEALTI_FXX_0985_6746_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	0985_6746	985	6746
68	RGEALTI_FXX_0985_6747_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	0985_6747	985	6747
68	RGEALTI_FXX_0985_6748_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	0985_6748	985	6748
68	RGEALTI_FXX_0985_6749_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	0985_6749	985	6749
68	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6750_MNT_LAMB93_IGN69.asc	0985_6750	985	6750
68	RGEALTI_FXX_0985_6751_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6755_MNT_LAMB93_IGN69.asc	0985_6751	985	6751
68	RGEALTI_FXX_0985_6752_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6755_MNT_LAMB93_IGN69.asc	0985_6752	985	6752
68	RGEALTI_FXX_0985_6753_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6755_MNT_LAMB93_IGN69.asc	0985_6753	985	6753
68	RGEALTI_FXX_0985_6754_MNT_LAMB93_IGN69.asc	RGEALTI_FXX_0985_6755_MNT_LAMB93_IGN69.asc	0985_6754	985	6754

On se retrouve donc avec une liste de toutes les tuiles de 1 km² qui seraient nécessaires pour couvrir l'emprise du MNT Alti. Certaines de ces tuiles existent déjà en tant que tuiles lidar mais il faut vérifier si c'est le cas. De plus, de très nombreuses tuiles lidar sont incomplètes, en particulier là où

seules des zones à risque d'inondation étroites ont été couvertes (vallée de la Moder dans le Bas-Rhin par exemple). On dresse donc aussi la liste des tuiles lidar en vérifiant lesquelles sont incomplètes, et on confronte les deux listes. S'il n'existe pas de tuile lidar correspondant à un nom de tuile listé, ou que la tuile lidar est bien présente mais qu'au moins un de ses pixels est vide, il faudra créer une tuile correspondante à partir du MNT Alti.

Vu la taille des tuiles Alti de 25 km², elles débordent largement de la région et certaines des tuiles de 1 km² qui en dérivent sont entièrement en dehors. Un test d'intersection entre chaque tuile et la couche vectorielle correspondant à la région permet de ne pas traiter ces tuiles non pertinentes. De la même manière, on utilise la couche présence/absence de forêts pour déterminer quelles tuiles ne recourent aucune zone boisée.

3.1.2. Création des tuiles MNT manquantes

Les tuiles de 5 km de côté issues de la BD Alti ont une résolution de 5 m, au lieu de la résolution de 1 m souhaitée. Avant de les découper pour obtenir les tuiles voulues, il faut donc les faire passer à la bonne résolution (fonction *disaggregate*). Vu le poids de ces données, le calcul est long et on cherche à l'effectuer uniquement si nécessaire : si au moins l'une des tuiles de 1 km² théoriques recouvrant la tuile MNT Alti est nécessaire, mais que sa version lidar est incomplète ou manquante.

Malheureusement, le découpage du MNT Alti est décalé de 2 m vers le nord et 2 m vers l'ouest par rapport aux tuiles lidar, sur lesquelles on souhaite s'appuyer. Les futures tuiles de 1 km² situées en bordure est ou sud d'une tuile Alti dépassent ainsi de 2 m sur la tuile Alti voisine, donc on aura besoin de celle-ci aussi à la bonne résolution ; si elles se trouvent au coin sud-est, on aura besoin de plus de la tuile Alti située au sud-est. Toutes ces situations doivent donc être testées pour chaque future tuile pour savoir quelles tuiles Alti traiter, et rallongent d'autant le temps de calcul.

Cependant, comme les tuiles Alti peuvent ne pas être toutes disponibles, certaines tuiles MNT se retrouveront incomplètes (Figure 13). Vu les faibles surfaces concernées et comme ces bandes ou coins seront en dehors de la région, les tuiles seront quand même utilisées telles quelles.

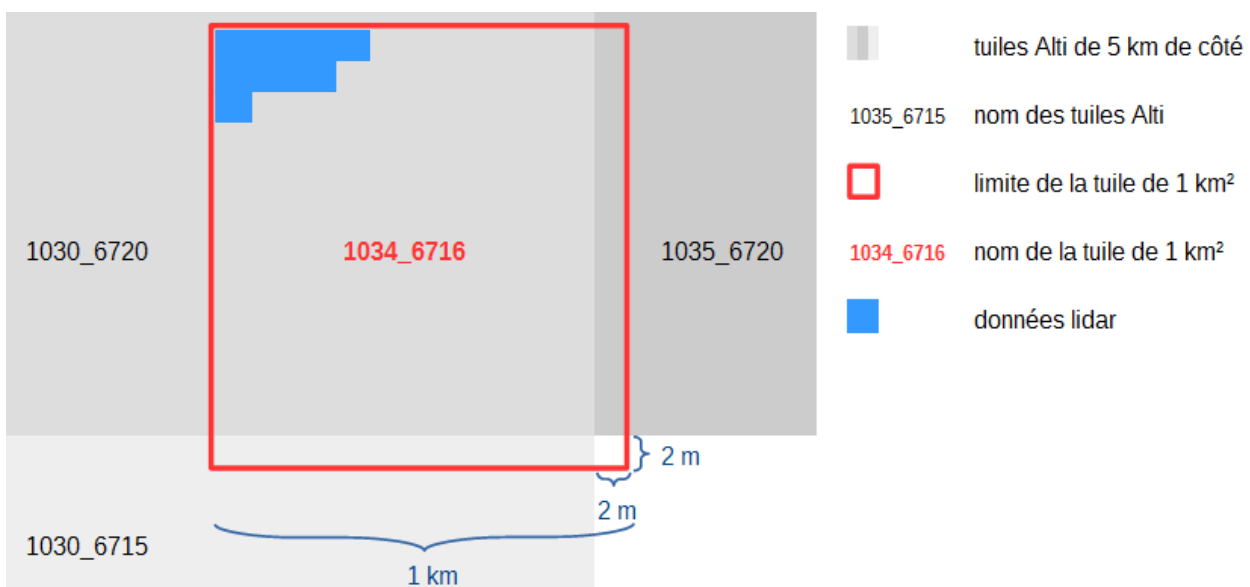


Figure 13 : exemple de tuile située à la fois en limite sud-est de la tuile Alti correspondante et en limite sud-est de la zone de travail (le schéma n'est pas à l'échelle). Ici, une des tuiles Alti nécessaires (1035_6715) n'existe pas dans le jeu de données, il manquera donc un coin de 4 m² à la version Alti de 1034_6716.

3.1.3. Mise en correspondance MNE-MNT et calcul du MNH

Après ces opérations, on dispose d'une couverture uniforme de la région par des tuiles MNT de même taille et même résolution, partout où c'est nécessaire, et lidar si possible. Il faut ensuite sélectionner les tuiles MNE et associer à chacune la donnée MNT appropriée. On commence par vérifier la pertinence (en Alsace et en forêt) de chaque tuile MNE, comme on l'a fait pour le MNT.

À ce stade, la liste des tuiles MNE comporte des doublons à la limite entre les deux départements. En effet, les MNE ont été créés département par département, en dépassant largement les frontières de ceux-ci, mais les tuiles en bordure sont toutes incomplètes du point de vue des données : la plupart ont une emprise normale (et ne sont donc pas détectables par un test simple), mais une partie de la tuile est vide (Figure 14).

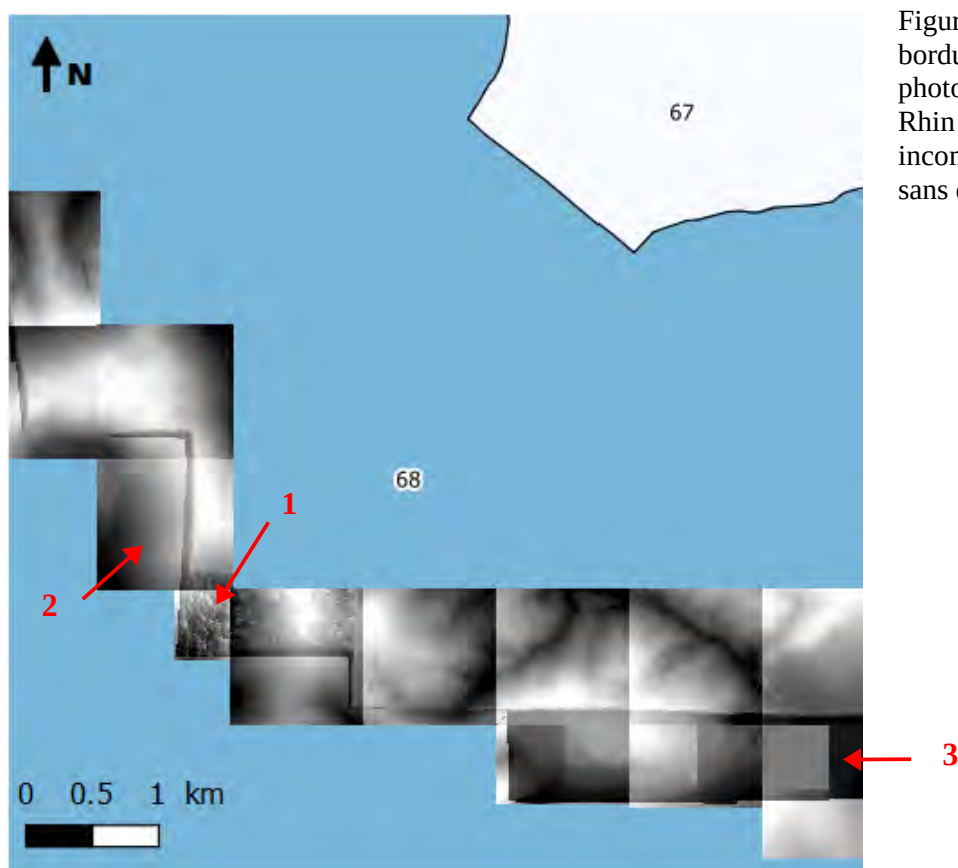


Figure 14 : quelques tuiles en bordure sud du chantier photographique pour le Bas-Rhin (67), montrant des tuiles incomplètes (1) et des zones sans données (2, 3).

Si l'on élimine simplement les doublons par ordre alphabétique du département, on garde ces tuiles incomplètes, qui peuvent couvrir plusieurs rangs. Aussi, pour l'éviter, on introduit un test qui vérifie si la tuile est entièrement incluse dans un département différent de celui auquel elle est censée appartenir. Comme c'est le cas des tuiles de bordure, on élimine les tuiles pour lesquelles ce test est vérifié (qui sont toutes des doublons) avant d'éliminer les doublons restants.

On crée ensuite une table de correspondance entre tuiles MNE sélectionnées et tuiles MNT, où l'on privilégie la tuile lidar si elle existe et est complète (Table 2).

Table 2 : début de la table de correspondance. Noter l'origine différente des données MNT lidar ou non.

coord	test_mne	dep	mne_20cm	x_tuile.x	y_tuile.x	test_NA	mnt
0987_6752	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_987_6751.TIF	987	6752	FALSE	E:\Documents Sgeorges\Data\MNT_1m\MNT_1m_RGEALTI_FXX_0987_6752_MNT_LAMB93_IGN69.asc
0987_6753	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_987_6752.TIF	987	6753	FALSE	E:\Documents Sgeorges\Data\MNT_1m\MNT_1m_RGEALTI_FXX_0987_6753_MNT_LAMB93_IGN69.asc
0987_6754	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_987_6753.TIF	987	6754	FALSE	E:\Documents Sgeorges\Data\MNT_1m\MNT_1m_RGEALTI_FXX_0987_6754_MNT_LAMB93_IGN69.asc
0987_6755	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_987_6754.TIF	987	6755	FALSE	E:\Documents SGeorges\Data\MNT_1m\MNT_1m_RGEALTI_FXX_0987_6755_MNT_LAMB93_IGN69.asc
0988_6751	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_988_6750.TIF	988	6751	FALSE	\\Dgs1211n001/mns/MNT_Alsace\MNT_1m\1_DONNEES_LIVRAISON_2017-03-00106\RGEALTI_MNT_1M_ASC_LAMB93_IGN69_42_20170307\RGEALTI_FXX_0988_6751_MNT_LAMB93_IGN69.asc
0988_6752	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_988_6751.TIF	988	6752	FALSE	\\Dgs1211n001/mns/MNT_Alsace\MNT_1m\1_DONNEES_LIVRAISON_2017-03-00106\RGEALTI_MNT_1M_ASC_LAMB93_IGN69_42_20170307\RGEALTI_FXX_0988_6752_MNT_LAMB93_IGN69.asc
0988_6753	TRUE	68	\\Dgs1211n001/mns/FD68/Dalles/MNE_FD68_988_6752.TIF	988	6753	FALSE	\\Dgs1211n001/mns/MNT_Alsace\MNT_1m\1_DONNEES_LIVRAISON_2017-03-00106\RGEALTI_MNT_1M_ASC_LAMB93_IGN69_42_20170307\RGEALTI_FXX_0988_6753_MNT_LAMB93_IGN69.asc

Il faut ensuite redécouper les tuiles MNE, plus grandes, et les amener à une résolution de 1 m au lieu de 20 cm. Ainsi, pour chaque case, on a une tuile MNT et une tuile MNE parfaitement calées et de la même étendue (sauf les éventuelles données MNT manquantes à l'est et/ou au sud). Le changement de résolution est obtenu avec la fonction *aggregate*, paramétrée de manière à ce que les nouveaux pixels prennent la plus forte valeur parmi les 25 pixels d'origine, car la hauteur maximale est la donnée pertinente pour l'exploitation ultérieure du MNH. Il est alors possible de calculer le MNH tuile par tuile en soustrayant le MNT au MNE.

Seules les données MNH à l'intérieur des zones de forêt telles que définies dans la BD Forêt V2 sont intéressantes pour la stratification. La dernière opération consiste donc à intersecter les tuiles MNH avec la couche forêt pour ne conserver que ces zones. Après cette dernière opération, on dispose d'un MNH qui couvre entièrement et uniquement les zones boisées de la région, sans trou ni chevauchement (Figure 15).



Figure 15 : le MNH découpé (à droite) comparé à l'orthophoto seule (à gauche). Remarquer les zones de hauteur inférieure plus foncées, plus difficilement lisibles sur l'orthophoto.

De 10 300 tuiles de 1 km² potentielles définies initialement, on est passé à 7 972 tuiles MNH réellement créées en éliminant les tuiles hors région ou ne comportant pas du tout de forêts.

3.2. Opérations statistiques sur le MNH et hauteur dominante

3.2.1. Statistiques descriptives sur le MNH

Cette deuxième phase de travail prépare directement la stratification proprement dite. La première partie concerne l'acquisition de données statistiques sur le MNH. On souhaite en effet non seulement vérifier sa qualité (puisque les données, encore trop lourdes à ce stade, ne peuvent pas être visualisées en totalité dans QGIS), mais aussi connaître la répartition des valeurs pour créer un modèle simplifié qui soit pertinent.

On stocke dans un tableau, pour chaque valeur de -100 à $+400$, le nombre de pixels par tuile qui y correspondent. Cependant, le tableau obtenu est trop lourd (3 993 972 lignes) pour pouvoir être lu. On s'en sert donc pour créer deux tableaux plus exploitables : un qui donne les valeurs minimale et maximale par tuile pour pouvoir remonter à celles qui présentent des valeurs aberrantes, et un qui synthétise le nombre d'observations par classe de valeur pour tout le MNH. Ce dernier permet notamment de réaliser des graphes (Figure 17) et de lire rapidement les extrema.

Ces données ont permis de se faire une meilleure idée de la qualité générale du MNH. Ainsi, il s'avère que le calage vertical entre MNT et MNE est de qualité variable. Une large majorité des tuiles MNH présente des valeurs minimales négatives (7 577 sur 7 972, mais ce minimum n'est inférieur ou égal à -1 m « que » pour 5 546 d'entre elles). Pour 940 tuiles, le minimum atteint au moins -10 m : la consultation de certaines de ces tuiles dans QGIS confirme que ces valeurs négatives représentent dans certains cas une proportion relativement importante de leur surface.

Cependant, les statistiques globales relativisent ce problème : si ces 940 tuiles représentent près de 12 % des tuiles, les surfaces avec une altitude négative restent limitées. Seuls 1,3 % des pixels appartiennent aux classes de valeur inférieures à -1 . Les tuiles vérifiées dans QGIS, qui sont les cas les plus extrêmes, se situent pour l'essentiel au nord-ouest de la région, dans une zone issue de la numérisation de la BD Alti, soit la plus mauvaise qualité de données MNT disponible. Le MNH dans son ensemble est donc malgré tout d'une qualité satisfaisante.

3.2.2. Du MNH au modèle de hauteur dominante

Les étapes suivantes n'utilisent plus directement le MNH, mais un modèle dérivé qui présente deux grandes différences. Premièrement, la résolution n'est plus que de 25 m au lieu de 1 m. En effet, une résolution de 25 m permet de se rapprocher de l'échelle des mesures réalisées sur les arbres des placettes de terrain (sur 15 m de rayon) afin que le modèle leur soit comparable. Une résolution de 30 m serait encore plus adaptée, mais 25 m est une division entière des tuiles existantes.

Deuxièmement, plutôt que la hauteur brute, on souhaite utiliser une métrique qui serait fortement corrélée au volume. Pour ce faire, on s'inspire d'une variable très utilisée en foresterie : la hauteur dominante H_0 . Celle-ci est généralement définie pour une zone donnée comme la hauteur moyenne des 100 plus gros arbres (soit globalement ceux qui forment la canopée) à partir d'un diagramme fréquence / diamètre. Il s'agit donc de données purement de terrain et très limitées spatialement, non accessibles directement à partir du MNH.

On peut cependant extraire, à partir des données de hauteur du MNH, un modèle similaire en termes de corrélation au H_0 des forestiers. La métrique retenue est donc un simple quantile à 99 % à partir de la hauteur, que par abus de langage nous considérerons comme la hauteur dominante, et qu'il est facile d'intégrer aux fonctions du package inventR. En revanche, celles-ci ont été conçues pour

travailler avec des « data tables » et ne proposaient aucun moyen d'exporter les résultats. Les fonctions nécessaires ont donc été rajoutées, et ont depuis été intégrées au package (Figure 16).



Figure 16 : à gauche, la tuile MNH avant traitement ; à droite, la tuile résolue à 25 m avec « H_0 ».

3.2.3. Hauteur dominante par classe

Le modèle numérique de hauteur dominante n'est pas non plus utilisé directement : les valeurs sont tout d'abord reclassées. En s'appuyant sur le graphe issu des statistiques sur le MNH, huit classes ont été définies : hauteur « nulle » (-3 à 4 m, pour tenir compte des imprécisions du MNH ; on voit sur le graphe que le pic correspondant s'étend de part et d'autre de 0), 4 à 10 m, 10 à 15 m, 15 à 19 m, 19 à 22 m, 22 à 26 m, 26 à 30 m et plus de 30 m (Figure 17). Une fonction a été définie qui associe à chaque valeur initiale du raster une valeur de 1 à 8 sur ces critères.

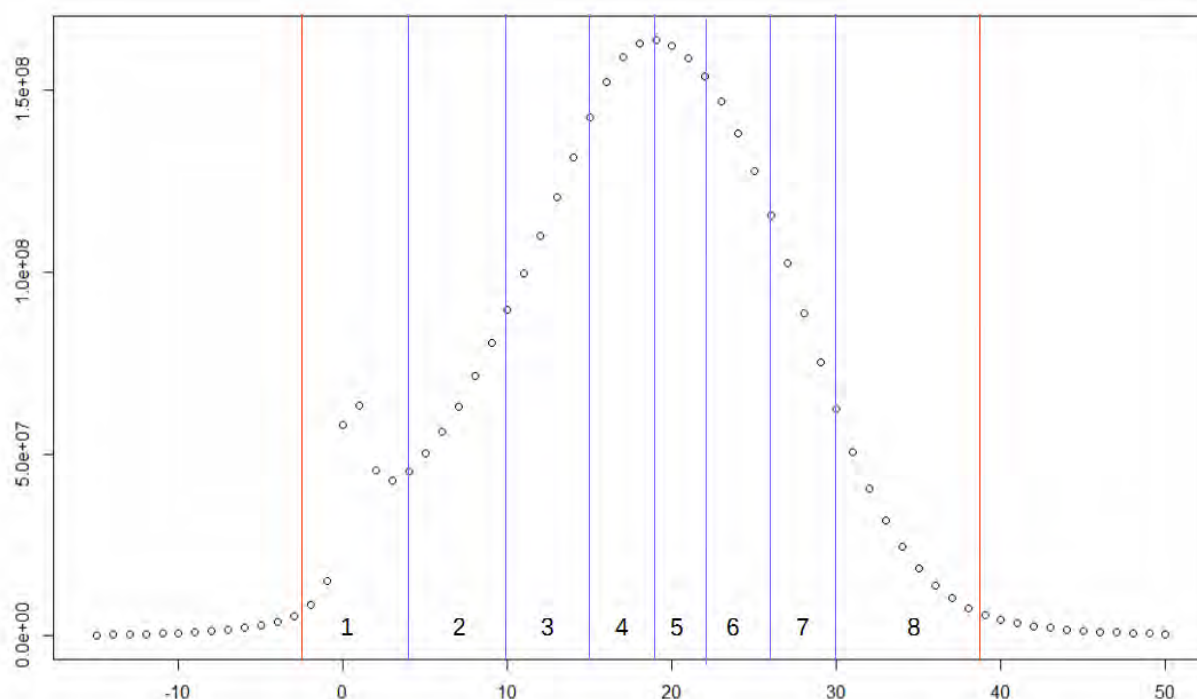


Figure 17 : histogramme de distribution des hauteurs dominantes avec les huit classes définies.

Les tuiles résultantes sont plus légères que celles du modèle de hauteur dominante non reclassé : 10 Ko maximum contre près du triple. Par conséquent, elles ont pu être toutes visualisées, ce qui n'était pas possible précédemment et a permis de repérer certaines aberrations (des problèmes de

géométrie sur des polygones troués, repris sous QGIS). Les tuiles ont ensuite été fusionnées en une seule couche, plus simple à manipuler (elle ne pèse que 144 Mo) (carte en annexe).

3.2.4. Exploration de la BD Forêt et choix des regroupements

Parallèlement aux opérations sous R, des statistiques ont été faites sur la couche BD Forêt V2 avec l'extension « Group Stats » de QGIS pour déterminer quels essences ou mélanges d'essences correspondent en Alsace à des surfaces suffisantes pour qu'une stratification soit intéressante. D'emblée, certains types de formation ou essences peu fréquents dans la région sont apparus comme à regrouper. La nouvelle nomenclature retenue au final est la suivante (Table 3) :

Table 3 : regroupements effectués sur les essences ou types d'essences de BD Forêt V2.

Code	Libellé d'origine	Regroupements	Surface totale (ha)		
FF1-00	Feuillus purs en îlots	--	9217,63		
LA6	Formation herbacée	Peuplements non arbustifs	10021,05		
LA4	Lande ligneuse				
FF2-64-64	Douglas pur	--	12826,47		
FO2	Forêt ouverte de conifères purs	Autres conifères	15364,85		
FF2-51-51	Pin maritime pur				
FF2G53-53	Pin laricio ou pin noir pur				
FF2-00	Conifères purs en îlots				
FF2-91-91	Autre conifère pur autre que pin				
FF2-63-63	Mélèze pur				
FF2-80-80	Mélange de pins purs				
FF2-90-90	Mélange d'autres conifères				
FO1	Forêt ouverte de feuillus purs			Autre feuillu pur	21198,10
FF1-10-10	Châtaignier pur				
FP	Peupleraie				
FF1-14-14	Robinier pur				
FF1-49-49	Autre feuillu pur				
FO0	Forêt ouverte avec coupe rase ou incident	Coupe rase ou incident	28433,47		
FF0	Jeune peuplement ou coupe rase ou incident				
FF2-00-00	Mélange de conifères	--	31738,91		
FF2-52-52	Pin sylvestre pur	--	45004,60		
FF1-09-09	Hêtre pur	--	56525,51		
FF1G01-01	Chênes décidus purs	--	68081,75		
FF32	Mélange de conifères prépondérants et feuillus	--	85783,59		
FF2G61-61	Sapin ou épicéa pur	--	92126,78		
FO3	Forêt ouverte à mélange de feuillus et de conifères	Feuillus prépondérants et conifères	93072,85		
FF31	Mélange de feuillus prépondérants et conifères				
FF1-00-00	Mélange de feuillus	--	230891,34		

On peut remarquer que le regroupement « peuplements non arbustifs » ne correspondra pas à des données MNH car, dans la couche forêt/non forêt qui a servi à le découper, ces formations avaient

été exclues. À partir de la couche BD Forêt V2 limitée à l'Alsace, une version de celle-ci regroupée suivant cette nouvelle nomenclature a été créée. L'opérateur « regrouper » a ensuite été utilisé sur le champ correspondant afin de n'avoir plus qu'un seul multi-polygone par essence ou regroupement d'essences (carte en annexe).

Par ailleurs, il a été décidé de travailler à l'échelle des GRECO. Trois d'entre elles sont présentes en Alsace : Grand Est semi-continentale (C) en plaine d'Alsace et à la pointe ouest du Bas-Rhin, Vosges (D), et Jura (E) à l'extrême sud de la région. Étant donné la faible superficie de cette dernière GRECO en Alsace, il a été décidé de la regrouper avec une des deux principales. Elle a été regroupée avec D pour former la zone DE car la composition en essences de E s'en rapprochait davantage : hêtre et sapins ou épicéas comme principales essences hors mélanges (contre chênes décidus et autres feuillus purs pour C), environ 15 % de feuillus seuls en mélange (contre 53 % pour C), au profit de tous les mélanges contenant des conifères (seuls, prépondérants ou dominés).

4. Base de données pour la stratification

4.1. Détermination des surfaces

Une fois les données disponibles et les regroupements actés, il était nécessaire de déterminer la surface des strates initiales qui découlent des croisements entre les variables retenues : essence ou regroupement d'essences, zone C ou DE, et classe de hauteur dominante de 1 à 8. La couche BD Forêt regroupée a été intersectée avec les zones sous QGIS. De plus, les deux couches ainsi obtenues se sont vu rajouter le restant de la zone (avec « union » puis correction de la table d'attributs) comme multi-polygone supplémentaire présentant la valeur « autres » pour le regroupement d'essences. On dispose ainsi d'une couverture complète de la région. « Group Stats » a alors permis de connaître rapidement la surface par essence (la surface en « autres » étant déduite du reste).

Le découpage supplémentaire obtenu par croisement avec les classes de hauteur n'a été appliqué qu'aux combinaisons regroupement d'essences x zone présentant une surface minimale de 1 000 ha (déjà bien en dessous des futurs seuils de regroupement). La couche raster correspondant à la hauteur dominante classée a été intersectée avec chacun des croisements concernés : l'extension QGIS « Landscape Ecology » permet alors de calculer et d'exporter (avec la fonction « Landscape statistics ») la surface valeur par valeur dans chacune de ces couches.

Du fait des intersections sur une couche raster, on perd quelques-uns des pixels de bordure. En faisant la somme des surfaces calculées par « Landscape statistics » et en la comparant à la surface du polygone correspondant, il s'avère que la différence atteint 0,14 % de la surface raster finale au maximum, et moins de 0,1 % dans tous les autres cas. Vu la faible probabilité que des points s'y trouvent par la suite, ces faibles surfaces perdues ont été attribuées à « autres ».

Cependant, il est apparu que la somme des surfaces des strates (835 285,41 ha) était supérieure à la surface de la région (832 353,74 ha), un point de blocage pour la suite des opérations. Sur les 2931,67 ha de différence, une partie correspond probablement au fait que les pixels de bordure dépassent en partie des frontières de la région, mais ces pixels ne représentent qu'une surface de 606,63 ha lorsqu'on les isole par intersection. Certaines erreurs ont pu être induites par l'utilisation d'intersections entre vecteur et raster (peut-être des pixels ont-ils pu se retrouver dans plusieurs strates ?). Toutes les couches utilisant la projection Lambert 93, le problème se situe ailleurs.

Faute d'explication définitive et donc de véritable solution, les surfaces des strates ont été recalculées avec un produit en croix afin que leur somme corresponde bien à la surface de la région.

4.2. Entrée des données dans la base pour stratification

Avec toutes ces données, on peut maintenant définir des strates dans la base de données de l'inventaire. Pour ce faire, il faut mettre en base de données les informations suivantes :

- Une table contenant les surfaces par strates, les strates étant définies par un regroupement d'essences, une zone (C ou DE) et une classe de hauteur (de 1 à 8, ou « non_stratifié » si la surface était trop faible). On obtient ainsi une table nommée « e1sigmnh » où les regroupements d'essences sont codés regrt, la zone C ou DE z_eco et la classe de hauteur hdom_cl (Table 4) :

Table 4 : premières lignes et colonnes de *e1sigmnh* telle qu'elle apparaît en base.

regrt	z_eco	hdom_cl	surface_sig
FF1-00	C	1	690692,28
FF1-00	DE	1	348771,58
FF1-00	C	2	2217190,72
FF1-00	DE	2	1323463,56

- Les deux couches vectorielles correspondant à la BD Forêt regroupée par zone, pour savoir par la suite dans quel croisement regroupement x zone se situe chaque point tiré au sort (il n'a pas été possible de regrouper les deux couches dans QGIS, l'opération étant trop longue, ni dans R à cause de problèmes sur certains polygones, hérités de la BD Forêt V2).
- La couche raster pour le modèle de hauteur dominante classée, pour attribuer de même une classe de hauteur à chaque point en fonction de sa localisation.

Les colonnes regrt, z_eco et hdom_cl ne correspondant pas à des variables déjà définies dans la base (comme l'est la nomenclature BD Forêt V2 normale par exemple), il a fallu les y ajouter avec les métadonnées correspondantes pour pouvoir s'en servir par la suite comme critères de sélection des strates. Toutes ces opérations ont été effectuées par le développeur des outils de stratification Benoît Pesty.

5. Mise en place des stratifications de test

Lors de la mise en place d'une nouvelle stratification, les phases 1 et 2 sont construites séparément. Une stratification de phase 1 est créée en différentes étapes lors desquelles toutes les strates qui ne comportent pas de point ou sont de surface nulle sont regroupées avec une autre (par exemple, avec nos regroupements d'essence, une essence feuillue trop peu fréquente sera regroupée avec « autres feuillus purs »). Tous les points sont ainsi attribués à une strate et une seule. La stratification est alors publiée et devient disponible comme base pour une ou des stratifications de phase 2.

En phase 2, un nombre minimal de points par strate plus important est indiqué, mais, contrairement à la phase 1, la stratification peut être publiée même si certaines strates ne respectent pas cette condition. Une règle de regroupement définie ne s'applique quoi qu'il en soit que si les strates

concernées ont été définies comme regroupables et ne respectent pas encore les conditions de nombre minimal de points.

L'objectif des premières tentatives de stratification avec nos données était de comparer les résultats obtenus avec les regroupements d'essences et avec les classes de hauteur. Le jeu de données test était celui des points de phase 1 de l'échantillon 2010, c'est-à-dire les points ayant servi à la photo-interprétation lors de la campagne 2010 (1247 points en Alsace, à conserver tous lors des différentes étapes de stratification). Dans un premier temps, une stratification a été tentée utilisant les regroupements d'essences ainsi que les classes de hauteur comme critères regroupables, et les zones écologiques Z_ECO (GRECO regroupées) comme critère non regroupable (Figure 18).

Stratification - Tables utilisées - Critères de sélection - Critères de validité

Critères de sélection

Ajouter des critères de sélection

Donnée	Z_ECO	➔	Schéma	Table	Donnée	Regroupable	Supprimer
Schéma	Inv_exp_nm		INV_EXP_NM	TE1POINT	HDOM_CL	OUI	
Table	TE1POINT		INV_EXP_NM	TE1POINT	REGRT	OUI	
Regroupable	non		INV_EXP_NM	TE1POINT	Z_ECO	NON	

[Suivant](#)

Les critères non regroupables seront conservés jusqu'au bout et serviront de ventilation pour la stratification. Les critères regroupables disparaîtront et serviront pour la fusion des strates ne répondant pas aux critères de validité.

Figure 18 : aperçu de l'interface de mise en place des stratifications : choix des critères de sélection.

Sur 184 strates définies, 84 ne respectaient pas les critères de la phase 1 (Figure 19).

Nombre de strates: **184** Surface de domaine : 8 323 537 363 m² Stratification: **ALSACE_2010_1**

Nombre de strates OK: **100** Nb points échantillon : 1247 **Conformité: En cours**

Nombre de strates KO: **84** [Exporter au format CSV](#)

N°	Identifiant	Nom	Surface	Nb points	Etat	Regroupable	Règle	Site	HDOM_CL	Z_ECO	REGRT
1	2196261	2196261	3 315 198	1	OK	Oui		F	2	DE	coniferes
2	2196262	2196262	17 921 254	3	OK	Oui		F	4	DE	FF1-09-09
3	2196263	2196263	195 561	0	KO	Oui		F	1	C	FF1G01-01
4	2196264	2196264	33 312 668	6	OK	Oui		F	7	C	FF1G01-01
5	2196265	2196265	4 993 662	0	KO	Oui		F	5	C	coupe
6	2196266	2196266	6 291 590	0	KO	Oui		F	6	DE	coupe
7	2196267	2196267	731 798	0	KO	Oui		F	2	DE	FF1G01-01
8	2196268	2196268	7 712 211	1	OK	Oui		F	3	C	FF2-52-52
9	2196269	2196269	11 728 690	0	KO	Oui		F	8	DE	coniferes

Figure 19 : interface de consultation des strates, montrant le nombre de strates non conformes.

Des regroupements ont donc été effectués, mais n'ont pu être menés à terme car limités par l'impossibilité de fusionner les strates entre zones. Ainsi, la strate correspondant aux chênes de classe de hauteur 1 en zone C (troisième ci-dessus) ne contenait pas de point par exemple, il était possible de la fusionner avec celle des chênes de classe 7 en zone C (quatrième), ou avec les

feuillus de hauteur 1 en zone C, mais pas avec une strate en zone DE. Il est donc apparu que ce découpage était trop fin par rapport à la surface et au nombre de points disponibles.

À la place, des stratifications plus simples ont été mises en place. Les opérations de regroupement restent cependant contraintes par le fonctionnement des règles de regroupement : bien que l'ordre des règles soit pris en compte, on ne peut pas en avoir deux affectant la même strate lors de la même étape (A dans B puis B dans A, A dans B puis B dans C, ou A et B dans C).

En phase 2, en plus des critères déjà mentionnés, les résultats de la photo-interprétation ont été utilisés pour certaines stratifications, comme dans les stratifications par défaut de l'inventaire. Pour ce faire, on utilise les critères suivants : couverture du sol photo-interprétée (CSO), point occulté ou non (OCC), présence ou non de segments ligneux hors forêt (PLHF), taille du massif (TM), utilisation du sol (UTO), indicateur de zone forestière détaillé (ZFORIFND) et altitude (ZPN). On définit grâce à eux les règles de regroupement de l'étape 0, en se basant sur la syntaxe SQL utilisée pour la référence (des changements apportés à l'interface depuis la création de celle-ci permettant au passage une syntaxe moins lourde). La plupart des strates résultantes sont désignées non regroupables (Figure 20). On élimine ainsi certains cas particuliers : eau, landes d'altitude...

Règles de regroupement de l'étape 0

[Voir les strates](#)

Ordre	Nom	Détails	Libellé	Regroupable
1	CS2_F50	OCC = '1' AND (UTO != '2' OR UTO IS NULL) AND CSO IN ('1', '3')	CS2_F50	OUI
2	CS2_P	CSO = '5' AND OCC = '1' AND (UTO != '2' OR UTO IS NULL)	CS2_P	OUI
3	CS2_LA	OCC = '1' AND (UTO != '2' AND ((CSO = '4' AND ZPN < '1700') OR (CSO = '4' AND ZFORIFND NOT IN ('3','6')) OR UTO is null AND ((CSO = '4' AND ZPN < '1700') OR (CSO = '4' AND ZFORIFND NOT IN ('3','6')))))	CS2_LA	OUI
4	CS2_LHF+BOSQ	OCC = '1' AND ((UTO != '2' AND (CSO = '2' OR PLHF = '1')) OR (UTO is null AND (CSO = '2' OR PLHF = '1')))	CS2_LHF+BOSQ	OUI
5	CS2_F1B	OCC = '1' AND TM = '3' AND UTO = '2'	CS2_F1B	NON
6	CS2_F2B	OCC = '1' AND TM = '2' AND UTO = '2'	CS2_F2B	NON
7	CS2_F3B	UTO = '2' AND OCC = '1' AND CSO IN ('5','2')	CS2_F3B	NON
8	CS2_FRICHE	UTO = '2' AND CSO = '4' AND OCC = '1' AND ZPN < '1700'	CS2_FRICHE	NON
9	CS2_LANDALT	PLHF = '0' AND OCC = '1' AND CSO = '4' AND ZPN >= '1700' AND ZFORIFND IN ('3','6')	CS2_LANDALT	NON
10	CS2_AV	CSO = '6' AND OCC = '1' AND (UTO = '2' OR PLHF = '0')	CS2_AV	NON
11	CS2_SVA	CSO = '7' AND OCC = '1' AND (UTO = '2' OR PLHF = '0')	CS2_SVA	NON
12	CS2_SVN	CSO = '8' AND OCC = '1' AND (UTO = '2' OR PLHF = '0')	CS2_SVN	NON
13	CS2_EAU	CSO = '9' AND OCC = '1' AND (UTO = '2' OR PLHF = '0')	CS2_EAU	NON
14	CS2_OCC	OCC = '0'	CS2_OCC	NON

Figure 20 : étape 0 (définition des strates) d'une stratification de phase 2 basée sur la photo-interprétation.

À partir de ces principes, toute une série de stratifications de test a été mise en place, en faisant varier les éléments suivants :

- En **phase 1** : on utilise soit les classes de hauteur, soit nos regroupements d'essence, soit une nomenclature simplifiée obtenue à partir de ces mêmes regroupements et qui correspond à peu près à la nomenclature FAO 2 (la différence étant que dans cette dernière les forêts ouvertes sont considérées à part, tandis qu'ici elles ont été regroupées).
- En **phase 2** : on utilise soit les critères issus de la photo-interprétation (PI), soit nos regroupements d'essence seuls, soit les classes de hauteur seules, soit la photo-interprétation croisée avec les regroupements d'essence ou les classes de hauteur.
- Conserver les GRECO regroupées (**Z_ECO**) s'est avéré une contrainte importante, obligeant à regrouper davantage de strates (par exemple, en zone C, même en regroupant entre elles toutes les modalités concernant les conifères, il a souvent fallu regrouper le résultat avec les feuillus faute de points). Pour chacune des combinaisons, il a donc été décidé de tester une version avec les zones non regroupables et une n'utilisant pas du tout celles-ci.

On obtient donc toutes les possibilités suivantes (Figure 21) :

			stratifications de phase 1			
			basées sur nom	nomenclature FAO 2 moins les forêts ouvertes	regroupements de BD forêt V2 : REGRT	8 classes de hauteur : HDOM_CL
basées sur		avec Z ECO				
stratifications de phase 2	photo-interprétation (CS2_x)	oui	ALSACE_201x_FAO	ALSACE_201x_R1	ALSACE_201x_H1	
		non	ALSACE_201x_FPI	ALSACE_201x_R2	ALSACE_201x_H2	
	regroupements de BD forêt V2 : REGRT	oui	ALSACE_201x_FPI2	ALSACE_201x_RPI2	ALSACE_201x_HP2	
		non	ALSACE_201x_FR2	ALSACE_201x_RR2	ALSACE_201x_HR2	
	regroupements de BD forêt V2 : REGRT croisé avec PI	oui	ALSACE_201x_FRE2	ALSACE_201x_RRE2	ALSACE_201x_HRE2	
		non	ALSACE_201x_FRPZ	ALSACE_201x_RRPZ	ALSACE_201x_HRPZ	
	8 classes de hauteur : HDOM_CL	oui	ALSACE_201x_FRP2	ALSACE_201x_RRP2	ALSACE_201x_HRP2	
		non	ALSACE_201x_FH2	ALSACE_201x_RH2	ALSACE_201x_HH2	
	8 classes de hauteur : HDOM_CL croisé avec PI	oui	ALSACE_201x_FHC2	ALSACE_201x_RHC2	ALSACE_201x_HHC2	
		non	ALSACE_201x_HCL2	ALSACE_201x_RHP2	ALSACE_201x_HHP2	
			non	ALSACE_201x_FHP2	ALSACE_201x_RHI2	ALSACE_201x_HPI2

Figure 21 : stratifications mises en place. 201x correspond à chaque fois aux années 2010 à 2015.

Pour chacune de ces combinaisons, six stratifications ont été créées, correspondant aux années les plus récentes permettant d’avoir une fenêtre glissante de cinq ans complète (2010 à 2014, centrée sur 2012) plus une pour les points de 2015. Ce dernier échantillon a cependant été exclu des analyses car nettement plus petit (480 points en phase 1 et 412 en phase 2, contre plus de 1200 en phase 1 et environ 900 en phase 2 les autres années). On a donc un total de 15 stratifications de phase 1, sur lesquelles sont basées 150 stratifications de phase 2 (sans compter 2015).

Par ailleurs, tous les autres paramètres sont identiques d’une stratification à l’autre. En particulier, le nombre de points minimum pour valider une strate en phase 2 est ici fixé à 10. Il s’agit simplement de la valeur qui est utilisée pour les stratifications de référence de l’inventaire. Cependant, dans le cas présent, vu la zone d’étude restreinte et donc le faible nombre de points, cette règle n’est pas toujours appliquée strictement. Les regroupements se poursuivent cependant au moins de manière à laisser au plus deux strates qui ne la respectent pas par stratification, mais qui contiennent au moins 8 points chacune (le plus souvent 9).

De même, les étapes de stratification ont été menées de la manière la plus uniforme possible. Pour chaque ligne du tableau, l’étape initiale de la phase 2 est la même pour toutes les phases 1 et toutes les années. Les étapes suivantes ont aussi été standardisées au maximum :

- Pour les stratifications qui utilisent la photo-interprétation, les regroupements suivent ce qui est habituellement pratiqué à l’inventaire (landes avec les bosquets, puis si nécessaire ceux-ci avec la forêt, peupliers également avec la forêt). Trois étapes toujours identiques suffisent pour la photo-interprétation seule (Figure 22).

Détails de la stratification ALSACE_2010_H2

Propriétés		Critères de validité	
STRATIFICATION		nb_point > 9	
Nom	ALSACE_2010_H2	Critères de sélection	
Propriétaire	IFN	inv_exp_nm.te1point.cso	
Libellé	ALSACE_2010_H2	inv_exp_nm.te1point.z_eco	
Définition	Stratification de phase 2	inv_exp_nm.te1point.occ	
Phase	2	inv_exp_nm.te1point.plhf	
Etat	Publiée	inv_exp_nm.te1point.tm	
Parent	ALSACE_2010_H1	inv_exp_nm.te1point.uto	
ECHANTILLON	id : 92	inv_exp_nm.te1point.zforifnd	
Domaine	Zonage quelconque TRAD_MANQUANTE	inv_exp_nm.te1point.zpn	
Surface domaine	8 323 537 363 m²	Tables utilisées	
Taille	933	inv_exp_nm.te2point	
Type d'enquête	Surfacique	inv_exp_nm.te1point	
Type d'unité	Point	Etapes de stratification	
Début	2010	<input type="radio"/>	Incrément
Fin	2010	<input type="radio"/>	Libellé
Phase statistique	2	<input type="radio"/>	0
Propriétaire	IFN	<input type="radio"/>	1
Parent	Echantillon Alsace 2010 - phase 1	<input type="radio"/>	2
		<input type="radio"/>	3
			E0
			Regroupement: LHF+BOSQ = LHF+BOSQ + LA
			Regroupement: CS2_F50 = CS2_F50 + CS2_P
			Regroupement: CS2_F50 = CS2_F50 + CS2_LHF+BOSQ

Figure 22 : étapes de regroupement pour la photo-interprétation, donnant un aperçu de l'interface de consultation des détails des stratifications.

- Dans le cas des classes de hauteur, de l'extérieur vers l'intérieur (1 vers 2, 8 vers 7) avec la classe 4 comme classe « par défaut » avec laquelle regrouper la classe « non stratifié ». Les classes les plus basses, étant les moins nombreuses, se sont souvent trouvées regroupées, notamment pour les stratifications les plus complexes. Autant que possible, le regroupement n'a pas été poussé plus loin que 1, 2 et 3 ensemble (si ces trois classes ensemble atteignent 8 points par exemple, on passe outre la règle des 10 points minimum).
- Dans le cas des essences, ont été privilégiés des regroupements logiques : du particulier vers le général (des pins sylvestres vers les autres conifères purs par exemple, même si ceux-ci ont moins de points). Comme mentionné plus haut, cela n'a pas toujours été possible. En général, les landes et coupes ont été regroupées dans la catégorie « feuillus majoritaires », mais vu le faible nombre de points concernés cela a peu de conséquences.

6. Méthode de comparaison entre stratifications

Pour vérifier la pertinence d'une nouvelle stratification, on se base sur le facteur d'efficacité RE, déjà utilisé lors d'études précédentes [1, 4]. C'est un simple rapport de variances :

$$RE = \frac{V(\bar{Y}_{ref})}{V(\bar{Y}_{str})}$$

où $V(\bar{Y})$ est l'estimation de la variance de la moyenne de la variable aléatoire Y , str est la stratification testée et ref est une stratification de référence (ou l'absence de toute stratification). Si RE est supérieur à 1, la stratification testée améliore la précision des résultats par rapport à la référence [1].

Ce calcul est applicable à la variance totale, mais aussi à chacune de ses composantes V_{SD} , V_{ST} et V_{HT} lorsque celles-ci sont définies ; en effet, comme vu précédemment, V_{HT} n'existe pas lorsque la variable y estimée est une surface. De plus, V_{ST} est nulle sur les surfaces lorsque les mêmes strates cartographiques (REGRT ou HDOM_CL) sont utilisées en phase 1 et en phase 2. Dans les deux cas, on ne peut pas calculer le RE correspondant.

Il est également intéressant de vérifier le coefficient de variation :

$$CV = \sigma (T_i(y)) / M_i(y)$$

où $M_i(y)$ est la moyenne de la grandeur y . Le CV est généralement donné en % et donne une idée de la qualité de l'estimation.

Le package `inventR` permet de récupérer le total, le nombre de points, la variance totale et les composantes de celle-ci pour une stratification ou une série de stratifications sur plusieurs années. Par exemple :

```
fao_pi_z <- ocre_nm(ans = 2010:2014, vars = c("ST", "V", "GTOT", "NT", "PV"),
                  stratif = paste0("ALSACE_", 2010:2014, "_FPI"),
                  detvar = TRUE, test = TRUE)
```

permet de récupérer ces résultats pour les années 2010 à 2014 de `ALSACE_201x_FPI`, qui est la série de stratifications avec "nomenclature FAO" en phase 1, la photo-interprétation seule en phase 2 et les GRECO regroupées, pour les grandeurs suivantes : surface, volume, surface terrière (la surface représentée par la section de tous les troncs), nombre de tiges et production en volume.

De premiers essais ont permis de vérifier qu'il existait effectivement des différences entre stratifications, mais pas forcément dans le même sens sur toutes les composantes de la variance. Le résultat s'est avéré dépendre aussi de la grandeur mesurée, avec des différences surtout entre la surface et les autres grandeurs (davantage corrélées au volume). C'est donc la surface et le volume qui ont été utilisés pour effectuer des comparaisons systématiques entre stratifications.

IV. Résultats : comparaison entre les stratifications créées et la référence

1. Comparaisons systématiques sans ventilation

1.1. Facteurs étudiés

Les stratifications correspondant à toutes les combinaisons de variables mises en place et aux années 2010 à 2014 ont été appelées dans R avec la fonction `ocre_nm` comme ci-dessus. Celle-ci présente l'avantage de fournir un résultat sous forme de « data table ». Des manipulations simples permettent donc d'en récupérer le contenu et d'effectuer des opérations dessus pour obtenir, pour une grandeur donnée, la valeur totale, le CV, la variance totale et ses composantes, et calculer les RE correspondants (valant 1 pour les stratifications servant de référence). À ce niveau, les résultats n'ont pas été ventilés selon des variables supplémentaires.

Des tableaux de comparaison ont donc été créés de manière à faire varier un seul paramètre à la fois (Table 5). Les stratifications sont désignées suivant leur phase 1 :

- **REGRT** : stratification avec les regroupements d'essence
- **FAO** : « types FAO 2 » reconstitués à partir de REGRT
- **HDOM_CL** : stratification avec les classes de hauteur

et leur phase 2 :

- **PI** : photo-interprétation
- **REGRT**

- **REGRT x PI** : on utilise les deux critères à la fois en phase 2
- **HDOM_CL** : stratification avec les classes de hauteur
- **HDOM_CL x PI**

Table 5 : comparaisons mises en place.

Référence	Comparée à	Figure
phase 2 PI seule	toutes autres phases 2	23
stratifications avec Z_ECO	stratifications sans Z_ECO	24
phase 2 PI seule	PI associée à REGRT ou HDOM_CL	25
REGRT ou HDOM_CL seuls	REGRT x PI ou HDOM_CL x PI	26
phase 2 REGRT avec ou sans PI	phase 2 HDOM_CL correspondante	27
phase 1 FAO	autres phases 1	28
phase 1 REGRT	phase 1 HDOM_CL correspondante	-

Si l'on compare les phases 2 entre elles (à phase 1 identique) en utilisant PI seule comme référence, on constate que le comportement des stratifications peut être très différent entre surfaces et volumes. Ainsi, pour les volumes, le RE global est supérieur à 1 pour toutes les stratifications et présente une différence nette entre REGRT et REGRT x PI d'une part, et HDOM_CL et HDOM_CL x PI d'autre part. Pour les surfaces en revanche, la situation est très différente (Figure 23). C'est le cas avec ou sans Z_ECO, et sur les composantes de la variance. Par la suite, les résultats sont donc présentés pour les surfaces et pour les volumes chaque fois que nécessaire.

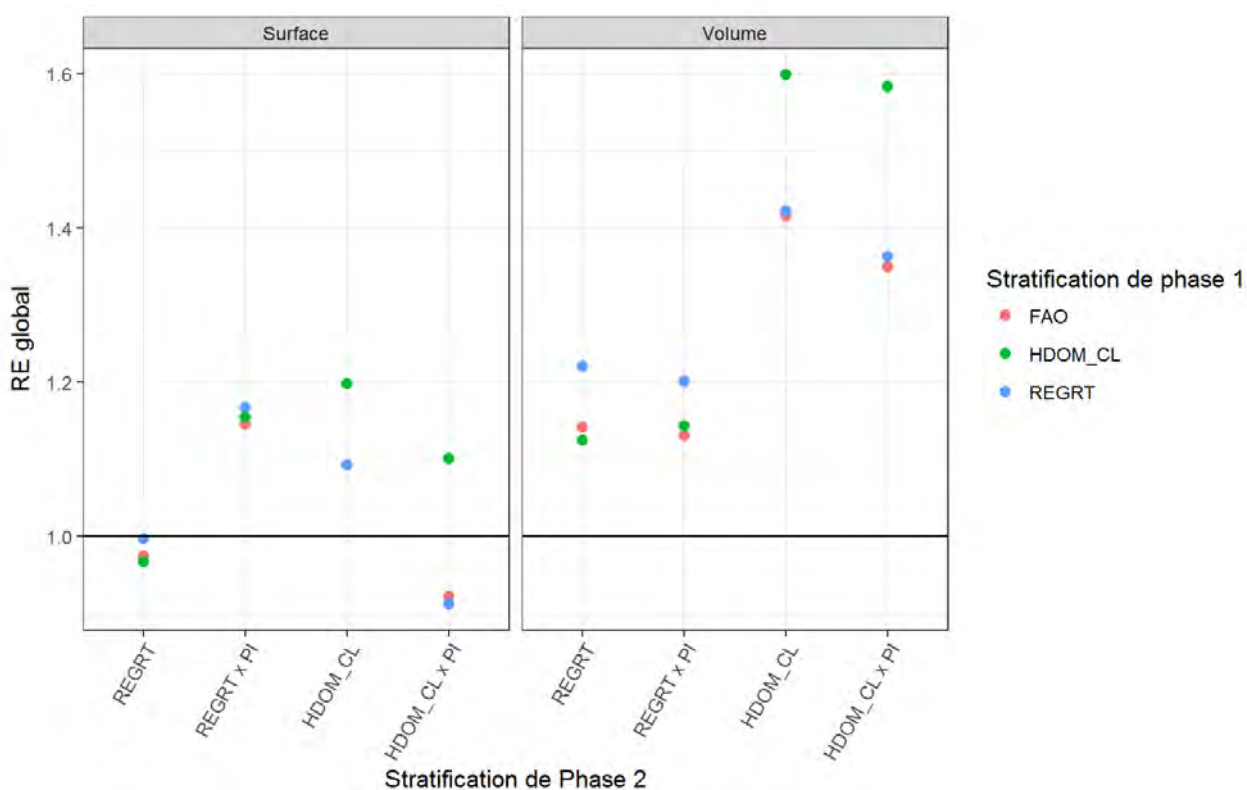


Figure 23 : RE global avec PI seule comme référence, stratifications sans Z_ECO. La référence (RE global = 1) n'est pas représentée pour alléger le graphe ; il en sera de même dans les figures suivantes.

1.2. Effet de Z_ECO

Lors de l'estimation de surfaces, RE est toujours supérieur à 1 si l'on compare les stratifications deux à deux en prenant celles avec Z_ECO comme références. Autrement dit, on augmente systématiquement la variance totale en rajoutant Z_ECO (bien que pas forcément toutes ses composantes, selon les stratifications). La différence est plus importante pour les stratifications qui utilisent la photo-interprétation, seule ou non. L'effet est plus variable pour les volumes, y compris pour les composantes de la variance (Figure 24).

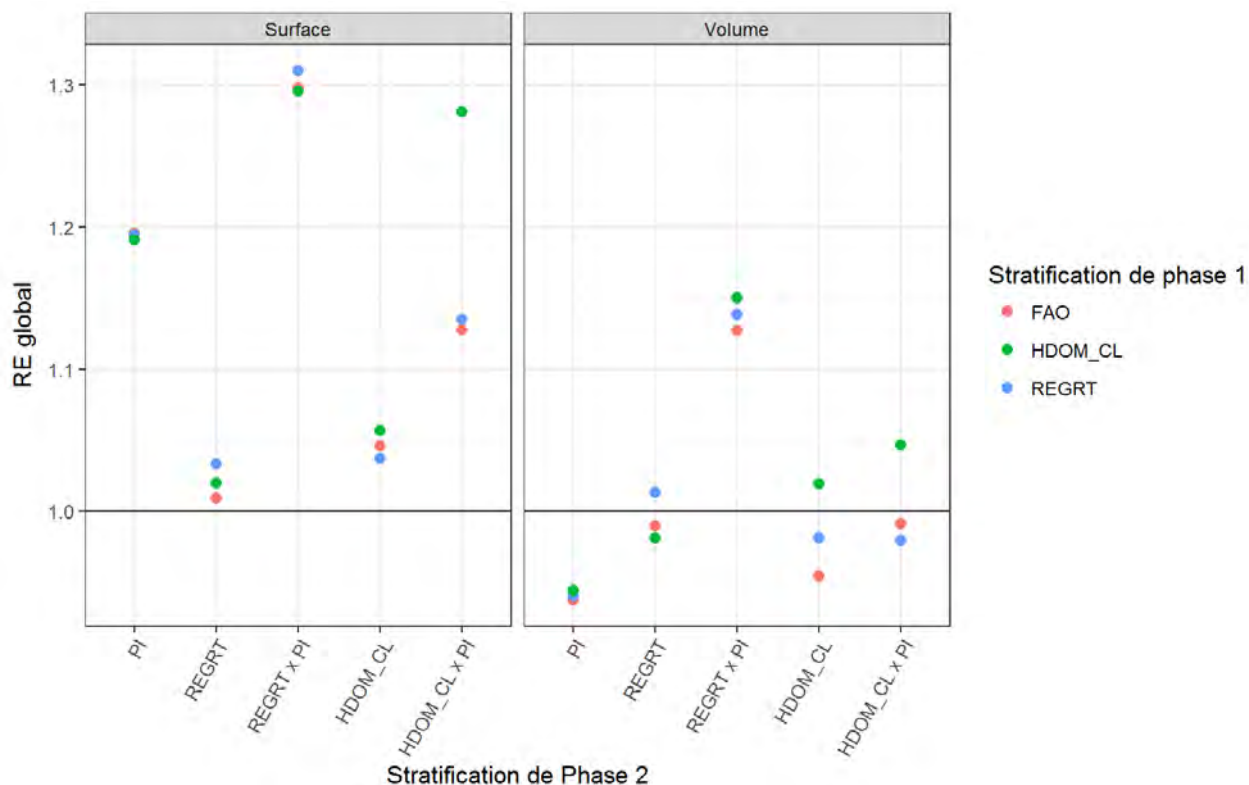


Figure 24 : facteur d'efficacité pour la variance globale lorsque les stratifications avec Z_ECO sont utilisées comme références (non représentées).

Les autres comparaisons ont malgré tout été menées avec toutes les stratifications, mais les résultats avec Z_ECO, généralement moins intéressants (et confirmant donc le résultat ci-dessus), ne seront qu'évoqués brièvement et ne seront pas montrés.

1.3. Effet des associations REGRT x PI et HDOM_CL x PI en phase 2

Par rapport à la photo-interprétation seule, REGRT x PI et, dans une moindre mesure, HDOM_CL x PI avec HDOM_CL en phase 1 améliorent la précision de l'estimation pour les surfaces. Pour les volumes en revanche, c'est HDOM_CL x PI qui présente les meilleures valeurs (surtout avec HDOM_CL en phase 1, le RE atteignant 1,6) (Figure 25).

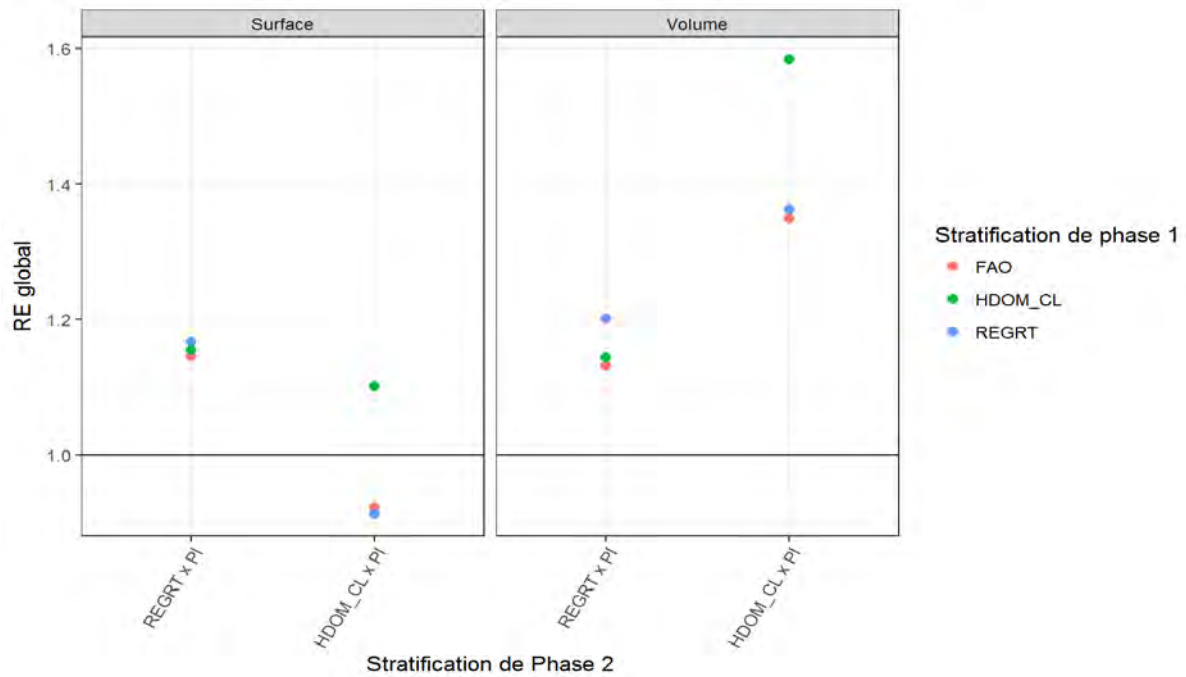


Figure 25 : RE global avec PI seule comme référence (sans Z_ECO).

Les composantes de la variance suivent ce patron, sauf la V_{SD} , qui suit celui des surfaces pour les deux variables (mais avec des valeurs de RE toujours supérieures à 1 pour les volumes). Si l'on rajoute Z_ECO, REGRT x PI améliore le résultat (et HDOM_CL le dégrade légèrement) par rapport à la stratification sans PI, quelle que soit la grandeur considérée.

En revanche, REGRT x PI et HDOM_CL x PI dégradent la variance par rapport à REGRT ou HDOM_CL seuls, REGRT pour les surfaces faisant exception (Figure 26). L'effet est plus important avec Z_ECO que sans, et peut être différent d'une composante à l'autre (la V_{SD} est en fait le plus souvent améliorée, surtout sur les volumes, mais la V_{ST} et la V_{HT} sont toujours dégradées).

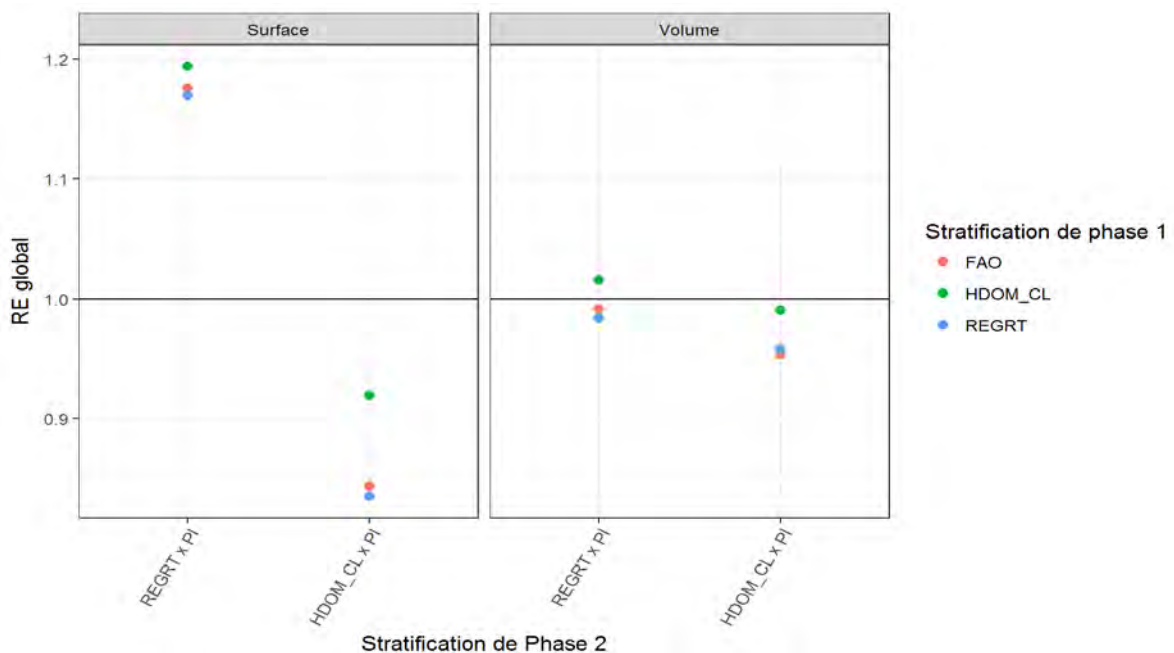


Figure 26 : RE global avec REGRT ou HDOM_CL seuls comme référence (sans Z_ECO).

1.4. Comparaison entre REGRT et HDOM_CL en phase 2

La variance globale pour les surfaces est meilleure pour HDOM_CL seul que pour REGRT seul (RE supérieur à 1 avec REGRT seul comme référence), mais moins bonne pour HDOM_CL x PI que pour REGRT x PI. Pour les volumes en revanche, elle est toujours meilleure pour HDOM_CL que pour REGRT (mais ce n'est pas le cas pour la V_{SD} ni la V_{ST}). Sans Z_{ECO} , les stratifications avec HDOM_CL en phase 1 présentent de plus un RE plus élevé, à phase 2 identique (Figure 27).

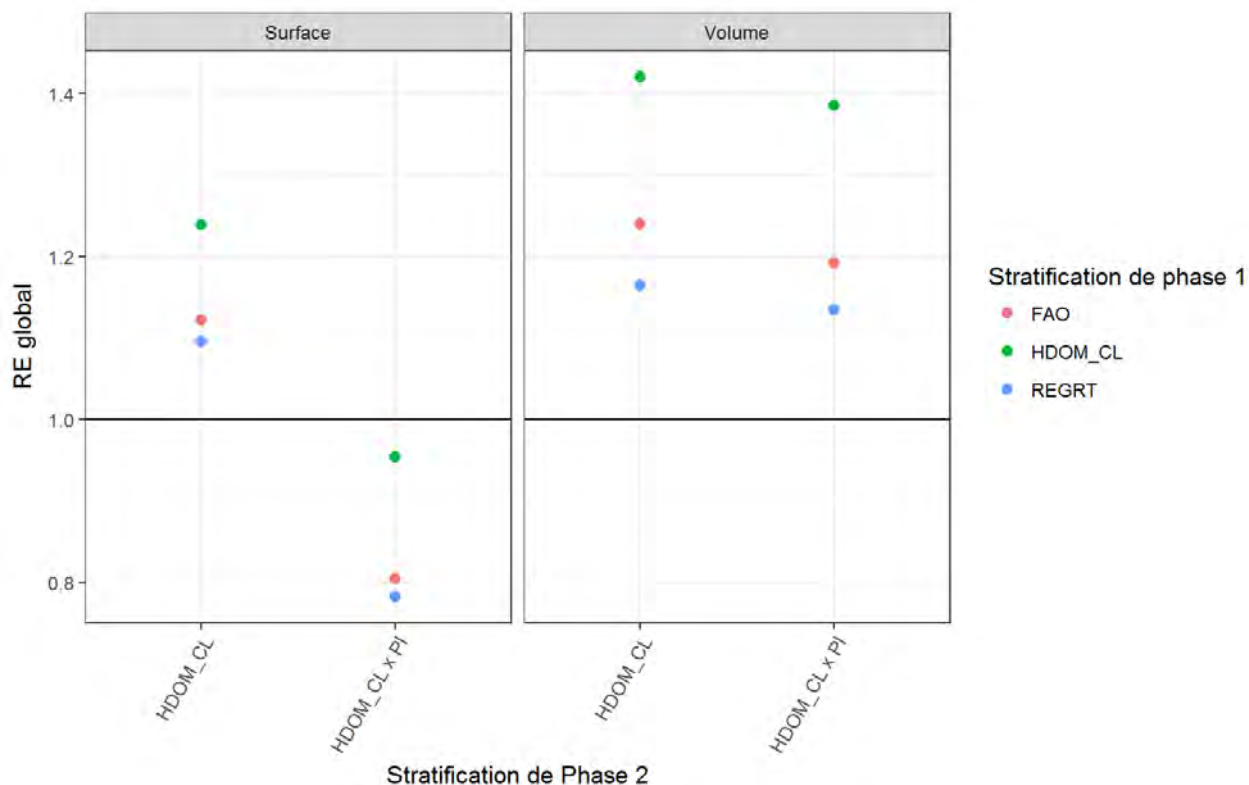


Figure 27 : RE global avec REGRT (seul ou REGRT x PI) comme référence (sans Z_{ECO}).

1.5. Comparaison entre phases 1

Les différences entre les trois phases 1 (FAO, REGRT ou HDOM_CL) sont faibles comparées aux différences entre phases 2. Les stratifications avec FAO ou REGRT en phase 1 tendent à être plus proches l'une de l'autre que de celles qui utilisent HDOM_CL, quelle que soit la phase 2 (logique puisque les types FAO sont reconstruits à partir de REGRT). Il est par ailleurs plus difficile de faire des observations générales avec Z_{ECO} que sans.

De manière générale, les stratifications avec les mêmes strates en phases 1 et 2 (avec ou sans photo-interprétation) sont les seules à présenter une plus faible variance que celles qui utilisent les types FAO en phase 1. L'écart entre phases 1 est plus important pour HDOM_CL (Figure 28). Ceci reste valable que l'on observe les surfaces ou les volumes (graphe très similaire, donc seules les données pour les surfaces sont présentées), et sur toutes les composantes.

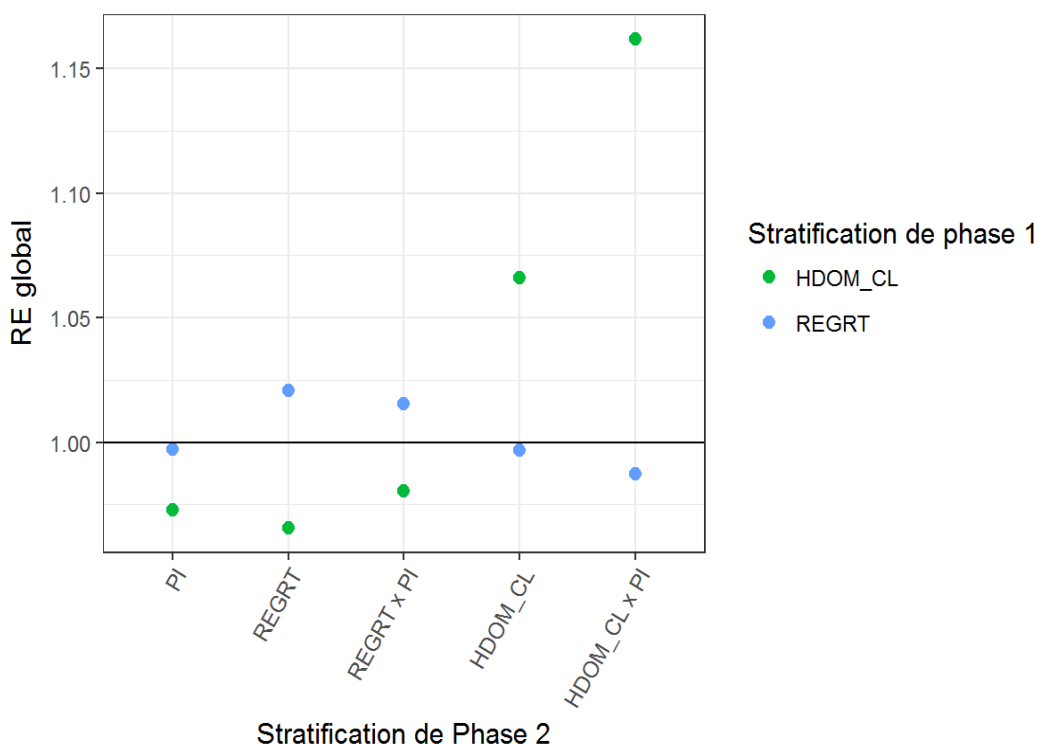


Figure 28 : RE global avec les stratifications utilisant le type FAO en phase 1 comme référence (surfaces).

1.6. Bilan des comparaisons systématiques

À la lumière de ces premiers résultats, certaines pistes peuvent déjà être abandonnées ou confortées : ainsi, l'ajout de Z_{ECO} augmente de manière générale la variabilité, surtout sur les surfaces, et rajoute un bruit de fond aux autres comparaisons effectuées. Les stratifications avec Z_{ECO} ne seront donc pas utilisées par la suite. Elles seraient probablement plus pertinentes sur un échantillon France entière, tandis qu'ici leur présence a nécessité des regroupements plus drastiques entre essences ou entre classes de hauteur.

On constate aussi que, du fait de l'interaction entre les phases de stratification 1 et 2, les meilleurs résultats sont souvent obtenus lorsque les mêmes strates sont utilisées en phases 1 et 2. Il semble donc préférable de ne pas tenter de nouveau le croisement REGRT x HDOM_CL en phase 2, même sans Z_{ECO} . Enfin, REGRT et HDOM_CL en phase 2 sont globalement meilleurs seuls qu'associés à la photo-interprétation, et HDOM_CL meilleur que REGRT sur les volumes.

Les analyses suivantes utilisent donc uniquement trois des stratifications mises en place : REGRT seul en phases 1 et 2, de même pour HDOM_CL, ainsi que la nomenclature « FAO » en phase 1 associée à la photo-interprétation en phase 2, qui sert de référence supplémentaire, car elle est similaire aux stratifications de référence de l'IFN tout en étant construite de la même manière que les stratifications testées (mêmes échantillons de points).

Par ailleurs, il a été confirmé lors de cette étape que, lorsque les stratifications utilisent les mêmes strates (basées sur REGRT ou HDOM_CL) en phases 1 et 2, la V_{ST} est nulle (puisque les surfaces des strates de phase 2 ne sont pas estimées mais connues à partir de la phase 1). Le RE correspondant ne pourra donc pas être calculé sur ces stratifications.

2. Comparaisons avec ventilation

La ventilation des résultats utilise des variables déjà disponibles dans OCRE, et qui sont définies soit au niveau de la placette de notations de terrain, soit au niveau de l'arbre, soit les deux. Les surfaces ne sont pas disponibles lorsqu'on ventile au niveau arbre. Ici, n'ont été utilisées que les quelques variables qui apparaissent les plus pertinentes. Étant donné que l'on cherche à réaliser des statistiques à un niveau infra-régional, donc déjà très fin, tous les critères de ventilation ne permettent pas forcément d'avoir suffisamment de points par sous-domaine. Pour la même raison, les ventilations n'utilisent ici qu'un seul de ces critères à la fois, mais il serait envisageable de les croiser si l'on disposait d'un jeu de données plus important.

Les stratifications utilisées et citées par la suite sont donc les suivantes :

1. **REGRT** : stratification avec uniquement les regroupements d'essence (REGRT) à la fois en première et deuxième phase
2. **HDOM_CL** : stratification avec uniquement les classes de hauteur (HDOM_CL) à la fois en première et deuxième phase
3. **référence simple** : stratification avec les « types FAO 2 » reconstitués à partir de REGRT en première phase, et la photo-interprétation en deuxième phase
4. **référence inventaire** : stratification par défaut de l'inventaire, avec les types FAO 2 officiels ventilés par département et par type de propriété en première phase, et la photo-interprétation ventilée par département et par type de propriété en deuxième phase.

Les grandeurs que l'on cherche à estimer sont de nouveau les surfaces et les volumes, auxquelles s'ajoute la production de bois en volume.

2.1. Par type de propriété forestière

Il s'agit des mêmes types de propriété qui sont employés dans la construction de la référence inventaire (et dans les restitutions de résultats d'inventaire). Ils correspondent à trois classes : forêts domaniales, autres forêts publiques (désignées simplement comme « publiques » dans la suite du document), et forêts privées. Les forêts publiques, communales notamment, représentent près de la moitié des forêts alsaciennes (49,6 %), le reste se répartissant à peu près équitablement entre forêts domaniales (24,1 %) et privées (26,3 %). Cette situation est particulière au Nord-Est et à certaines régions de montagne ; dans la plupart des autres régions, les forêts privées dominent. Lorsqu'on analyse les données, cela se traduit par un CV nettement plus faible pour la catégorie « forêts publiques » (4 %, contre environ 7 % pour les autres catégories).

Quelle que soit la grandeur mesurée, HDOM_CL et REGRT représentent une amélioration par rapport à la référence simple, mais une dégradation par rapport à la référence inventaire (c'est moins le cas pour les forêts privées). HDOM_CL présente de meilleurs RE que REGRT pour la production dans les forêts domaniales, et surtout pour les volumes dans tous les types de forêts (Figure 29).

Cependant, les stratifications testées présentent finalement peu d'intérêt pour ce critère de ventilation : la référence inventaire intègre les types de propriété, il est donc logique qu'elle donne des résultats meilleurs ou équivalents à ceux des stratifications testées dans tous les cas.

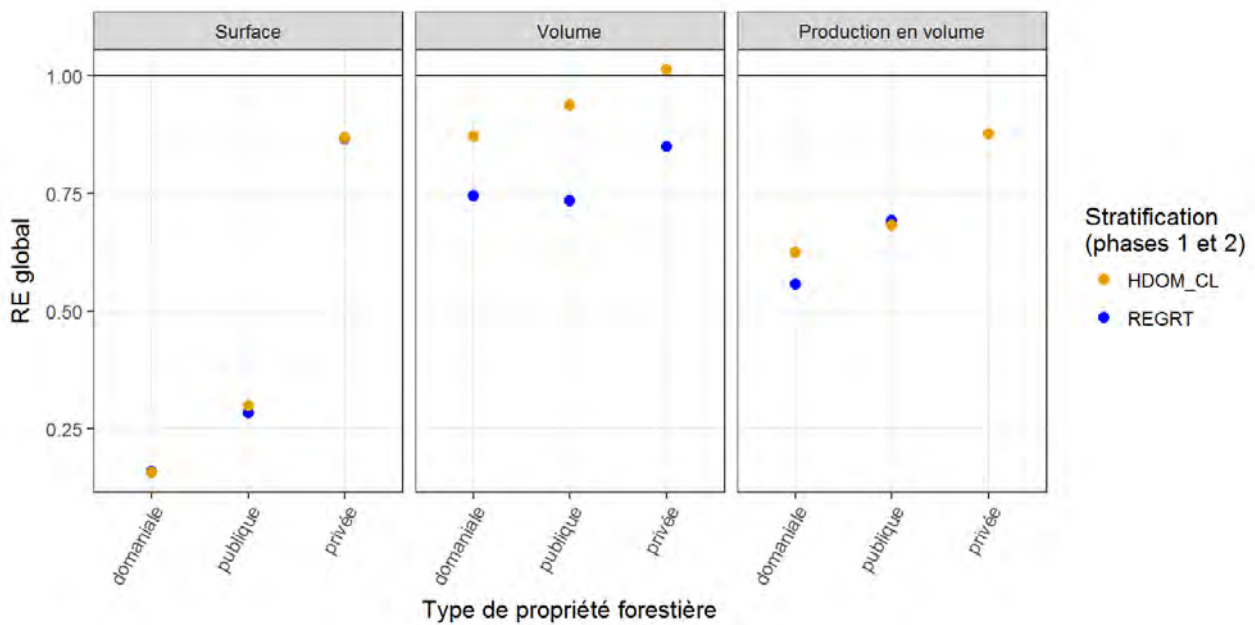


Figure 29 : RE global pour les stratifications REGRT et HDOM_CL sélectionnées par rapport à la référence inventaire, données ventilées par type de propriété.

2.2. Par essence ou groupe d'essences

Bien que l'essence soit un critère de ventilation important, il pose le problème d'un découpage risquant d'être trop fin. Plusieurs niveaux de regroupement ont donc été testés :

- essence principale, c'est-à-dire celle qui présente le plus fort taux de couvert au niveau du point, sans regroupement
- composition du peuplement en feuillus / résineux, purs ou mélangés (cinq classes, plus « inclassable »)
- simple opposition feuillus / résineux

2.2.1 Par essence principale

Cette ventilation a été tentée au niveau placette (28 essences) et au niveau arbre (33 essences). On ne dispose que de peu de données pour les moins fréquentes d'entre elles (parfois un seul point), aussi les CV correspondants atteignent des valeurs extrêmes, jusqu'à plus de 400 % si l'on ventile au niveau placette. Les écarts de CV et le nombre d'essences rendent le graphe peu lisible. Dans le cas de la ventilation au niveau placette pour les surfaces, on peut distinguer un groupe d'essences plus fréquentes (au moins 11 points), avec un CV d'au plus 31 % (Figure 31), également visible de manière moins nette pour les volumes et la production, ainsi qu'avec la ventilation au niveau arbre. Les graphes qui suivent (figures 32 à 35) montrent donc uniquement ce sous-ensemble d'essences, même si les analyses ont aussi été menées avec toutes les essences.

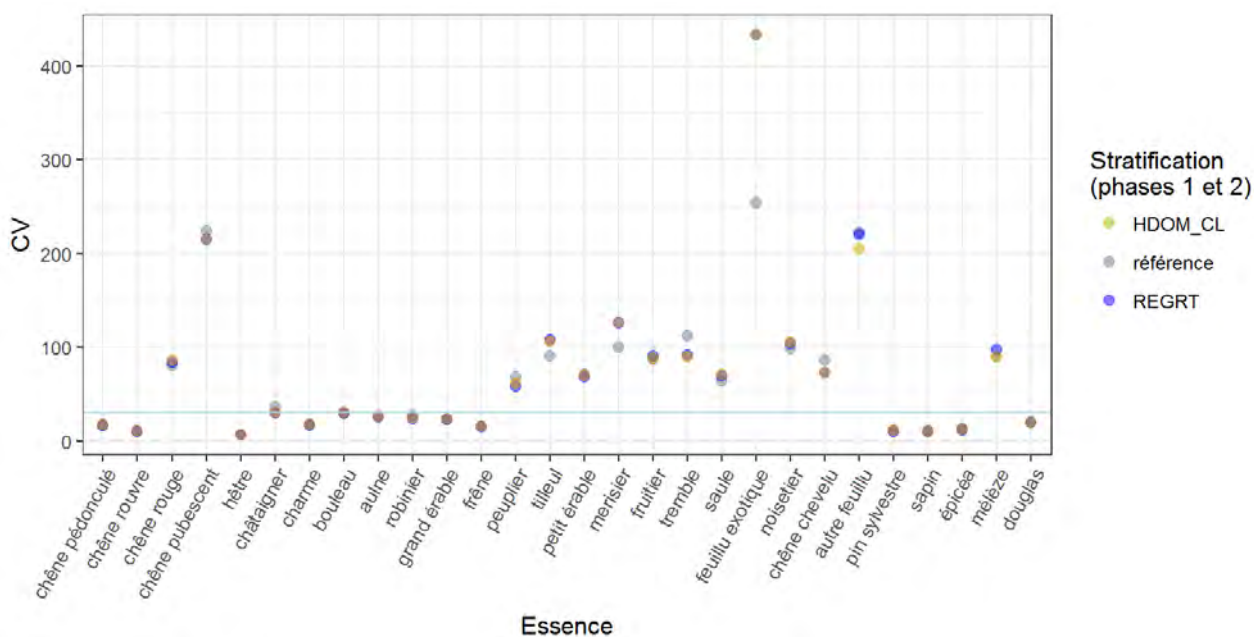


Figure 31 : CV avec les données ventilées par essence principale au niveau placette (surfaces). En bleu, la limite à 31 %.

Pour l'estimation des **surfaces**, REGRT et HDOM_CL sont supérieures ou équivalentes aux deux références pour la grande majorité des essences, dont les essences les plus fréquentes en Alsace. Les deux références se comportent de manière similaire. Pour les essences les plus rares, le RE global atteint quelquefois des valeurs extrêmes (plus de 600 % pour « autres feuillus » avec la référence simple, 400 % avec la référence inventaire). REGRT est systématiquement meilleure sur les conifères, les écarts entre REGRT et HDOM_CL étant plus importants avec la référence inventaire. Pour les autres essences, la hiérarchie entre les deux stratifications testées ne présente aucune constante. On peut remarquer le cas du charme, seule essence relativement fréquente pour laquelle les RE sont toujours inférieurs à 1 : c'était déjà le cas lors de l'étude précédente [5] (Figure 32).

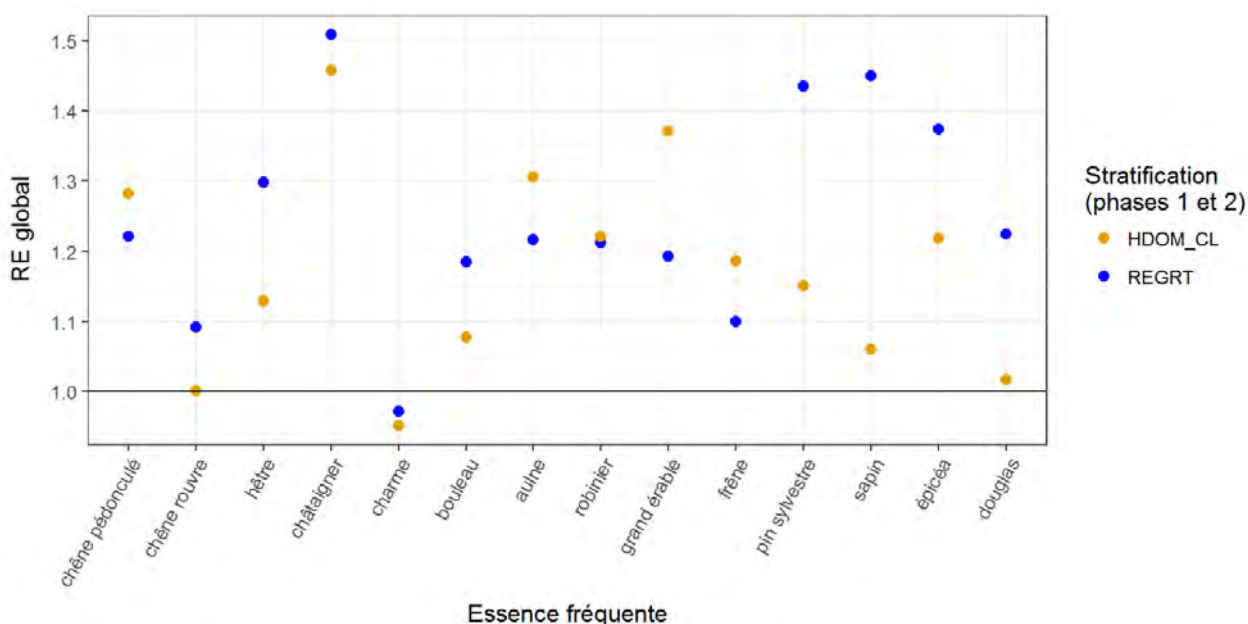


Figure 32 : RE global pour les surfaces par essence avec la référence inventaire, ventilation par placette.

Pour l'estimation des **volumes** par placette en revanche, le RE global tend à être plus bas, et est plus souvent que pour les surfaces inférieur à 1 pour au moins une des deux stratifications, notamment sur des essences moyennement fréquentes (chêne pubescent, bouleau, aulne, robinier...). Avec la ventilation par arbre, c'est même une majorité des essences qui sont dans ce cas. REGRT est de nouveau supérieure à HDOM_CL sur les conifères relativement fréquents (Figure 33).

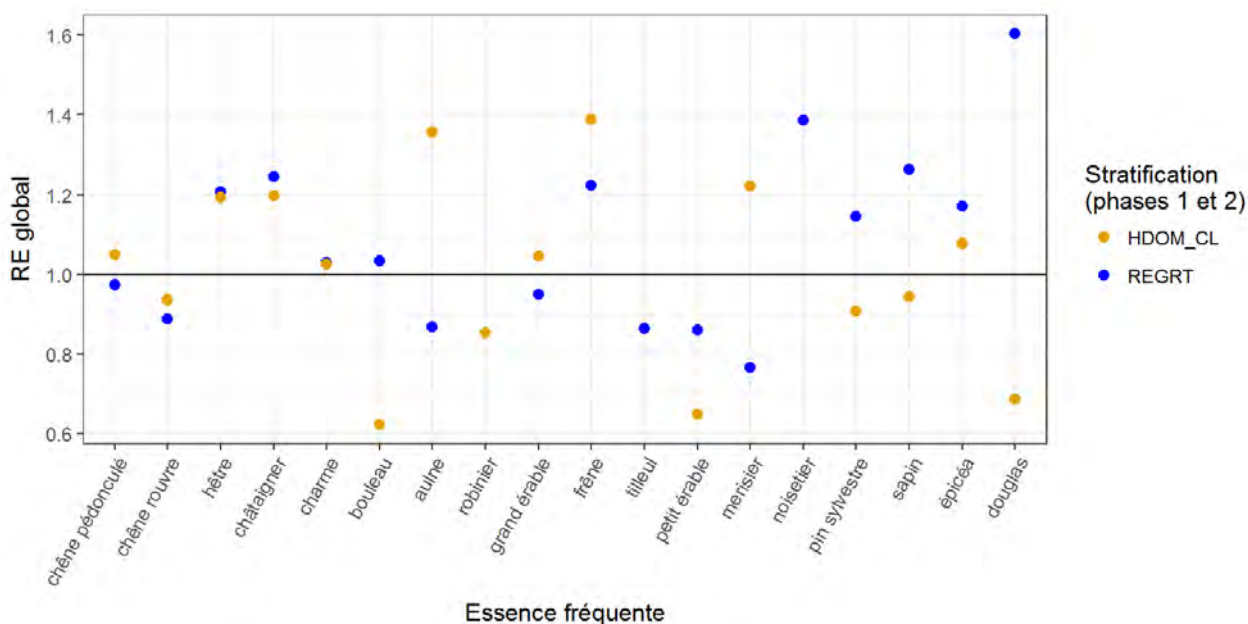


Figure 33 : RE global pour les volumes par essence avec la référence inventaire, ventilation par arbre.

L'écart entre stratifications est plus faible pour la V_{HT} avec la ventilation au niveau placette, rendant plus lisibles les différences entre essences (Figure 34). Avec la ventilation au niveau arbre, HDOM_CL est systématiquement supérieure à REGRT pour la V_{SD} , mais la proportion de V_{SD} est très faible (moins de 5 % de la variance totale quelle que soit l'essence).

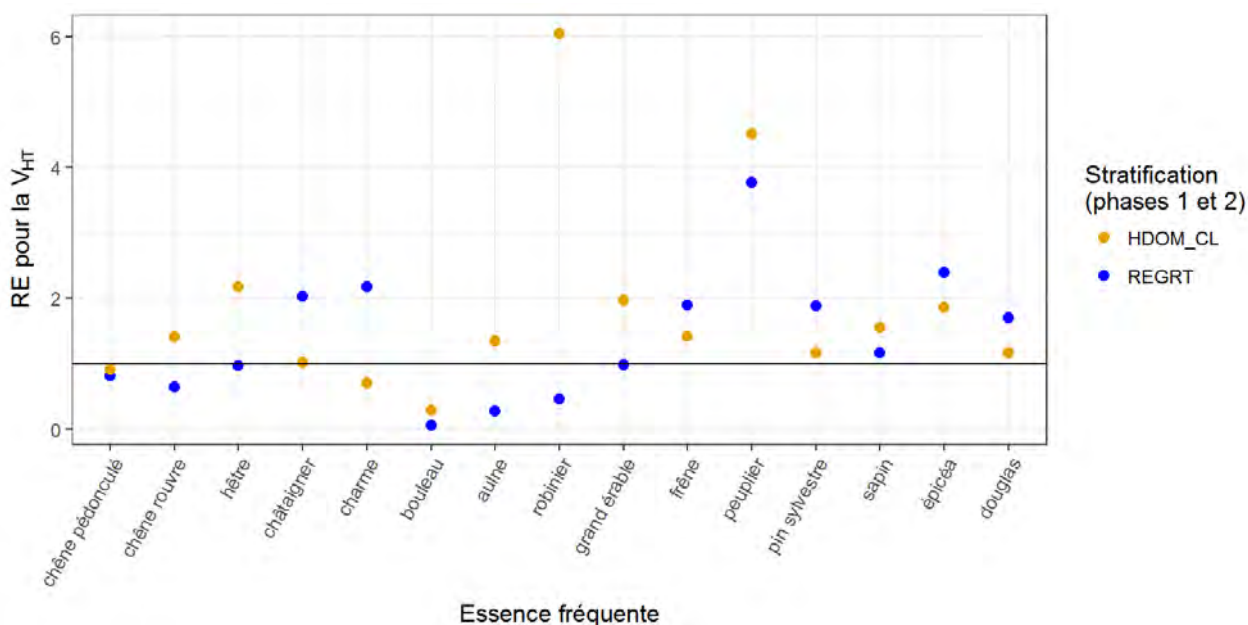


Figure 34 : RE pour la V_{HT} des volumes par essence avec la référence inventaire, ventilation par placette.

Pour la **production en volume**, les valeurs de CV sont plus hautes que pour les autres variables (de 13 % pour les essences les plus fréquentes à plus de 600 % pour les feuillus exotiques avec la ventilation par placette, et 300 % avec la ventilation par arbre). La majorité des RE globaux sont inférieurs à 1 avec la référence simple, mais pas ceux des essences les plus communes. Avec la référence inventaire, davantage d'essences ont un RE global positif pour au moins une stratification. REGRT est supérieure à HDOM_CL pour les conifères fréquents, et légèrement supérieure pour les chênes et le hêtre avec la ventilation par arbre (les différences sont similaires mais un peu plus visibles sur la V_{HT} ; la V_{SD} présente des résultats différents).

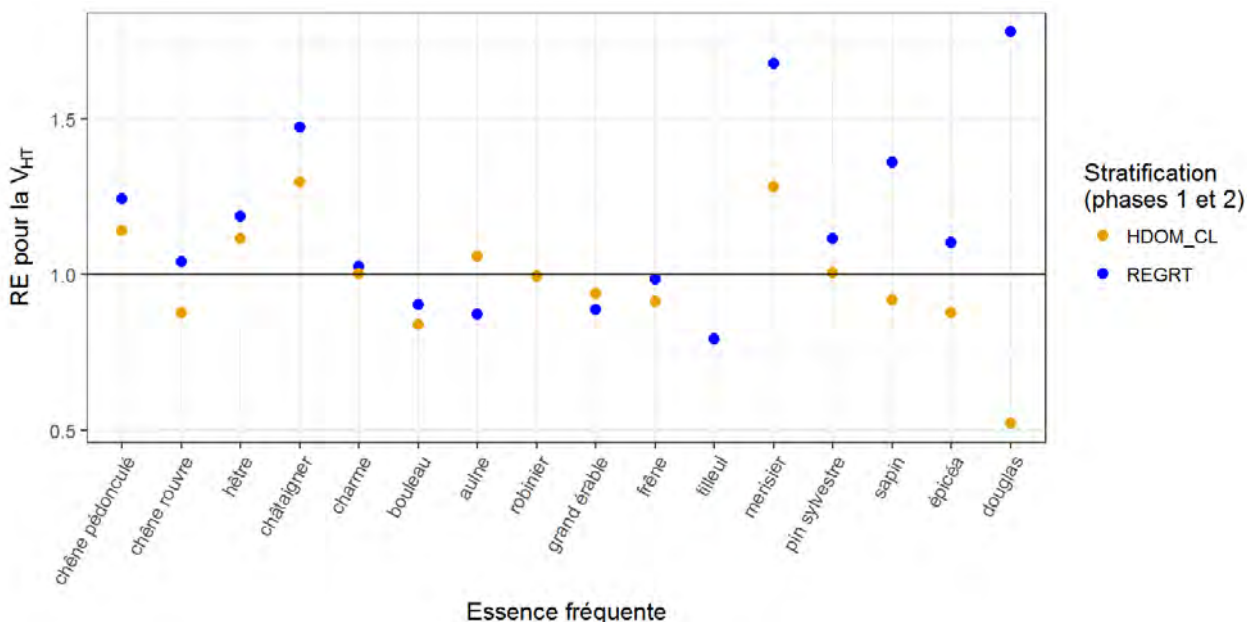


Figure 35 : RE pour la V_{HT} de la production en volume par essence avec la référence inventaire, ventilation par arbre.

Le résultat est donc très variable en fonction des essences, mais globalement intéressant sur les surfaces avec les deux stratifications testées. Pour les volumes et la production en volume, les résultats sont plus variables. Dans tous les cas, ils sont probablement peu fiables sur de nombreuses essences trop peu représentées. Il est difficile de préciser laquelle des deux stratifications est la meilleure en général, mais REGRT est systématiquement meilleure sur les conifères relativement fréquents. L'écart est plus important pour les conifères que pour les autres essences, et encore plus important pour la production en volume.

2.2.2. Par cinq classes de composition de peuplement

Étant donné les résultats moyens pour le volume et la production avec les essences, on souhaite vérifier si des ventilations moins fines sont plus pertinentes pour les stratifications testées. Le critère « composition du peuplement » offre une ventilation au niveau placette uniquement, avec six catégories : une seule essence feuillue, plusieurs essences feuillues, une seule essence conifère, plusieurs essences conifères, feuillus et conifères en mélange, et « inclassable ».

Le CV est beaucoup plus élevé pour la classe « inclassable » (14 points seulement) et dans une moindre mesure « plusieurs conifères », surtout pour les volumes (l'écart de CV entre la classe « inclassable » et les autres y atteint un facteur 10). Par conséquent, le RE pour cette classe se comporte différemment des autres classes sur toutes les variables.

Pour les **surfaces**, le RE global est toujours supérieur à 1, et REGRT est supérieure à HDOM_CL, sauf pour « inclassable ».

Le RE global pour les **volumes** est supérieur ou égal à 1, sauf pour REGRT et « inclassable », où il est même très bas avec la référence simple (0,4 pour REGRT). HDOM_CL est supérieure à REGRT sur les feuillus et « inclassable », et inférieure sur les résineux et mélanges. En revanche, pour la V_{SD} , REGRT est supérieure partout, sauf sur « inclassable ».

La **production en volume** ne présente un RE global inférieur à 1 que pour HDOM_CL sur conifères (et pour « inclassable » avec la référence simple). HDOM_CL est équivalente à REGRT sur les feuillus. La V_{HT} présente des RE similaires mais plus faibles ; en revanche, la V_{SD} est plus similaire à celle observée pour les surfaces.

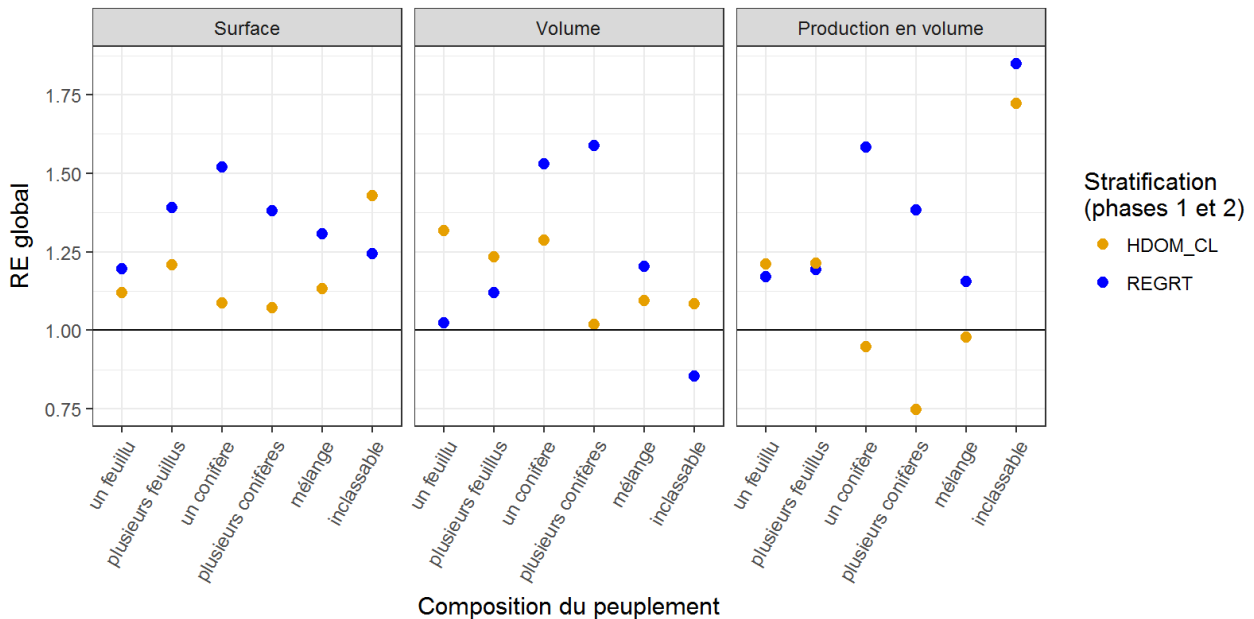


Figure 36 : RE global par catégorie de composition du peuplement avec la référence inventaire.

Avec une décomposition limitée à six classes, on retrouve donc de nouveau les meilleurs résultats pour les surfaces, mais aussi des résultats intéressants pour les autres grandeurs, notamment pour les volumes. En ne prenant pas en compte la classe « inclassable », il n'y a que pour la production en volume avec HDOM_CL sur conifères que l'on observe une dégradation de la variance globale.

2.2.3. Par opposition feuillus / conifères

Les feuillus représentent 67,4 % de la superficie en Alsace (mais « seulement » 58,7 % du volume) et ce déséquilibre se traduit par un CV plus élevé pour les conifères (4,5 % à 5,5 % pour les surfaces, contre 3 % pour les feuillus). La ventilation a été effectuée au niveau placette et au niveau arbre. Les deux références présentent peu de différences.

Pour les **surfaces**, REGRT présente des RE globaux élevés (environ 1,6 sur feuillus et 2,15 sur conifères), nettement supérieurs à HDOM_CL, surtout sur conifères (1,15 à 1,2).

Pour les **volumes**, les RE globaux restent supérieurs à 1, mais HDOM_CL présente les meilleurs résultats sur feuillus, REGRT sur conifères (avec un RE global atteignant 2 ou presque). REGRT est par contre meilleure sur les deux groupes pour la V_{HT} . Avec une ventilation au niveau arbre, la

différence entre les deux stratifications se trouve réduite sur feuillus et augmentée sur conifères, si bien que sur cette catégorie HDOM_CL n'est plus différente de la référence inventaire.

Cet écart entre stratifications plus important pour les conifères est exacerbé pour la **production en volume** : alors que le RE global est proche de 1,2 pour les feuillus avec les deux stratifications, il atteint environ 1,8 sur conifères pour REGRT (1,4 avec la référence simple), mais est peu différent de 1 avec HDOM_CL, quel que soit le mode de ventilation. Pour la V_{HT} en revanche, les valeurs de RE sont plus faibles et HDOM_CL est nettement supérieure à REGRT (Figure 37).

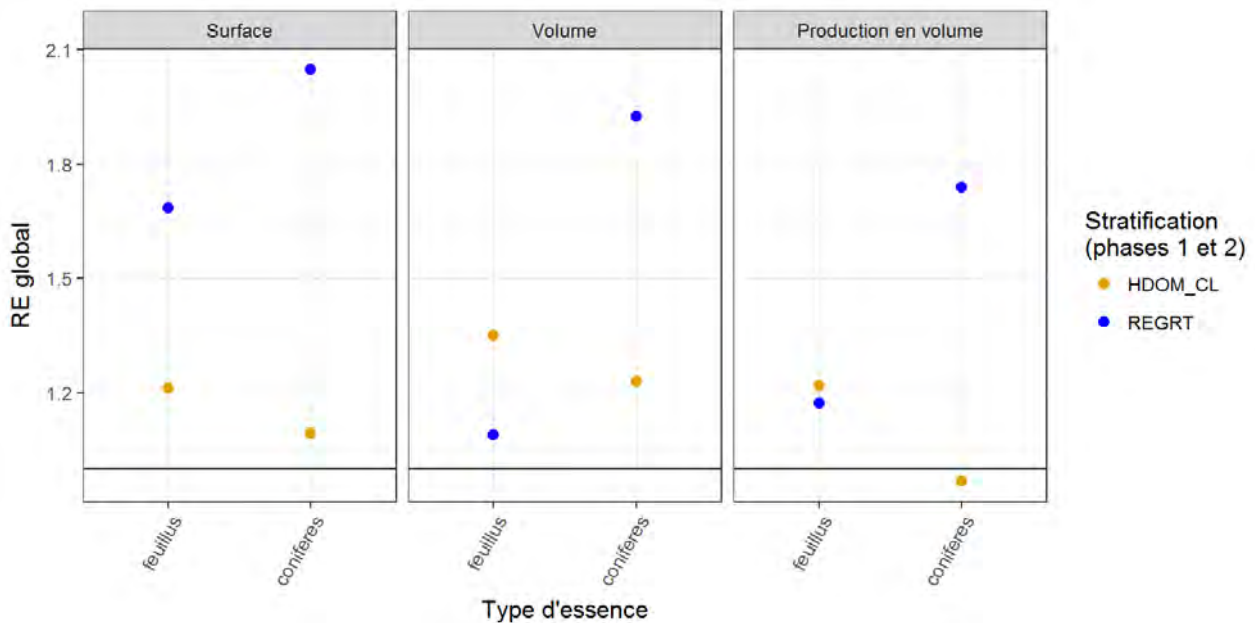


Figure 37 : RE global par grand type d'essence avec la référence inventaire, ventilation au niveau placette.

REGRT présente donc ici des résultats très positifs sur les surfaces. Pour les volumes, les deux stratifications sont intéressantes, la meilleure dépendant du type d'essence, donc on peut être amené à privilégier l'une ou l'autre en fonction du peuplement. Ici, vu la plus faible différence sur les feuillus, on pourrait plutôt choisir REGRT qui améliore davantage les résultats sur conifères. Pour la production, les différences sont encore plus importantes sur conifères, HDOM_CL ne présentant pas forcément d'amélioration par rapport aux références.

2.3. Par classes d'exploitabilité

Ce critère ventilé au niveau placette synthétise la plus ou moins grande facilité à exploiter la parcelle pour son bois (distance de débardage, accessibilité du site aux engins d'exploitation, pente...), avec quatre classes allant de « très facile » à « difficile ». La classe « moyennement difficile » est en fait celle qui présente le moins de points (36), le CV est donc beaucoup plus élevé.

Le RE global est supérieur ou égal à 1 quelles que soient la stratification, la référence et la grandeur évaluée, avec deux exceptions : REGRT pour le volume et pour la production en volume, sur la classe « moyennement difficile » uniquement ; mais vu le fort CV sur cette classe, il est sans doute préférable de ne pas trop prendre en compte ce résultat.

Pour les **surfaces**, le RE atteint toujours au moins 1,1. Pour les trois premières classes, HDOM_CL et REGRT sont très similaires ; REGRT est légèrement supérieure à HDOM_CL pour la classe « difficile ». Pour les **volumes** et la **production en volume**, HDOM_CL est un peu supérieure à REGRT pour « très facile » et « facile » ; c'est l'inverse pour la production en volume et « difficile ». Les résultats sont un peu inférieurs pour la V_{SD} .

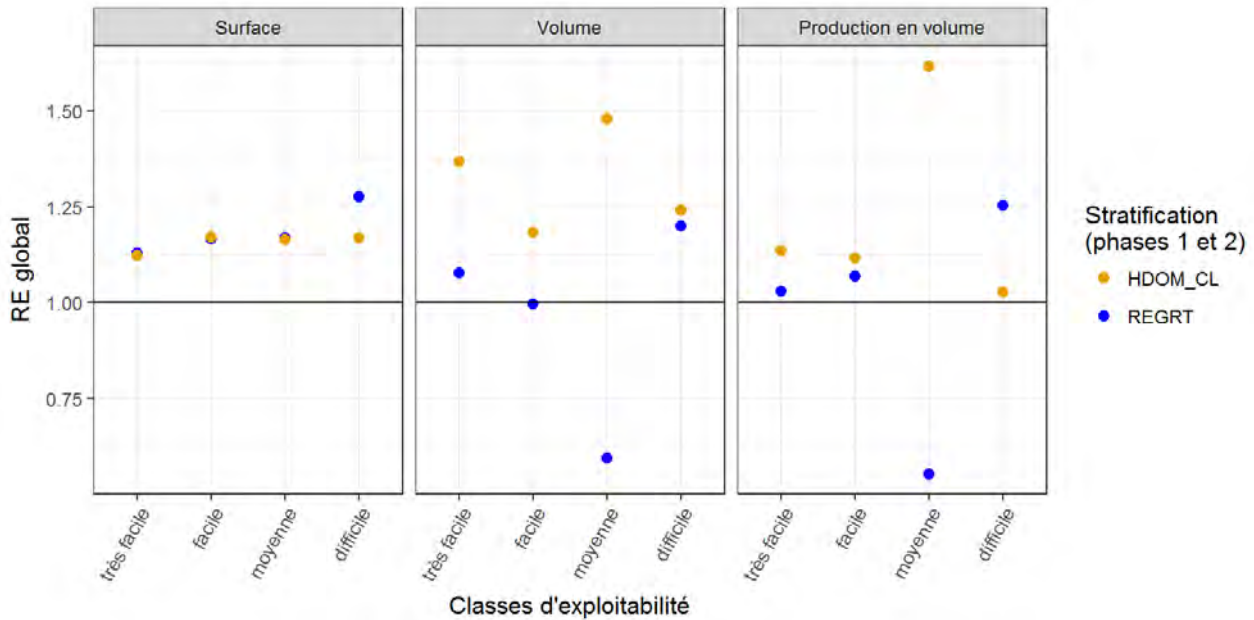


Figure 38 : RE global pour les volumes par classe d'exploitabilité avec la référence inventaire.

Les deux stratifications améliorent donc globalement les résultats par rapport à la référence inventaire. Il est par contre difficile de les départager, sauf pour les volumes où HDOM_CL est meilleure. Pour les volumes et la production, l'écart entre les deux stratifications est très important pour la classe « moyennement difficile », mais ce résultat est peu fiable.

2.4. Par classe de diamètre

Ce critère ventilé au niveau arbre existe avec différents classements. Celui qui a été retenu sépare les arbres en petit bois (diamètre à 1,30 m de haut compris entre 7,5 cm et 22,5 cm, sachant que les arbres plus petits ne sont pas recensés), moyen bois (22,5 à 37,5 cm de diamètre), gros bois (37,5 à 52,5 cm de diamètre) et très gros bois (plus de 52,5 cm de diamètre). Le CV est le plus faible pour les gros bois, suivis des très gros bois et des moyens bois, et le plus élevé pour les petits bois. Pour la production en volume, le CV est extrêmement faible pour les gros et très gros bois, souvent moins de 0,5 %, alors qu'il est aux alentours de 6 % pour les petits et moyens bois.

Le RE global est supérieur à 1 pour pour les deux stratifications et toutes les catégories. Pour les petits et moyens bois, HDOM_CL est légèrement supérieure à REGRT (volume) ou très légèrement inférieure ou équivalente (production en volume). Pour les gros et très gros bois en revanche, HDOM_CL est nettement supérieure à REGRT. Le RE pour la V_{HT} est plus faible avec la référence inventaire, surtout pour REGRT pour les volumes et les deux stratifications pour la production en volume (il ne dépasse pas 1 pour les petits bois). Pour la V_{SD} au contraire (12 % maximum de la variance totale), il est inférieur à 1 avec la référence simple.

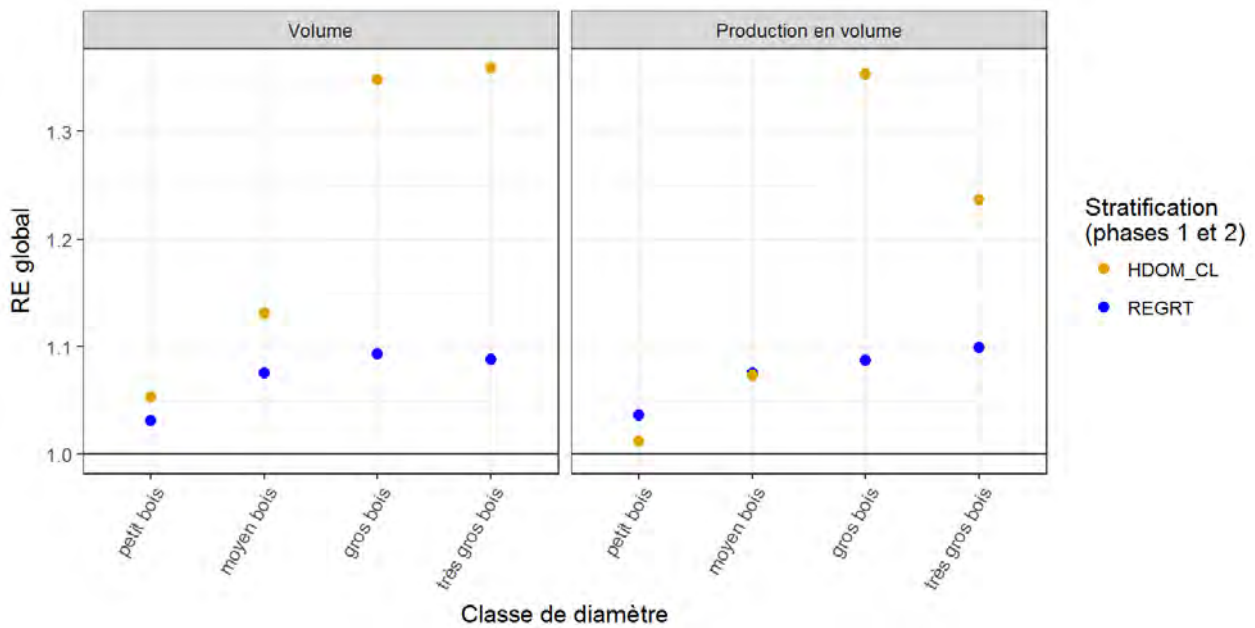


Figure 39 : RE global par classe de diamètre avec la référence inventaire, ventilation par arbre.

Les deux stratifications sont globalement supérieures aux références, et au moins équivalentes sur les composantes de la variance. C'est notamment le cas pour les gros et très gros bois, où, de plus, HDOM_CL est très au dessus de REGRT, atteignant un RE global relativement élevé (le plus souvent supérieur à 1,3) même face à la référence inventaire.

V. Discussion

1. Limites des données disponibles et de leur traitement

Bien que fournissant des possibilités nouvelles de croisement d'informations d'origine et de nature différentes, les couches d'informations géographiques ne sont pas forcément faciles à exploiter. Tout d'abord, le positionnement des points d'inventaire avec la BD Ortho® n'est précis qu'à 3 à 4 m près. Les points peuvent donc en réalité se trouver hors de leur strate théorique. Par ailleurs, les observations de terrain s'effectuant sur des placettes et non des points, elles peuvent être à cheval sur une limite de strate ou même de milieu. De même, plus le décalage temporel entre les données de stratification et les observations est grand, plus la variabilité est augmentée par le changement d'état de certaines parcelles [1].

Cependant, dans le cadre du présent travail, le problème le plus concret est celui de la qualité du MNT, dont on a vu qu'il se répercutait sur le MNH (le MNE est beaucoup plus uniforme). Bien que l'Alsace ait été choisie comme zone de travail, entre autres, du fait qu'elle présentait davantage de données de qualité que la plupart des autres régions, le MNT lidar y représente moins de la moitié de la surface de la région. De plus, il n'est pas de la même nature partout : à l'est du Haut-Rhin et dans la majorité de la plaine du Rhin, ainsi que dans les forêts du Haut-Rhin, il s'agit de données acquises en interne, mais dans les vallées il est issu d'acquisitions externes. Dans tous les cas, la définition n'est pas la même partout, avec une densité théorique allant de 1 à 7 points au m². Le

MNT issu de la BD Alti est également de qualité très inégale, puisqu'il peut être issu d'une saisie photogrammétrique ou de la vectorisation, c'est-à-dire la numérisation manuelle ou par scannage des courbes de niveau et points cotés, dans les deux cas avec des pas variables [6].

La méthode utilisée ici pour obtenir un MNH complet a permis de privilégier les données lidar, mais dans une certaine mesure : seules ont pu être conservées les tuiles lidar complètes, ce qui a conduit à abandonner une grande quantité de ces données. De plus, des tuiles lidar jouxtent des tuiles Alti. Idéalement, il aurait fallu reconstituer des tuiles complètes en conservant toutes les données lidar, et ménager des zones de raccord entre les deux types de données, mais cela aurait largement excédé le cadre de ce stage, d'autant qu'il ne s'agissait que d'une étape préparatoire. Aucun effet de bord n'ait été observé, ce qui ne prouve pas qu'ils n'existent pas, mais leurs effets éventuels sont probablement limités par les traitements appliqués ensuite au MNH.

2. Stratifications utilisées

Les résultats obtenus sont largement tributaires des choix effectués lors de la mise en place des stratifications et, pour les données ventilées, des stratifications retenues. Comme pour la stratification par défaut de l'inventaire, celles-ci ne peuvent fournir des surfaces exactes que suivant les découpages qui sont intégrés à leur construction. Cependant, il ne serait pas possible d'intégrer tous les critères envisageables dans la construction des stratifications, un découpage trop fin faisant de nouveau perdre en précision faute d'un nombre suffisant de points par strate.

Dans le cas présent, la taille du jeu de données n'a tout simplement pas permis de construire les stratifications les plus complexes envisagées. En effet, travailler sur la seule région Alsace était nécessaire pour pouvoir mettre en place un modèle de hauteur dans un délai raisonnable, mais cela constitue une limite en termes de taille d'échantillon, surtout si l'on veut pouvoir ventiler les résultats. Les classes de hauteur auraient aussi pu être moins nombreuses, ce qui aurait peut-être été le cas si elles n'avaient pas été définies manuellement.

Les nombreux regroupements entre strates lors de la mise en place des stratifications montrent déjà que les strates créées sont un peu trop fines. Les découpages trop fins du MNH classé peuvent aussi avoir posé problème au niveau des pixels de bord lors de la création des strates d'origine par intersection entre le MNH (raster) et les regroupements d'essences (vecteur). Ceci pourrait expliquer les incohérences constatées à cette étape sur la surface de la région.

Les ventilations basées sur les essences montrent bien que le nombre de points d'échantillonnage est une limite dans le cas présent : avec une ventilation directement par essence principale (une trentaine d'essences), les stratifications testées sont intéressantes par rapport à la référence pour estimer les surfaces, mais elles dégradent les résultats presque aussi souvent qu'elles les améliorent pour les volumes et la production en volume. Par type de peuplement (six classes), les résultats sont meilleurs, avec une réduction de la variance dans presque tous les cas. Quand on se limite à deux classes seulement, feuillus et conifères, la variance est toujours réduite ou (rarement) inchangée.

De manière générale, des stratifications qui se sont avérées infaisables ou peu pertinentes ici car trop compliquées pourraient se montrer beaucoup plus pertinentes sur un échantillon France entière. C'est notamment le cas de celles qui incorporent Z_ECO. De même, les ventilations n'utilisent ici qu'un seul critère à la fois, mais il serait envisageable de les croiser. Il faudrait pour cela disposer d'un jeu de données plus important, correspondant à une densité d'échantillonnage renforcée. Ce pourrait être le cas dans le cadre d'une étude locale à façon comme l'inventaire en propose.

3. Interprétation des résultats

Les résultats ne dépendent cependant pas uniquement de la taille de l'échantillon et de la finesse de la stratification ou de la ventilation. En effet, si l'on étudie l'influence du nombre de points par essence, on constate qu'il existe bien une relation entre celui-ci et le CV, mais pas entre le nombre de points (ou le CV) et le RE. De plus, des ventilations avec un nombre de classes similaire ne donnent pas forcément les mêmes résultats. Les différences constatées entre critères de ventilation et, pour un même critère, entre grandeurs mesurées montrent bien l'intérêt d'utiliser des stratifications spécifiques pour optimiser des types de résultats spécifiques.

Dans le cas de la stratification basée sur les classes de hauteur, il n'est pas surprenant d'obtenir des résultats intéressants pour les classes de diamètre : en effet, le diamètre influence directement le volume de bois. Or, ces classes de hauteur ont été bâties à partir d'une métrique fortement corrélée au volume. De même, c'est souvent lorsqu'on estime des volumes ou des productions en volume que cette stratification est supérieure à celle basée sur les regroupements d'essences. C'est le cas par exemple lorsqu'on ventile par type de propriété forestière (où par ailleurs la référence inventaire reste la meilleure stratification), ou par difficulté d'exploitation.

Les regroupements d'essence montrent pour leur part un intérêt lorsqu'on ventile par classe de diamètre (quoique moins que les classes de hauteur), par opposition feuillus / résineux et par composition du peuplement, surtout pour les surfaces. C'est aussi le cas avec une ventilation par essence, mais seulement pour les conifères. Dans le premier cas, on conçoit aisément qu'il y ait une relation puisque toutes les essences n'ont pas le même potentiel de croissance. Les autres ventilations concernent toute l'essence, donc il est normal que les regroupements d'essence améliorent les résultats. Dans le cas des conifères, il faut remarquer que les regroupements ont souvent été importants faute de points, ce qui pourrait expliquer pourquoi ils se comportent tous de manière semblable.

VI. Conclusion générale

Ce stage avait un objectif à la fois méthodologique et en termes de résultats avec les données d'inventaire : d'une part, réaliser et documenter toutes les étapes allant des données de hauteur brutes à leur utilisation en stratification ; d'autre part, mettre en place des stratifications nouvelles et vérifier si elles permettent d'obtenir des résultats plus précis qu'avec la stratification déjà utilisée.

Les données de hauteur initiales (toutes au format raster) présentaient la particularité d'être très lourdes et impossibles à vérifier de manière exhaustive à l'avance : près de deux To incluant un modèle numérique de hauteur lidar partiel, un MNH issu de la BD Alti complet, mais de qualité inférieure, et des modèles d'élévation par département avec des tuiles de bordure incomplètes, tous dans une définition différente, et parfois non calés les uns par rapport aux autres. De nombreux problèmes sont donc apparus au cours du traitement, mais ils ont été pris en compte par les scripts R qui ont permis, au final, d'automatiser entièrement les opérations menant des données de base au MNH que nous souhaitons utiliser (cf. les scripts en annexe pour plus de détails).

Cette méthode peut donc être appliquée telle quelle à des jeux de données composés de manière identique, ou adaptée relativement simplement aux particularités de données un peu différentes. Si un travail ultérieur similaire devait être mené, dans une autre région par exemple, cette méthode devrait permettre au minimum de détecter et corriger plus rapidement les difficultés qui se

présenteraient. Le présent travail a aussi permis de mieux appréhender la qualité et les limites de MNT de différentes sources, ainsi que des paramètres concrets comme le temps de calcul et l'espace disque nécessaires. Certains éléments spécifiques ont de plus déjà été intégrés à de futures utilisations en devenant partie intégrante du package R à usage interne inventR.

En ce qui concerne la mise en place de stratifications, la méthode en elle-même n'est pas nouvelle, mais n'avait pas été étendue à des croisements de données systématiques. De plus, l'inventaire travaille habituellement avec des échantillons France entière. Ici, le jeu de données restreint à l'Alsace a représenté une limite en termes de finesse de stratification ou de ventilation, ce qui a conduit à mettre de côté certains croisements de données. Ainsi, certaines pistes de réflexion, comme l'utilisation de zones écologiques en plus de la hauteur des peuplements, pourraient être explorées, mais à condition de travailler à une échelle supérieure. Lors de prochains travaux à l'échelle d'une petite région en revanche, elles pourront être ignorées d'emblée.

Pour ce qui est des résultats proprement dits, les stratifications qui ont démontré le plus d'intérêt lors des phases de comparaison sans ventilation (c'est-à-dire une estimation de grandeurs avec le moins d'erreur possible) sont finalement les plus simples, faisant appel soit aux essences ou regroupements d'essences, soit à des classes de hauteur maximale du peuplement, mais pas les deux simultanément. Ceci est sans doute dû à la quantité de données disponibles comme mentionné ci-dessus, mais s'appliquerait probablement aussi dans toute situation où les modèles les plus robustes sont préférables. Selon les critères de ventilation testés, trois situations sont apparues :

- **les stratifications testées n'apportent rien ou peu** par rapport à celle qu'utilise déjà l'IFN dans la publication de ses résultats : c'est le cas pour les classes de propriété forestière (très bien prises en compte par la méthode actuelle), et pour certaines essences
- **la stratification par regroupements d'essences est la plus intéressante** : c'est le cas si l'on ventile les données par composition du peuplement, selon une simple opposition feuillus / conifères (sauf pour le volume de bois sur les feuillus), pour la surface de la plupart des essences et ponctuellement sur le volume et la production de certaines essences
- **la stratification par classes de hauteur est la plus intéressante** : c'est le cas là aussi lorsqu'on ventile par essence pour certaines d'entre elles, pour le volume de bois sur les feuillus, pour le volume et la production en volume avec une ventilation par difficulté d'exploitation, et pour la répartition par classe de diamètre, surtout pour les plus gros arbres.

Ainsi, il est vérifié qu'une stratification utilisant une donnée corrélée au critère de ventilation permet d'obtenir de meilleurs résultats qu'avec d'autres méthodes, et notamment la méthode de référence actuelle. Par exemple, aucune stratification testée n'emploie les types de propriété dans la définition de ses strates, aussi la stratification utilisée actuellement reste la meilleure si l'on ventile par ce critère. À l'inverse, les classes de hauteur que nous avons employées en stratification sont basées sur un modèle de hauteur fournissant une métrique que l'on sait bien corrélée au volume ; et la stratification résultante produit effectivement souvent des résultats plus intéressants pour les volumes.

Les stratifications créées et testées au cours de ce stage permettraient donc d'ores et déjà d'affiner plusieurs des résultats publiés par l'inventaire, notamment lorsque ceux-ci sont ventilés par essences simplifiées (composition du peuplement, feuillus ou conifères) ou par classe de diamètre. Ces stratifications et leur méthode d'obtention s'avèreraient peut-être pertinentes aussi dans d'autres situations ou d'autres régions, offrant à l'inventaire des résultats toujours plus précis.

Bibliographie

L'inventaire et ses méthodes, tentatives précédentes de stratification par des modèles de hauteur :

[1] Belouard, Thierry, Py, Nicolas, Jolly, Anne et Renaud, Jean Pierre, juin 2013. Stratification et autres considérations méthodologiques relatives à l'échantillonnage de placettes de lever pour les trois zones biogéographiques de l'étude, projet ANR-10-BIOE-008 : Caractérisation de la ressource forestière pour les bioénergies, tâche 310, jalon J31 (diffusion interne)

[2] Belouard, Thierry, Py, Nicolas et Renaud, Jean Pierre, février 2015. Illustration complémentaire sur le département de la Nièvre (58), projet ANR-10-BIOE-008 : Caractérisation de la ressource forestière pour les bioénergies, tâche 310, jalon J31 (diffusion interne)

[3] Hervé, Jean-Christophe, Wurpillot, Stéphanie, Vidal, Claude et Roman-Amat, Bernard, 2014. L'inventaire des ressources forestières en France : un nouveau regard sur de nouvelles forêts. *Revue Forestière Française LXVI (3)*, pp. 247-260

[4] Ketchiamen Tchatchoua, Emily Linda, 2016. Utilisation des couches cartographiques pour l'amélioration des estimations statistiques de l'inventaire forestier national français, Mémoire de fin d'études présenté en vue de l'obtention du diplôme de master international en écologie appliquée, option gestion des écosystèmes, université de Poitiers

[5] Morneau, François, Wurpillot, Stéphanie et Hervé, Jean-Christophe, 2017. Enquête « L'inventaire forestier national » : dossier de présentation au Comité du label de la statistique publique - demande de label d'intérêt général et de qualité statistique avec caractère obligatoire, séance du 16 mai 2017

[6] Pauthonnier, Alexandre et équipe produits altimétriques de l'IGN, 2014. Spécifications du produit interne RGE ALTI, édition 1.1, 18 juin 2014

[7] Site internet public de l'IFN : <http://inventaire-forestier.ign.fr/spip/>

Logiciels et packages R et leur documentation :

Allaire, JJ, Cheng, Joe, Xie, Yihui, McPherson, Jonathan, Chang, Winston, Allen, Jeff, Wickham, Hadley, Atkins, Aron et Hyndman, Rob, 2016. *rmarkdown: Dynamic Documents for R*. R package version 1.3. <https://CRAN.R-project.org/package=rmarkdown>

Bivand, Roger, Keitt, Tim et Rowlingson, Barry, 2016. *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 1.2-5. <https://CRAN.R-project.org/package=rgdal>

Dowle, Matt et Srinivasan, Arun, 2017. *data.table: Extension of `data.frame`*. R package version 1.10.4. <https://CRAN.R-project.org/package=data.table>

Dowle, Matt, 2017. *Data analysis the data.table way : The official Cheat Sheet for the DataCamp course*. © 2017 DataCamp Inc.

Dowle, Matt, 2 octobre 2014. *FAQs about the data.table package in R*.

Hijmans, Robert J., 2016. *raster: Geographic Data Analysis and Modeling*. R package version 2.5-8. <https://CRAN.R-project.org/package=raster>

Morneau, François, 2017. *inventR : Fonctions R pour l'Inventaire Forestier*. R package version 3.3.4-1.

R Core Team, 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

RStudio Team, 2016. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. <http://www.rstudio.com>.

RStudio Team, 2014. *R Markdown Reference Guide*, © 2014 RStudio, Inc.

Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilke, Claus O., 2016. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.7.0. <https://CRAN.R-project.org/package=cowplot>

Annexes

1. Calcul des composantes de la variance

La variance d'une grandeur (surface, volume...) estimée au niveau d'un sous-domaine (par exemple les feuillus) peut être décomposée en trois composantes. Soient P_{fk} la proportion du sous-domaine f dans la strate k , et $M_{fk}(y)$ la moyenne de la grandeur y dans f et k , et $V_{fk}(y)$ la variance de la grandeur y dans le sous-domaine f et la strate k . Les composantes de la variance s'écrivent alors :

V_{SD} : la variance « de sous-domaine » introduite par l'incertitude sur la surface du sous-domaine.

Pour une surface : $V(\bar{Y})$

Pour le total d'une autre grandeur : $RE = \frac{V(\bar{Y}_{ref})}{V(\bar{Y}_{str})}$

V_{ST} : la variance liée à la stratification, introduite par les erreurs d'échantillonnage sur les surfaces de strates, lorsque celles-ci sont estimées à partir du nombre de points (nulle si les surfaces sont connues).

Pour une surface : $V_{SD}(S_f) = \sum_k \left(\frac{(S_k^2 + V(S_k)) P_{fk} (1 - P_{fk})}{n'_k - 1} \right)$

Pour le total d'une autre grandeur : $V_{SD}(T_f(y)) = \sum_k \left(\frac{(S_k^2 + V(S_k)) M_{fk}(y)^2 P_{fk} (1 - P_{fk})}{n'_k - 1} \right)$

V_{HT} : la variance « d'hétérogénéité » de la grandeur considérée, due à sa variabilité intra-strate (non définie si la grandeur est la surface du sous-domaine).

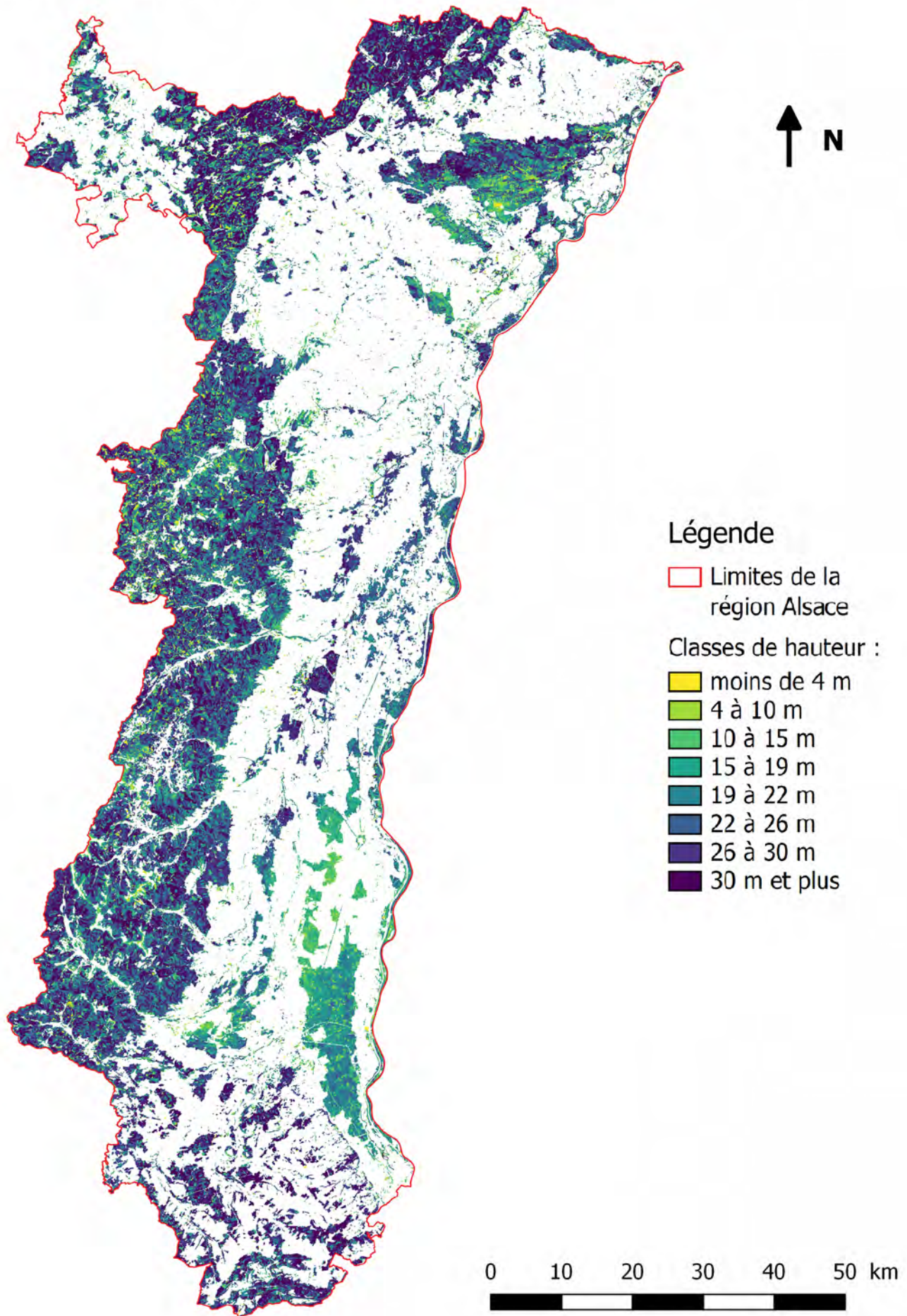
Pour le total : $V_{ST}(S_f) = \sum_k \sum_l P_{fk} P_{fl} cov(S_k, S_l)$

Son estimation prend en compte le fait que les variables considérées en pratique sont presque toujours positives ou nulles, et valent souvent 0 si l'on ventile par un critère avec de nombreuses catégories (par exemple, le volume d'une essence en particulier vaut 0 sur tous les points où elle n'est pas présente, soit en général la majorité). Si l'on respecte bien $y \geq 0$, on a alors :

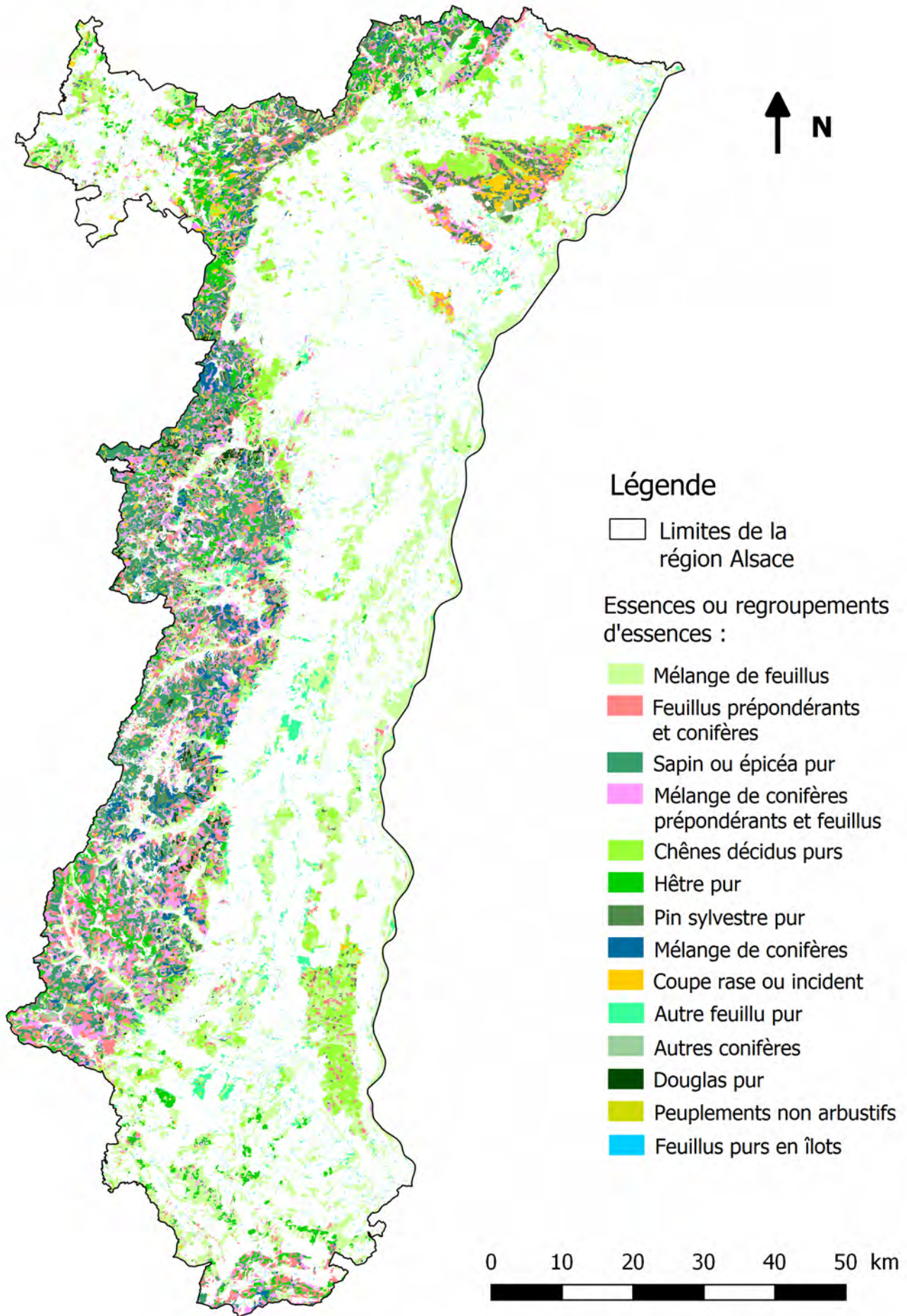
$$V_{fk}(y) = (1 - P_{fk}(y = 0)) (M_{fk}(y > 0))^2 \times (P_{fk}(y = 0) + (Cv_{fk}(y > 0))^2)$$

avec $P_{fk}(y = 0)$ la proportion de la surface dans k et f pour laquelle $y = 0$, $M_{fk}(y > 0)$ la moyenne de y dans f et k lorsque $y > 0$, et $Cv_{fk}(y > 0)$ le coefficient de variation de y dans k et f lorsque $y > 0$. Ce dernier est calculé avec l'hypothèse que les valeurs des coefficients de variation recherchés ne dépendent pas des strates [5].

Classes de hauteur définies en Alsace à partir du MNH forêt complet



Essences et regroupements d'essences définis en Alsace à partir de BD Forêt V2



Processus de création des tuiles MNH

Sophie Georges

10 avril 2017

Objectif

Le but de ce script est de créer, pour l'ensemble des forêts de l'ancienne région Alsace, un MNH complet et de la meilleure qualité possible. Pour ce faire, on dispose des données suivantes (issues de l'IGN) sur le serveur de l'IFN :

- Un MNT complet sur la région, formé de tuiles de 5 x 5 km et de 5 m de résolution, obtenu à partir de la BD Alti par numérisation manuelle ou vectorisation.
- Un MNT LiDAR disponible sur une grande partie de la région, mais loin d'être exhaustif. Les données disponibles couvrent notamment les zones inondables comme la plaine du Rhin, mais aussi des zones de forêt. Elles prennent la forme de tuiles de 1 km de côté et 1m de résolution, mais beaucoup d'entre elles sont incomplètes vu l'étroitesse de certaines des zones couvertes. Elles sont accompagnées d'une grille vectorielle qui indique leur position et sert de référence pour la suite.
- Un MNE d'une résolution de 20 cm, sous forme de tuiles de 1 km de côté, disponible sur toute la région. Elles sont originaires au format .TIF, contrairement aux autres données qui sont en .asc. Elles sont aussi plus grandes, dépassant de 2 m des tuiles MNT lidar dans toutes les directions, si bien qu'elles se chevauchent. Toutes ces tuiles comportent dans leur nom les coordonnées de leur sommet nord-ouest (de la forme ... *x_y*...).

La base de données de l'inventaire dispose aussi de données postGIS. Celles-ci ne sont pas utilisées directement ici (mais ce serait envisageable). Elles ont servi à créer deux couches vectorielles : une pour les contours de l'Alsace (à partir de la couche "communes") et une pour la présence/absence de forêts (à partir de la BD Forêt V2, qui cartographie les formations forestières par département).

Ici, on cherche au final à avoir un MNH ramené à une résolution de 1 m et uniquement sur les zones de forêt de l'Alsace. Comme les données MNT LiDAR sont de bien meilleure qualité que le MNT Alti, elles sont à utiliser en priorité.

Préparation de l'espace de travail

Packages utilisés

Outre les packages nécessaires pour manipuler les fichiers raster utilisés et produits (.TIF et .asc), une grande partie du code repose sur l'utilisation du package **data.table**, qui permet de manipuler de grandes quantités d'information rapidement (une version plus optimisée de ce script pourrait peut-être être obtenue en exploitant au mieux ce package pour éviter les boucles *for*). **rpostgis** n'est pas utilisé ici, mais pourrait l'être si les couches vectorielles utilisées provenaient d'un serveur.

inventR est un package créé et utilisé en interne à l'IFN (auteur : François Morneau). Ici, il est utilisé pour la reprojection des couches vectorielles du référentiel Lambert II étendu vers le Lambert 93.

```
library(raster)
library(gtools)
library(data.table)
library(rgdal)
library(rgeos)
```

```
#library(rpostgis)
library(inventR)
```

Chemins d'accès

Certains chemins d'accès aux données sur serveur reviennent plusieurs fois. Comme de plus ils sont longs, il vaut mieux les définir comme variables :

```
# chemins alternatifs suivant le poste de travail
path_partage <- "~/Documents/SIG/data/IGN/MNS/mns_Alsace/smb-share:server=dgs1211n001,
                share=mns,user=fmorneau/"      # poste de François (sous Linux)
path_partage <- "\\Dgs1211n001/mns/"          # mon poste (sous Windows)

# Chemins d'accès aux différentes sources de données serveur
path_lidar <- paste0(path_partage, "MNT_Alsace/MNT1m/1_DONNEES_LIVRAISON_2017-03-00106/
                        RGEALTI_MNT_1M_ASC_LAMB93_IGN69_42_20170307/")
path_mne_67 <- paste0(path_partage, "FD67/Dalles/")
path_mne_68 <- paste0(path_partage, "FD68/Dalles/")
path_5m_67 <- paste0(path_partage, "MNT_Alsace/RGEALTI_2-0_5M_ASC_LAMB93-IGN69_D067_2016
                        -02-03/RGEALTI/1_DONNEES_LIVRAISON_2016-03-00227/RGEALTI_MNT_5M_ASC_
                        LAMB93_IGN69_D067/")
path_5m_68 <- paste0(path_partage, "MNT_Alsace/RGEALTI_2-0_5M_ASC_LAMB93-IGN69_D068_2015
                        -08-18/RGEALTI/1_DONNEES_LIVRAISON_2016-03-00227/RGEALTI_MNT_5M_ASC_
                        LAMB93_IGN69_D068/")
```

Couches vectorielles

Ces couches correspondent aux contours de l'Alsace et à la présence de forêts. Elles vont servir à détecter si une tuile MNT les intersecte bien toutes les deux. La projection *lambert93* utilisée est celle définie dans inventR. On récupère les départements individuels à partir de la région.

```
foret <- readOGR("E:\\Documents SGeorges\\masque_foret.shp")
foret_93 <- spTransform(foret, CRS(lambert93))
Alsace <- readOGR("E:\\Documents SGeorges\\Alsace_ref93_tout.shp")
Alsace_93 <- spTransform(Alsace, CRS(lambert93))
hautrhin <- Alsace_93[Alsace_93@data$num_dep == "68",]
basrhin <- Alsace_93[Alsace_93@data$num_dep == "67",]
```

Data table des tuiles MNT potentielles

L'objectif de cette étape est de créer et remplir une *data table* avec une ligne par tuile MNT de 1 km² pouvant être définie pour l'Alsace en s'appuyant sur la grille, et qui contienne toutes les informations nécessaires. Comme le MNT Alti couvre largement la région de travail, les tuiles de 5 km de côté (tuiles Alti) peuvent servir de base pour définir par leurs coordonnées toutes les tuiles de 1 km² correspondantes. Par la suite, des tests permettront de vérifier lesquelles de ces tuiles de 1 km² il sera pertinent de créer.

Data tables des données disponibles

On commence par faire la liste des données disponibles, également sous forme de *data tables*, ce qui permettra de gérer ces listes facilement et rapidement. Les *data tables* sont toutes indexées (fonction *setkey*) pour

accélérer les traitements.

```
# on liste les tuiles au pas de 5m (en gardant l'information département d'origine)
tuiles_5m_67 <- list.files(path = path_5m_67, pattern = "\\\\.asc$", recursive = TRUE)
# le pattern choisi permet de sélectionner uniquement les fichiers .asc, sans ceux
# en .asc.aux.xml générés quand le fichier est ouvert dans QGIS
tuiles_5m_68 <- list.files(path = path_5m_68, pattern = "\\\\.asc$", recursive = TRUE)

tuiles_5m <- rbindlist(list(data.table(dep = 67, tuile = tuiles_5m_67),
                           data.table(dep = 68, tuile = tuiles_5m_68)))

# on élimine les tuiles doublons (à la frontière des deux départements)
tuiles_5m <- tuiles_5m[!duplicated(tuiles_5m[, tuile])]
setkey(tuiles_5m, tuile) # on indexe la data table

# On liste les tuiles MNT LiDAR
tuiles_lidar <- data.table(lidar = list.files(path = path_lidar, pattern = "\\\\.asc$",
                                             recursive = TRUE))
setkey(tuiles_lidar, lidar) # on indexe la data table
```

Liste des tuiles potentielles

À partir de la table *tuiles_5m*, on va pouvoir générer la liste principale : *tuiles_1m*. On génère le nom de chaque tuile de 1 km² potentielle à partir de celui de la tuile Alti correspondante.

Initialisation

Pour commencer, on crée une *data table* avec tous les champs qui serviront pour la suite (initialisés à 0 ou FALSE), basée sur *tuiles_5m* :

```
tuiles_1m <- data.table(id = 0, dep = tuiles_5m[, dep], nom = tuiles_5m[, tuile],
                      mnt_5m = tuiles_5m[, tuile], coord = 0, x_tuile = 0, y_tuile = 0,
                      lidar = "", absent = FALSE, NAs = FALSE, alsace = FALSE,
                      foret = FALSE, coin_NO_x = 0, coin_NO_y = 0, coin_NE_x = 0,
                      coin_NE_y = 0, coin_SE_x = 0, coin_SE_y = 0, coin_SO_x = 0,
                      coin_SO_y = 0)
```

Coordonnées x et y des tuiles potentielles

Chaque tuile Alti correspond à 25 tuiles de 1 km de côté potentielles. On répète donc d'abord cette première liste 5 fois en concaténant 5 copies de la table (donc 5 lignes par tuile Alti) :

```
tuiles_1m <- rbindlist(list(tuiles_1m, tuiles_1m, tuiles_1m, tuiles_1m, tuiles_1m))
```

Puis on définit les 5 valeurs de *x* qui correspondent à chaque tuile Alti. Pour ce faire, on utilise l'opérateur modulo (%) avec la coordonnée en *x* fournie par le nom de la tuile Alti et le numéro de ligne. Par exemple, RGEALTI_FXX_0995_6865_MNT_LAMB93_IGN69.asc correspond aux valeurs de *x* 995, 996, 997, 998 et 999, donc les cinq lignes qui lui correspondent se voient attribuer ces valeurs dans le champ *x_tuile*.

Remarque le fonctionnement des *data tables* : on utilise la syntaxe **DT[i, j := a, by = b]**, où l'opérateur **:=** sert à mettre à jour le champ *j* de la table *DT* avec l'expression *a*, pour la ligne *i* (qui peut être toutes les lignes si on ne met rien avant la virgule, ou une condition), en groupant par *b*.

```

tuiles_1m[, coord := sub("_MNT_LAMB93_IGN69.asc", "", sub("RGEALTI_FXX_", "", nom))]
tuiles_1m[, x_tuile := as.numeric(substr(coord, 1, 4))]
tuiles_1m[, id := .I] # on récupère le n° de ligne
tuiles_1m[, x_tuile := x_tuile + id %% 5]

```

Le raisonnement est le même en y . On a 5 valeurs possibles par tuile Alti **et** par valeur de x , donc on répète à nouveau la table 5 fois. Cette fois, la valeur de y obtenue avec $y_tuile - (((id - 1) \% \% 5) \% \% 5)$, où $\% \%$ est le reste entier de la division, se répète toutes les 25 lignes :

```

tuiles_1m[, y_tuile := as.numeric(substr(coord, 6, 9))]
tuiles_1m <- rbindlist(list(tuiles_1m, tuiles_1m, tuiles_1m, tuiles_1m, tuiles_1m))
tuiles_1m[, id := .I] # on rafraîchit le n° de ligne
tuiles_1m[, y_tuile := y_tuile - (((id - 1) \% \% 5) \% \% 5)]

```

Nom des tuiles potentielles

Maintenant qu'on a toutes les combinaisons possibles de x et y correspondant au MNT Alti (donc à la région en débordant par endroits), on peut les utiliser pour générer le nom des tuiles de 1 km² potentielles. À noter le cas particulier des quelques colonnes dont le nom commence par "09" : comme on a perdu cette information en passant en format numérique pour x et y (ce qui était nécessaire pour les utiliser dans des calculs), il faut restituer le "0" manquant s'il y a lieu. Ce détail revient plusieurs fois ; il est géré par la fonction **ifelse(condition, alors, sinon)** qui est souvent utilisé par la suite. Le nom résultant est de la même forme que celui des tuiles LiDAR : *RGEALTI_FXX_x_y_MNT_LAMB93_IGN69.asc*.

```

tuiles_1m[, coord := paste0(ifelse(x_tuile < 1000, paste0("0", as.character(x_tuile)),
                                as.character(x_tuile)), "_", as.character(y_tuile))]
tuiles_1m[, nom := paste0("RGEALTI_FXX_", coord, "_MNT_LAMB93_IGN69.asc")]
setkey(tuiles_1m, nom) # on rafraîchit l'indexation

```

Coordonnées des coins

Pour les tests d'intersection qui vont suivre, on définit les coordonnées x et y de chaque angle des tuiles potentielles :

```

tuiles_1m[, coin_NO_x := x_tuile * 1000]
tuiles_1m[, coin_NO_y := y_tuile * 1000]
tuiles_1m[, coin_NE_x := (x_tuile + 1) * 1000]
tuiles_1m[, coin_NE_y := y_tuile * 1000]
tuiles_1m[, coin_SE_x := (x_tuile + 1) * 1000]
tuiles_1m[, coin_SE_y := (y_tuile - 1) * 1000]
tuiles_1m[, coin_SO_x := x_tuile * 1000]
tuiles_1m[, coin_SO_y := (y_tuile - 1) * 1000]

```

Tests

Les tuiles MNT de 1 m de résolution et 1 km de côté listées dans `tuiles_1m` ne sont pas toutes pertinentes à créer. Une série de tests faisant appel aux autres données sont donc pratiqués pour déterminer si c'est le cas. La plupart requièrent un certain temps de calcul, mais une fois la *data table* correctement remplie elle peut être conservée en sauvant l'environnement de travail.

Donnée LiDAR complète sur la tuile

On n'a notamment pas besoin de générer la tuile à partir du MNT Alti s'il existe déjà à cet emplacement une tuile LiDAR et que celle-ci est complète. On recherche donc pour chaque tuile potentielle une tuile du même nom dans la liste des tuiles LiDAR, on récupère son nom (chemin compris) si elle existe, et sinon on met *TRUE* dans le test "absent" :

```
tuiles_1m[, lidar := ifelse(nom %in% tuiles_lidar[,lidar], paste0(path_lidar, "\\ ", nom),
                           "")] # la case reste vide si on ne trouve pas de fichier
tuiles_1m[, absent := !nom %in% tuiles_lidar[, lidar]]
```

Ensuite, il faut vérifier que cette tuile est complète : s'il manque ne serait-ce qu'un pixel (la tuile contient *NA* au moins une fois), il faudra créer et utiliser la tuile de 1 km² issue du MNT Alti. On définit une fonction de test, mais on l'appelle uniquement sur les lignes où la tuile LiDAR existe :

```
test_nodata <- fonction(fichier, nodata = -99999) {
  zz <- file(fichier) # on se connecte au fichier, mais sans le charger (plus rapide)
  on.exit(close(zz)) # on referme bien la connexion
  tp <- scan(zz, skip = 6, quiet = TRUE)
  res <- any(tp == nodata)
  return(res)
}
tuiles_1m[absent == FALSE, NAs := test_nodata(paste0(path_lidar, nom)), by = id]
```

Tuile en Alsace

Étant donné la taille des tuiles Alti d'origine, certaines tuiles de 1 km² qui en dérivent sont entièrement en dehors de la région. Pour ne pas les traiter, on crée le polygone correspondant à chaque tuile avec les coins définis plus haut, et on vérifie s'il intersecte la couche *Alsace_93* :

```
dans_region <- fonction(NO_x, NO_y, NE_x, NE_y, SE_x, SE_y, SO_x, SO_y) {
  coins <- rbind(c(NO_x, NO_y), c(NE_x, NE_y), c(SE_x, SE_y), c(SO_x, SO_y))
  # on crée le polygone à partir du vecteur de coordonnées de ses points (coins)
  case <- SpatialPolygons(list(Polygons(list(Polygon(coins)), 1)),
                          proj4string = Alsace_93@proj4string)
  en_alsace = gIntersects(case, Alsace_93)
  return(en_alsace)
}
tuiles_1m[, alsace := dans_region(coin_NO_x, coin_NO_y, coin_NE_x, coin_NE_y,
                                coin_SE_x, coin_SE_y, coin_SO_x, coin_SO_y), by = id]
```

Tuile en forêt

De même, on utilise la couche présence/absence de forêts *foret_93* pour déterminer avec une fonction quelles tuiles ne recoupent aucune zone boisée.

```
dans_foret <- fonction(NO_x, NO_y, NE_x, NE_y, SE_x, SE_y, SO_x, SO_y) {
  coins <- rbind(c(NO_x, NO_y), c(NE_x, NE_y), c(SE_x, SE_y), c(SO_x, SO_y))
  case <- SpatialPolygons(list(Polygons(list(Polygon(coins)), 1)),
                          proj4string = Alsace_93@proj4string)
  en_foret = gIntersects(case, foret_93)
  return(en_foret)
}
```


Par contre, on n'effectue le test que pour les tuiles dont on sait qu'elles sont non disponibles LiDAR (ou contiennent des *NA*) et bien en Alsace pour accélérer un peu le traitement. Par la suite, le test "présence de forêts" suffit donc : *foret = FALSE* (la valeur par défaut) correspond à tous les cas où traiter la tuile est non pertinent.

```
tuiles_1m[, foret := ifelse(absent == TRUE,
                           ifelse(alsace == TRUE,
                                   dans_foret(coin_NO_x, coin_NO_y, coin_NE_x, coin_NE_y,
                                             coin_SE_x, coin_SE_y, coin_SO_x, coin_SO_y),
                                   FALSE),
                           ifelse(NAs == TRUE,
                                   ifelse(alsace == TRUE,
                                           dans_foret(coin_NO_x, coin_NO_y, coin_NE_x,
                                                     coin_NE_y, coin_SE_x, coin_SE_y,
                                                     coin_SO_x, coin_SO_y),
                                           FALSE),
                                   FALSE)), by = id]
```

Création des tuiles MNT nécessaires

Tuiles MNT Alti à la bonne résolution

Fonction *disaggregate*

Les tuiles de 5 km de côté issues de la BD Alti ont une résolution de 5 m. Avant de les découper pour obtenir les tuiles voulues, il faut donc les faire passer à la résolution souhaitée au final. Vu le poids de ces tuiles, le calcul est long et on cherche à l'effectuer uniquement si nécessaire : si au moins l'une des tuiles de 1 km² théoriques recouvrant la tuile Alti est nécessaire, mais que sa version lidar est incomplète ou manquante. Pour charger les données Alti nécessaires, on doit gérer à chaque fois des chemins différents selon le département, c'est pourquoi cette information a été conservée dans *tuiles_1m*.

On ramène les tuiles Alti à 1 m de résolution avec la fonction *disaggregate*, qui permet de diviser les pixels d'un raster par un facteur entier (ici 5) pour passer à une taille de pixels inférieure. Comme on ne souhaite pas lisser les valeurs, on utilise la méthode par défaut de *disaggregate*, où les nouveaux pixels ont la valeur de leur pixel d'origine. Cette fonction oblige à sauver le résultat sous forme de fichier et ne permet pas de spécifier un chemin d'écriture, donc tout se fait dans le répertoire de travail.

Prise en compte du décalage entre sources de données

Une difficulté supplémentaire est posée par le fait que les tuiles LiDAR et Alti ne coïncident pas : le découpage du MNT Alti est décalé de 2 m vers le nord et 2 m vers l'ouest par rapport aux tuiles LiDAR, sur lesquelles on souhaite s'appuyer. Les tuiles de 1 km² situées en bordure est d'une tuile Alti (n° de colonne *x* finissant par 4 ou 9) dépassent ainsi de 2 m sur la tuile Alti juste à l'est, donc on aura aussi besoin de celle-ci à la bonne résolution pour y découper cette bande ; le problème est le même en bordure sud (n° de ligne *y* finissant par 1 ou 6). Si les deux conditions sont réunies, on aura de plus besoin de la tuile Alti située au sud-est pour y découper l'angle sud-est de la tuile de 1 km².

Pour gérer l'utilisation de *disaggregate*, on utilise une boucle *for* qui parcourt la *data table* *tuiles_1m* et qui, chaque fois qu'une tuile de 1 km² non LiDAR est nécessaire, effectue l'opération d'agrégation sur la tuile Alti correspondante, sauf si celle-ci a déjà été traitée (ce qui est vérifié avec *file.exists*). Suivant les coordonnées de cette tuile de 1 km², les tuiles Alti situées juste à l'est et/ou juste au sud et/ou au sud-est sont traitées aussi, là encore si ce n'est pas déjà fait.

Boucle complète

```
for (i in 1:nrow(tuiles_1m)){
  if (tuiles_1m[i, foret] == TRUE) { # seul test nécessaire pour savoir si créer la tuile
    if (tuiles_1m[i, dep] == 67) {
      # on vérifie que la tuile à 5m n'a pas déjà été traitée (le fichier n'existe pas encore)
      if (file.exists(paste0("disaggregate_", sub(path_5m_67, "", tuiles_1m[i, mnt_5m])))
          == FALSE) {
        # on va chercher la donnée sur le serveur (on est dans la région, elle existe forcément)
        mnt_init <- raster(paste0(path_5m_67, tuiles_1m[i, mnt_5m]))
      }
    } else { # mêmes opérations avec une tuile dans le département 68
      if (file.exists(paste0("disaggregate_", sub(path_5m_68, "", tuiles_1m[i, mnt_5m])))
          == FALSE) {
        mnt_init <- raster(paste0(path_5m_68, tuiles_1m[i, mnt_5m]))
      }
    }
    disaggregate(mnt_init, fact = 5, method = "",
                 filename = paste0("disaggregate_", tuiles_1m[i, mnt_5m]),
                 overwrite = TRUE)

    # même opération si la tuile Alti située à l'est est nécessaire et non encore traitée
    if (tuiles_1m[i, x_tuile] %% 5 == 4){
      if (tuiles_1m[i, dep] == 67) {
        nom_mnt_E <- ifelse(tuiles_1m[i, y_tuile] %% 5 == 0,
        # les coordonnées correctes de la tuile Alti à l'est doivent être calculée différemment
        # si le n° de ligne est multiple de 5 (y_tuile %% 5 valant alors 0)
          paste0(path_5m_67, "RGEALTI_FXX_",
                 ifelse(tuiles_1m[i, x_tuile] < 1000, "0", ""),
                 tuiles_1m[i, x_tuile] + 1, "_",
                 tuiles_1m[i, y_tuile] - (tuiles_1m[i, y_tuile] %% 5),
                 "_MNT_LAMB93_IGN69.asc"),
          paste0(path_5m_67, "RGEALTI_FXX_",
                 ifelse(tuiles_1m[i, x_tuile] < 1000, "0", ""),
                 tuiles_1m[i, x_tuile] + 1, "_",
                 tuiles_1m[i, y_tuile] - (tuiles_1m[i, y_tuile] %% 5)+5,
                 "_MNT_LAMB93_IGN69.asc"))
        # on vérifie que la tuile désagrégée n'existe pas encore
        if (file.exists(paste0("disaggregate_", sub(path_5m_67, "", nom_mnt_E))) == FALSE){
          # certaines des tuiles Alti nécessaires à l'est ne sont pas disponibles car trop en dehors
          # de la région, donc on vérifie qu'elles existent (tenter de traiter une tuile absente
          # interromprait toute la boucle for)
          if (file.exists(nom_mnt_E) == TRUE){
            mnt_init_E <- raster(nom_mnt_E)
            disaggregate(mnt_init_E, fact = 5, method = "",
                         filename = sub("\\\\", "", paste0("disaggregate_", sub(path_5m_67,
                         "", nom_mnt_E))), overwrite = TRUE)
          }
        }
      }
    } else {
```

```

nom_mnt_E <- ifelse(tuiles_1m[i, y_tuile] %% 5 == 0,
  paste0(path_5m_68, "RGEALTI_FXX_",
    ifelse(tuiles_1m[i, x_tuile] < 1000, "0", ""),
    tuiles_1m[i, x_tuile] + 1, "_",
    tuiles_1m[i, y_tuile] - (tuiles_1m[i, y_tuile] %% 5),
    "_MNT_LAMB93_IGN69.asc"),
  paste0(path_5m_68, "RGEALTI_FXX_",
    tuiles_1m[i, x_tuile] + 1, "_",
    tuiles_1m[i, y_tuile] - (tuiles_1m[i, y_tuile] %% 5)+5,
    "_MNT_LAMB93_IGN69.asc"))

if (file.exists(paste0("disaggregate_", sub(path_5m_68,"", nom_mnt_E))) == FALSE){
  if (file.exists(nom_mnt_E) == TRUE){
    mnt_init_E <- raster(nom_mnt_E)
    disaggregate(mnt_init_E, fact = 5, method = "",
      filename = sub("\\\\", "", paste0("disaggregate_", sub(path_5m_68,
        "", nom_mnt_E))), overwrite = TRUE)
  }
}
}

# même opération pour la tuile à 5m au sud-est, si elle est disponible
# (nb : ici, pas besoin de tester si le n° de ligne est multiple de 5)
if (tuiles_1m[i, y_tuile] %% 5 == 1){
  nom_mnt_SE <- paste0("RGEALTI_FXX_", as.character(tuiles_1m[i, x_tuile] + 1),
    "_", as.character(tuiles_1m[i, y_tuile] - 1),
    "_MNT_LAMB93_IGN69.asc")
  if (tuiles_1m[i, dep] == 67) {
    if (file.exists(paste0(path_5m_67, nom_mnt_SE)) == TRUE){
      if (file.exists(paste0("disaggregate_", nom_mnt_SE)) == FALSE){
        mnt_init_SE <- raster(paste0(path_5m_67, nom_mnt_SE))
        disaggregate(mnt_init_SE, fact = 5, method = "",
          filename = paste0("disaggregate_", nom_mnt_SE), overwrite = TRUE)
      }
    }
  }
} else {
  if (file.exists(paste0(path_5m_68, nom_mnt_SE)) == TRUE){
    mnt_init_SE <- raster(paste0(path_5m_68, nom_mnt_SE))
    disaggregate(mnt_init_SE, fact = 5, method = "",
      filename = paste0("disaggregate_", nom_mnt_SE), overwrite = TRUE)
  }
}
}
}

# même opération pour la tuile au sud si elle existe (pas de n° de ligne multiple de 5)
if (tuiles_1m[i, y_tuile] %% 5 == 1){
  if (test_mnt != tuiles_1m[i - 5, mnt_5m]) {
    nom_mnt_S <- paste0("RGEALTI_FXX_",
      as.character(tuiles_1m[i, x_tuile] - tuiles_1m[i, x_tuile]%5),

```

```

        "_", as.character(tuiles_1m[i, y_tuile] - 1),
        "_MNT_LAMB93_IGN69.asc")
if (tuiles_1m[i - 5, dep] == 67) {
  if (file.exists(paste0(path_5m_67, nom_mnt_S)) == TRUE){
    if (file.exists(paste0("disaggregate_", nom_mnt_S)) == FALSE){
      mnt_init_S <- raster(paste0(path_5m_67, nom_mnt_S))
      disaggregate(mnt_init_S, fact = 5, method = "",
                  filename = paste0("disaggregate_", nom_mnt_S), overwrite = TRUE)
    }
  }
} else {
  if (file.exists(paste0(path_5m_68, nom_mnt_S)) == TRUE){
    if (file.exists(paste0("disaggregate_", nom_mnt_S)) == FALSE){
      mnt_init_S <- raster(paste0(path_5m_68, nom_mnt_S))
      disaggregate(mnt_init_S, fact = 5, method = "",
                  filename = paste0("disaggregate_", nom_mnt_S), overwrite = TRUE)
    }
  }
}
}
}
}
}
}
}
}

```

Création des tuiles MNT de 1 km² proprement dites

Principe

Maintenant, on dispose de tuiles Alti à la même résolution que les tuiles LiDAR (la qualité réelle de l'image n'ayant évidemment pas changé). Il reste à découper ces tuiles de 5 km de côté en tuiles de 1 km² alignées sur les tuiles LiDAR et venant les remplacer là où elles sont absentes ou incomplètes, pour au final obtenir une couverture uniforme de la région par des tuiles MNT de même taille et même résolution. On parcourt donc à nouveau la liste des tuiles potentielles (la *data table tuiles_1m*) mais cette fois, lorsque le test indique que la tuile est à créer, on définit le polygone qui correspond à la future tuile et on s'en sert pour découper celle-ci à partir des tuiles Alti désagrégées obtenues précédemment.

Il faut remarquer qu'il était indispensable de générer les tuiles Alti désagrégées **avant** de pratiquer le découpage, car sinon celui-ci se baserait sur des pixels de 5 m dont, à cause du décalage de 2 m mentionné plus haut, le sommet ne correspond pas à celui des futures tuiles de 1 km². Étant donné la façon dont la fonction *crop* gère les pixels, le résultat final serait décalé par rapport à ce qui est attendu.

Décalages à gérer

Le décalage entre MNT Alti et LiDAR conduit, comme on l'a vu, à ce que certaines tuiles à créer débordent de la tuile Alti correspondante si elles sont en bordure est et/ou sud. Comme on l'a pris en compte pour générer les tuiles désagrégées concernées là où c'est nécessaire, on va pouvoir les utiliser en les chargeant s'il y a lieu et en les découpant elles aussi avec la case correspondant à la future tuile. Les tuiles MNT obtenues au final résulteront donc de l'assemblage (avec la fonction *merge*) de jusqu'à quatre parties issues d'autant de tuiles Alti différentes : la tuile principale, une bande de 2 m de largeur à l'est, une au sud et un coin de 4 m² au sud-est. Cependant, comme les tuiles Alti correspondant à ces bordures et coins peuvent ne pas être disponibles, certaines tuiles MNT se retrouveront incomplètes. Vu les faibles surfaces concernées et

comme ces bandes ou coins seront en dehors de la région (trop au sud ou à l'est pour que le MNT Alti soit disponible), les tuiles seront quand même utilisées telles quelles.

Par ailleurs, comme la grille de définition des tuiles LiDAR (en coordonnées entières) est basée sur le centre des pixels et non leur coin, une correction de - 0,5 m (un demi-pixel) doit être apportée en x aux coordonnées des coins pour que le bord de la case de découpe corresponde bien au bord du pixel, sans quoi la première colonne ne sera pas prise en compte (le paramétrage de *crop* fait que la même opération n'est pas nécessaire en y).

Boucle complète

```
for (i in 1:nrow(tuiles_1m)){
  if (tuiles_1m[i, foret] == TRUE){ # seul test nécessaire pour savoir si créer la tuile
# le décalage de 0,5 en x permet d'aligner les nouvelles tuiles sur les tuiles LiDAR
    coins <- rbind(c(tuiles_1m[i, coin_NO_x - 0.5], tuiles_1m[i, coin_NO_y]),
                  c(tuiles_1m[i, coin_NE_x - 0.5], tuiles_1m[i, coin_NE_y]),
                  c(tuiles_1m[i, coin_SE_x - 0.5], tuiles_1m[i, coin_SE_y]),
                  c(tuiles_1m[i, coin_SO_x - 0.5], tuiles_1m[i, coin_SO_y]))
    case <- SpatialPolygons(list(Polygons(list(Polygon(coins)), 1)),
                             proj4string = Alsace_93@proj4string)
# on récupère la tuile désagrégée principale, correspondant aux coordonnées
    mnt_1m <- raster(paste0("disaggregate_", tuiles_1m[i, mnt_5m]))
# on crée la nouvelle tuile MNT (mais on ne la sauve pas encore)
    tuile_mnt <- crop(mnt_1m, case)

# on vérifie si la tuile en cours de création est une tuile de bordure est et/ou sud
# et/ou un coin sud-est, et si oui on charge les tuiles désagrégées nécessaires,
# on les découpe avec la même case (si elles ont pu être créées) et on rajoute à tuile_mnt
# les éventuels compléments avec merge

    if (tuiles_1m[i, x_tuile] %% 5 == 4){
# tuile en limite est (noter une fois de plus le calcul différent si y est multiple de 5)
      nom_mnt_1m_E <- paste0("disaggregate_RGEALTI_FXX_",
                            ifelse(tuiles_1m[i, x_tuile] < 999, paste0("0",
                                                                           tuiles_1m[i, x_tuile] + 1),
                                   tuiles_1m[i, x_tuile] + 1), "_",
                            tuiles_1m[i, ifelse(tuiles_1m[i, y_tuile] %% 5 == 0,
                                                  (y_tuile - y_tuile %% 5),
                                                  (y_tuile - y_tuile %% 5 + 5))],
                            "_MNT_LAMB93_IGN69.asc")

      if (file.exists(nom_mnt_1m_E) == TRUE){
        mnt_1m_E <- raster(nom_mnt_1m_E) # on charge la tuile désagrégée complémentaire
        tuile_mnt_E <- crop(mnt_1m_E, case) # on la découpe pour obtenir la bande est
        tuile_mnt <- merge(tuile_mnt, tuile_mnt_E) # on la rajoute au morceau principal
      }
    }
    if (tuiles_1m[i, y_tuile] %% 5 == 1){ # tuile au coin sud-est
      nom_mnt_1m_SE <- paste0("disaggregate_RGEALTI_FXX_",
                              ifelse(tuiles_1m[i, x_tuile] < 999, paste0("0",
                                                                              tuiles_1m[i, x_tuile] + 1),
                                      tuiles_1m[i, x_tuile] + 1), tuiles_1m[i, x_tuile] + 1),
                              "_", tuiles_1m[i, y_tuile] - 1, "_MNT_LAMB93_IGN69.asc")
      if (file.exists(nom_mnt_1m_SE) == TRUE){
        mnt_1m_SE <- raster(nom_mnt_1m_SE)
      }
    }
  }
}
```

```

    tuile_mnt_SE <- crop(mnt_1m_SE, case)
    tuile_mnt <- merge(tuile_mnt, tuile_mnt_SE)
  }
}
}

if (tuiles_1m[i, y_tuile] %% 5 == 1){ # tuile en limite sud
  nom_mnt_1m_S <- paste0("disaggregate_RGEALTI_FXX_",
                        ifelse(tuiles_1m[i, x_tuile] < 1000, paste0("0",
                            tuiles_1m[i, (x_tuile - x_tuile %% 5)]),
                            tuiles_1m[i, (x_tuile - x_tuile %% 5)]), "_",
                            tuiles_1m[i, y_tuile] - 1, "_MNT_LAMB93_IGN69.asc")
  if (file.exists(nom_mnt_1m_S) == TRUE){
    mnt_1m_S <- raster(nom_mnt_1m_S)
    tuile_mnt_S <- crop(mnt_1m_S, case)
    tuile_mnt <- merge(tuile_mnt, tuile_mnt_S)
  }
}

# maintenant que la tuile MNT est aussi complète que possible, on la sauve au format ASC
writeRaster(tuile_mnt, filename = paste0("MNT_1m_", tuiles_1m[i, nom]),
            overwrite = TRUE)
}
}

```

Sélection des tuiles MNE et du MNT à utiliser

Maintenant que l'on dispose d'une tuile MNT complète et de 1 m de résolution pour chaque case de 1 km² sur toute la région Alsace, on va pouvoir les associer aux tuiles MNE pour calculer des tuiles MNH. Avant, il faut sélectionner les tuiles MNE pertinentes (s'assurer qu'elles sont bien en Alsace et comportent de la forêt) et associer à chacune la donnée MNT appropriée.

Liste des tuiles LiDAR

On commence par lister les tuiles MNT LiDAR sous forme d'une *data table* qui indique si elles sont pertinentes (c'est-à-dire complètes ; elles sont forcément en Alsace). À noter qu'on avait déjà une liste des tuiles LiDAR avec seulement leur nom, mais les *data tables* étant générées très rapidement, en refaire une pour éviter les confusions n'est pas un problème.

```

lidar_selec <- data.table(lidar = list.files(path = path_lidar, pattern = "\\\\.asc$",
    recursive = TRUE), coord = 0, x_tuile = 0, y_tuile = 0,
    NAs = FALSE, coin_NO_x = 0, coin_NO_y = 0, coin_NE_x = 0,
    coin_NE_y = 0, coin_SE_x = 0, coin_SE_y = 0,
    coin_SO_x = 0, coin_SO_y = 0)
setkey(lidar_selec, lidar) # on indexe la data table pour accélérer les traitements

# on récupère les coordonnées figurant dans le nom de la tuile
lidar_selec[, coord := sub("_MNT_LAMB93_IGN69.asc", "", sub("RGEALTI_FXX_", "", lidar))]
lidar_selec[, x_tuile := as.numeric(substr(coord, 1, 4))]
lidar_selec[, y_tuile := as.numeric(substr(coord, 6, 9))]

```

On applique ensuite le mêmes test que précédemment. Pour gagner du temps, il serait peut-être possible de récupérer dans `tuiles_1m` le résultat du test précédent (les tests sont longs bien qu'on ne charge pas de rasters). On vérifiera sur les tuiles MNE la présence de forêts, donc ce n'est pas nécessaire ici.

```
# on vérifie si elle contient des NA (avec la fonction définie plus haut)
lidar_selec[, NAs := test_nodata(paste0(path_lidar, lidar)), by = lidar]
```

Liste des tuiles MNE

De la même manière, on liste les tuiles MNE disponibles sur le serveur et on leur applique des tests pour vérifier leur pertinence : en Alsace et en forêt ; cette fois, il n'est pas nécessaire de rechercher des *NA* car toutes les tuiles sont complètes, sauf quelques-unes en bordure (qui sont en fait hors région).

```
mne_67 <- list.files(path = path_mne_67, pattern = "\\\\.TIF$", recursive = TRUE)
mne_68 <- list.files(path = path_mne_68, pattern = "\\\\.TIF$", recursive = TRUE)
mne_tout <- rbindlist(list(data.table(dep = 67, tuile = mne_67),
                           data.table(dep = 68, tuile = mne_68)))
mne_tout <- mne_tout[!duplicated(mne_selec[, tuile])] # on élimine les doublons
mne_selec <- data.table(dep = mne_tout[, dep], tuile = mne_tout[, tuile], coord = 0,
                      x_tuile = 0, y_tuile = 0, alsace = FALSE, foret = FALSE,
                      coin_NO_x = 0, coin_NO_y = 0, coin_NE_x = 0, coin_NE_y = 0,
                      coin_SE_x = 0, coin_SE_y = 0, coin_SO_x = 0, coin_SO_y = 0)
setkey(mne_selec, tuile)
```

Le nom des tuiles MNE contient bien des coordonnées comme pour les autres données, mais elles présentent un décalage de 1 vers le sud par rapport à leur position réelle. On en tient compte quand on calcule les coordonnées de la future case de découpage :

```
mne_selec[, coord := ifelse(dep == 67, sub(".TIF", "", sub("MNE_FD67_", "", tuile)),
                           sub(".TIF", "", sub("MNE_FD68_", "", tuile)))]
mne_selec[, x_tuile := ifelse(substr(coord, 1, 1) == "9", as.numeric(substr(coord, 1,3)),
                              as.numeric(substr(coord, 1, 4)))]
mne_selec[, y_tuile := ifelse(substr(coord, 1, 1) == "9", (as.numeric(substr(coord, 5,8))
                                                           + 1, (as.numeric(substr(coord, 6, 9))) + 1)]

mne_selec[, coin_NO_x := x_tuile * 1000]
mne_selec[, coin_NO_y := y_tuile * 1000]
mne_selec[, coin_NE_x := (x_tuile + 1) * 1000]
mne_selec[, coin_NE_y := y_tuile * 1000]
mne_selec[, coin_SE_x := (x_tuile + 1) * 1000]
mne_selec[, coin_SE_y := (y_tuile - 1) * 1000]
mne_selec[, coin_SO_x := x_tuile * 1000]
mne_selec[, coin_SO_y := (y_tuile - 1) * 1000]
```

Cette fois, il faut aussi vérifier que la tuile est au moins en partie en Alsace. Pour la forêt, on n'effectue le test que pour les tuiles dont on sait qu'elles sont bien en Alsace pour accélérer un peu le traitement (avec les mêmes fonctions que pour le MNT) :

```
mne_selec[, alsace := dans_region(coin_NO_x, coin_NO_y, coin_NE_x, coin_NE_y,
                                 coin_SE_x, coin_SE_y, coin_SO_x, coin_SO_y), by = tuile]

mne_selec[, foret := ifelse(alsace == TRUE,
                            dans_foret(coin_NO_x, coin_NO_y, coin_NE_x, coin_NE_y,
                                       coin_SE_x, coin_SE_y, coin_SO_x, coin_SO_y),
                            FALSE), by = tuile]
```

À ce stade, la liste comporte des doublons à la limite entre les deux départements. En effet, les MNE ont été créés département par département, en dépassant largement les frontières de ceux-ci, mais les tuiles en bordure sont toutes incomplètes du point de vue des données : la plupart ont une emprise normale (et ne sont donc pas détectables avec les *NA*), mais une partie de la tuile est vide. Si l'on élimine simplement les doublons par ordre alphabétique, on garde les tuiles incomplètes à l'extrême sud du chantier Bas-Rhin. Aussi, pour l'éviter, on introduit un test qui vérifie si la tuile est entièrement incluse dans un département différent de celui auquel elle est censée appartenir. Comme c'est notamment le cas des tuiles de bordure, on élimine les tuiles pour lesquelles ce test est vérifié (ce n'est pas un problème, car elles sont forcément des doublons), avant d'éliminer les doublons restants.

```
hors_bordure <- fonction(NO_x, NO_y, NE_x, NE_y, SE_x, SE_y, SO_x, SO_y, dept) {
  coins <- rbind(c(NO_x, NO_y), c(NE_x, NE_y), c(SE_x, SE_y), c(SO_x, SO_y))
  case <- SpatialPolygons(list(Polygons(list(Polygon(coins)), 1)),
    proj4string = Alsace_93@proj4string)
  en_bordure = FALSE
  if (dept == 67){ # la tuile fait partie du jeu de données du Bas-Rhin
    if (gIntersects(case, hautrhin) == TRUE & gIntersects(case, basrhin) == FALSE) {
      # mais elle intersecte le Haut-Rhin et pas le Bas-Rhin
      en_bordure = TRUE
    }
  } else {
    if (gIntersects(case, basrhin) == TRUE & gIntersects(case, hautrhin) == FALSE) {
      en_bordure = TRUE
    }
  }
  complet_ok = ifelse(en_bordure == TRUE, FALSE, TRUE)
  return(complet_ok)
}
mne_selec[, complet := hors_bordure(coin_NO_x, coin_NO_y, coin_NE_x, coin_NE_y,
  coin_SE_x, coin_SE_y, coin_SO_x, coin_SO_y, dep), by = tuile]
```

Correspondance MNE-MNT

Maintenant qu'on a sélectionné les tuiles MNE sur lesquelles intervenir, il faut savoir quelle source de données MNT utiliser : le MNT LiDAR, ou les tuiles MNT Alti créées pour le compléter. On crée donc une *data table* de correspondance entre MNE et MNT, en se basant d'abord sur la *data table* des tuiles MNE, puis sur les seules tuiles MNE sélectionnées :

```
corresp <- data.table(test_mne = mne_selec[, foret], dep = mne_selec[, dep],
  mne_20cm = mne_selec[, tuile], coord = "",
  x_tuile = mne_selec[, x_tuile], y_tuile = mne_selec[, y_tuile],
  test_NA = FALSE, mnt = "", mne_1m = "", coin_NO_x = 0,
  coin_NO_y = 0, coin_NE_x = 0, coin_NE_y = 0, coin_SE_x = 0,
  coin_SE_y = 0, coin_SO_x = 0, coin_SO_y = 0)

# on utilise les coordonnées corrigées définies auparavant pour pouvoir plus tard
# rechercher la bonne tuile MNT en comparant les noms
corresp[, coord := paste0(ifelse(x_tuile < 1000, paste0("0", as.character(x_tuile)),
  as.character(x_tuile)), "_", as.character(y_tuile))]

corresp <- corresp[!test_mne == FALSE] # on élimine les tuiles non sélectionnées
corresp <- corresp[!complet == FALSE] # on élimine les tuiles de bordure
corresp <- corresp[!duplicated(corresp[, mne_20cm])] # on élimine les doublons
```



```
# on récupère le chemin complet pour les tuiles MNE
corresp[, mne_20cm := ifelse(dep == 67, paste0(path_mne_67, mne_20cm),
                             paste0(path_mne_68, mne_20cm))]
```

Plutôt que d'utiliser des tests, on fusionne *corresp* avec la *data table lidar_selec* qui liste les tuiles LiDAR, en utilisant les coordonnées des tuiles comme référence et en précisant *all.x = TRUE* pour bien garder toutes les lignes de *corresp*, pas seulement celles qui trouvent une correspondance dans *lidar_selec*. Ensuite on récupère, pour chaque tuile MNE, la tuile LiDAR si elle existe et le résultat du test qui vérifie si elle comporte des *NA* ; si la tuile LiDAR n'existe pas ou que le test est vrai, on récupère à la place la tuile MNT issue du MNT Alti (qui est dans l'espace de travail).

```
corresp_tout <- merge(corresp, lidar_selec, by="coord", all.x = TRUE)
corresp_tout[, mnt := ifelse(is.na(corresp_tout[, lidar]) == TRUE),
# pas de tuile LiDAR à l'emplacement de cette tuile MNE
                             paste0("MNT_1m_RGEALTI_FXX_", coord, "_MNT_LAMB93_IGN69.asc"),
                             ifelse(corresp_tout[, NAs] == FALSE,
                                     paste0(path_lidar, corresp_tout[, lidar]),
                                     paste0("MNT_1m_RGEALTI_FXX_", coord,
                                             "_MNT_LAMB93_IGN69.asc")))]
# on élimine les doublons à la frontière entre les deux départements
corresp_tout <- corresp_tout[!duplicated(corresp_tout[, coord])]
```

Création du MNH

Découpage des tuiles MNE

On a vu que les tuiles MNE sont initialement un peu plus grandes que les tuiles MNT, donc on calcule les coordonnées de la case correspondant à la tuile MNT (LiDAR ou non) pour découper la tuile MNE avec. Ainsi, pour chaque case, on a une tuile MNT et une tuile MNE parfaitement calées et de la même étendue, sauf s'il manque des données MNT à l'est et/ou au sud (ce qui n'empêche pas le calcul du MNH et ne concerne que quelques tuiles à l'est et au sud de la région). De plus, ces tuiles appariées couvrent entièrement la région à l'exception des zones non boisées, sans trou ni chevauchement.

Il faut une fois de plus appliquer des corrections de 0,5 m pour être calé par rapport aux coins des pixels et non aux nœuds de la grille, cette fois dans les deux dimensions. C'est indispensable, car pour le moment les tuiles MNE sont encore à une résolution de 20 cm (là encore, l'ordre des opérations est imposé pour éviter des décalages dans le résultat final).

```
corresp_tout[, coin_NO_x.x := x_tuile.x * 1000 - 0.5]
corresp_tout[, coin_NO_y.x := y_tuile.x * 1000 + 0.5]
corresp_tout[, coin_NE_x.x := (x_tuile.x + 1) * 1000 - 0.5]
corresp_tout[, coin_NE_y.x := y_tuile.x * 1000 - 0.5]
corresp_tout[, coin_SE_x.x := (x_tuile.x + 1) * 1000 - 0.5]
corresp_tout[, coin_SE_y.x := (y_tuile.x - 1) * 1000 + 0.5]
corresp_tout[, coin_SO_x.x := x_tuile.x * 1000 - 0.5]
corresp_tout[, coin_SO_y.x := (y_tuile.x - 1) * 1000 + 0.5]
```

Soustraction et sauvegarde du MNH

Il est maintenant possible de calculer le MNH tuile par tuile. Pour cela, on utilise une fois de plus une boucle *for*, qui cette fois parcourt *corresp_tout*. À ce stade, de 10 300 tuiles de 1 km² potentielles listées dans

tuiles_1m, on est passé à 7 975 tuiles MNH réellement à créer dans *corresp_tout* en éliminant les tuiles hors région ou ne comportant pas du tout de forêts.

On commence par découper la tuile MNE en fonction des coordonnées définies ci-dessus, puis on lui applique la fonction *aggregate*, paramétrée avec un facteur 5 pour descendre à 1 m de résolution (on crée des pixels de côté 5 fois plus long) et *max* comme fonction d'agrégation (les nouveaux pixels prennent la plus forte valeur parmi les 25 pixels d'origine). En effet, la hauteur maximale est la donnée pertinente pour l'exploitation ultérieure du MNH, donc *max* est la fonction adéquate. On peut remarquer que le temps d'exécution d'*aggregate* dépend fortement de la fonction d'agrégation choisie, *max* étant parmi les plus rapides.

Il suffit ensuite de soustraire à ce MNE découpé et agrégé la tuile MNT qui lui correspond et de sauver le résultat. Dans le cas où le MNT est incomplet, R calcule le MNH sur l'étendue commune aux deux rasters pour peu que l'origine soit la même (ce qui n'est pas un problème ici car les données manquantes sont au sud-est), mais n'affiche de messages pour le signaler qu'à la fin de la boucle, sans préciser quelle ligne est concernée. Pour vérifier plus facilement quelles tuiles sont concernées, on affiche les coordonnées, plus facilement interprétables, lorsque c'est le cas pour pouvoir vérifier qu'elles correspondent à une tuile au sud-est.

```
# on découpe la tuile MNE en fonction de la tuile MNT correspondante
for (i in 1:nrow(corresp_tout)){
  coins <- rbind(c(corresp_tout[i, coin_NO_x.x], corresp_tout[i, coin_NO_y.x]),
               c(corresp_tout[i, coin_NE_x.x], corresp_tout[i, coin_NE_y.x]),
               c(corresp_tout[i, coin_SE_x.x], corresp_tout[i, coin_SE_y.x]),
               c(corresp_tout[i, coin_SO_x.x], corresp_tout[i, coin_SO_y.x]))
  case <- SpatialPolygons(list(Polygons(list(Polygon(coins)), 1)),
                          proj4string = Alsace_93@proj4string)
  mne_20cm <- raster(corresp_tout[i, mne_20cm])
  mne_20cm <- crop(mne_20cm, case)

# on calcule sa version dégradée à une résolution de 1m (sauvée au format ASC)
  mne_1m <- aggregate(mne_20cm, fact = 5, fun = max, expand = TRUE, na.rm = TRUE,
                      filename = paste0("MNE_1m_", corresp_tout[i, coord], ".asc"),
                      overwrite = TRUE)
  corresp_tout[i, mne_1m := paste0("MNE_1m_", corresp_tout[i, coord], ".asc")]

# et on lui soustrait le MNT correspondant
  mnt <- raster(corresp_tout[i, mnt])
  mnh <- mne_1m - mnt
  writeRaster(mnh, paste0("MNH_", corresp_tout[i, coord], ".asc"), overwrite = TRUE)

# si le MNT n'est pas complet, on affiche les coordonnées
  if (extent(mnt) != extent(mne_1m) == TRUE){
    print(corresp_tout[i, coord])
  }
}
```

Comme *disaggregate*, *aggregate* impose de sauvegarder le raster obtenu (et ne permet pas de définir un chemin). De plus, le calcul du MNH génère des fichiers temporaires qui posent rapidement problème si la place manque sur le disque (ici C:/) où se trouvent les fichiers temporaires créés par R : avec des milliers de tuiles à créer, le disque peut saturer en quelques heures à peine. Lorsque c'est le cas, R continue à faire fonctionner le script sans pouvoir en écrire correctement les résultats, générant des tuiles vides repérables à leur poids anormalement faible, et n'affichant de messages d'erreur que si le travail est interrompu ou fini. En pratique, il a donc fallu supprimer ces fichiers temporaires plusieurs fois au cours de l'exécution du script.

Intersection du MNH avec la couche forêt

Seules les données MNH à l'intérieur des zones de forêt telles que définies dans la BD Forêt V2 (sans les landes) sont intéressantes pour la stratification. La dernière opération consiste donc à intersecter les tuiles MNH avec la couche forêt pour ne conserver que ces zones.

En réalité, cette étape pourrait être intégrée directement à la boucle précédente. Cependant, on a vu qu'il pouvait être nécessaire de vérifier au fur et à mesure la qualité des données obtenues. Lorsque ce n'est pas un problème, on peut rajouter l'opération d'intersection dans la boucle de calcul du MNH et ainsi éviter l'écriture d'un raster supplémentaire par tour de boucle.

Dans un premier temps, l'utilisation de la fonction *mask* était précédée de la fonction *crop* (beaucoup plus rapide) pour alléger le travail. Cependant, on obtient, pour certaines tuiles situées à l'extrême fin de la zone de travail (point sud par exemple), une tuile MNH avec moins de lignes ou de colonnes que la normale. Ceci ne porte pas à conséquence pour le MNH découpé, mais lors d'étapes de travail ultérieures les tuiles générées à partir de ces tuiles incomplètes se sont retrouvées avec un en-tête en contradiction avec le contenu (annonçant 40 lignes et 40 colonnes alors que c'était faux), ce qui les rend illisibles par QGIS. Ce problème ne se pose pas en utilisant directement *mask*.

```
for (i in 1:nrow(corresp_tout)){
  mnh <- raster(paste0("MNH_", corresp_tout[i, coord], ".asc"))
  mnh_foret <- mask(mnh, foret_93)
  writeRaster(mnh_foret, paste0("MNH_foret_", corresp_tout[i, coord], ".asc"),
              overwrite = TRUE)
}
```

Statistiques sur les MNH et constitution du modèle par classes de hauteur

Sophie Georges

28 juillet 2017

Objectifs

Cette deuxième phase de travail prépare directement la stratification proprement dite. Elle requiert d'avoir au préalable généré, à partir des données brutes disponibles sur le serveur, un MNH couvrant l'ensemble des zones forestières d'Alsace, sous forme de tuiles de 1 km de côté et 1 m de résolution, découpées à partir des données BD Forêt V2 pour ne garder strictement que la forêt au sens de la BD (cf. le document précédent, *Processus de création des tuiles MNH*). Dans la suite de ce document, ces tuiles découpées seront simplement appelées "MNH". On distingue ici trois grandes étapes :

- l'acquisition de données statistiques sur le MNH, dans le but de vérifier sa qualité, ainsi que connaître la répartition des valeurs pour créer un modèle simplifié qui soit pertinent.
- l'application au MNH de fonctions du package **inventR** pour obtenir des tuiles moins résolues (25 m) dont la valeur corresponde à l'une des variables descriptives habituellement utilisées en foresterie. La résolution de 25 m permet de se rapprocher de l'échelle des mesures réalisées sur les arbres des placettes de terrain (sur 15 m de rayon), afin que le modèle soit comparable à ces données. Une résolution de 30 m serait encore plus adaptée, mais 25 m est une division entière des tuiles existantes.
- le découpage de ce nouveau modèle à faible résolution en classes de hauteur.

Parallèlement, des statistiques sont faites sur la couche BD Forêt V2 pour déterminer quels essences ou mélanges d'essences correspondent en Alsace à des surfaces suffisantes pour qu'une stratification soit intéressante, et lesquelles sont au contraire à regrouper (à noter que dans la couche forêt/non forêt qui en est dérivée et a servi à découper le MNH, on a déjà exclu les landes). Cette opération ne fait pas intervenir R.

Statistiques sur tous le MNH

Pour cette étape, on utilise encore le MNH directement. On passe une fois de plus par des *data tables* autant que possible. Pour ce faire, on utilise une fonction qui extrait les valeurs d'une tuile donnée sous forme d'un histogramme avec des breaks définis par pas de 1. Les bornes doivent être fixées de manière à inclure les extrema réels, sinon la fonction ne marche pas (elles l'ont été par essais/erreurs). Au lieu d'afficher les résultats sous forme de graphe, comme la fonction *hist* le fait par défaut, on les stocke dans une *data table* avec, pour chaque valeur entière correspondant au centre d'une classe de largeur 1, le nombre de pixels qui y correspondent.

```
# on liste les tuiles MNH, découpées pour ne garder que la forêt au sens de BD Forêt V2
liste_MNH <- list.files(pattern = "\\\\.asc$", recursive = TRUE)
MNH_stats <- data.table(fichier = liste_MNH)

# on définit la fonction qui crée, pour un fichier donné, l'histogramme correspondant
# avec des breaks de 1 en 1 entre -100.5 et 400.5, et au lieu de dessiner
# l'histogramme (plot = FALSE pour l'empêcher), on stocke les résultats
extract_count <- fonction(fich) {
  z <- read_asc(fich)$z
  hist_z <- hist(z, breaks = seq(-100.5, 400.5, 1), plot = FALSE)
```

```

res <- data.table(centre_classe = -100:400,
                 count = hist_z$counts,
                 min = min(z, na.rm = TRUE),
                 max = max(z, na.rm = TRUE)
                )
return(res)
}

# on applique la fonction tuile par tuile
Stats_z <- MNH_stats[, extract_count(fichier), by = fichier]

```

Bien que l'on ait pu s'affranchir de boucles pour n'utiliser que le formalisme des *data tables*, la lecture des tuiles pour en récupérer les valeurs prend beaucoup de temps (les autres opérations sont très rapides). La table obtenue est par ailleurs trop lourde (3 993 972 lignes) pour pouvoir être lue. On s'en sert donc pour créer deux *data tables* plus exploitables : une qui donne seulement les valeurs minimale et maximale par tuile pour pouvoir remonter à celles qui présentent des valeurs aberrantes, et une qui synthétise le nombre d'observations par classe de valeur pour tout le MNH. Cette dernière permet notamment de réaliser des graphes et de lire rapidement les extrema.

```

# pour voir seulement les min et max par tuile
minmax <- unique(Stats_z[, .(fichier, min, max)])
# pour voir seulement le nombre d'observations par valeur
stats_resume <- Stats_z[, sum(count), by = centre_classe]

```

Ces données ont permis de se faire une meilleure idée de la qualité du MNH, étant donné que celui-ci est beaucoup trop lourd pour être affiché en entier dans QGIS. Ainsi, on sait maintenant que le calage vertical entre MNT et MNE est de qualité variable. 7 577 tuiles MNH sur 7 972 présentent des valeurs minimales négatives (inférieur ou égal à - 1 m pour "seulement" 5 546 d'entre elles). Pour 940 tuiles, le minimum atteint au moins - 10 m : près de 12 % des tuiles mais finalement seulement 1,3 % des pixels.

La consultation de certaines de ces tuiles repérées grâce à *minmax* dans QGIS confirme que ces valeurs négatives représentent dans certains cas une proportion relativement importante de leur surface, mais aussi qu'elles se situent pour l'essentiel au nord-ouest de la région, dans une zone issue de la numérisation de la BD Alti, soit la plus mauvaise qualité de données disponible. Ces statistiques montrent donc une qualité d'ensemble du MNH satisfaisante.

Application de fonctions sur le MNH

Fonctions personnalisées

Avant d'appliquer des fonctions statistiques sur le MNH, il faut les définir, ainsi que toute fonction annexe nécessaire. En effet, on souhaite se baser sur ce qui existe dans le cadre d'**inventR**, mais avec des besoins spécifiques qui imposent de redéfinir légèrement certaines fonctions.

Ainsi, plutôt que la hauteur brute, on souhaite utiliser une métrique qui serait fortement corrélée au volume. Pour ce faire, on s'inspire d'une variable très utilisée en foresterie : la hauteur dominante H_0 . Celle-ci correspond à la hauteur des plus gros arbres (soit globalement ceux qui forment la canopée), est généralement définie pour une zone donnée comme la hauteur moyenne des 100 plus gros arbres à partir d'un diagramme fréquence / diamètre. Il s'agit donc de données purement de terrain et très limitées spatialement, donc non accessibles directement à partir du MNH.

On peut cependant définir, à partir des données de hauteur du MNH, un modèle similaire en termes de corrélation au volume à celui qui est déjà disponible, la fonction *hdom_lid_onf*, pour calculer H_0 . Cette fonction utilise un coefficient multiplicateur, 1.003145, obtenu empiriquement à partir d'une zone d'étude

particulière. Aussi, même si c'est le mode de calcul retenu par l'ONF, il est trop spécifique pour nos besoins et on redéfinit la fonction sans ce coefficient, sous le nom *hdom_lid_1* : c'est un simple quantile à 99 % à partir de la hauteur, que par abus de langage nous considérerons comme H_0 .

```
hdom_lid_1 <- fonction (z, na.rm = TRUE)
{
  output <- quantile(z, probs = 0.99, na.rm = na.rm)
  return(output)
}
```

Les calculs principaux que l'on souhaite effectuer reposent sur la fonction *mnh_calc* : celle-ci charge un fichier raster au format .asc (avec la fonction **inventR** *read_asc*, qui en fait directement une *data table*) et lui applique une série de fonctions, dont la hauteur dominante. Le résultat est une data table qui contient les valeurs de chaque fonction pour des cellules d'une dimension définie par défaut à 25 m, avec des seuils maximum (70 m par défaut) et minimum (toujours 0). Comme on a modifié la fonction qui calcule H_0 , il faut aussi définir une nouvelle version de *mnh_calc* qui utilise *hdom_lid_1* au lieu de *hdom_lid_onf*. Étant donné les imprécisions du MNH, on en profite pour ajouter l'argument **zmin** au cas où il soit nécessaire de changer le seuil minimum (mais on le laisse à 0 pour cette fois).

```
mnh_calc_custom <- fonction (file, z_func = list(h0 = hdom_lid_1, hmoy = hmean_lid,
        hsd = hsd_lid, hgap = hgap_lid), pas = 25, zmax = 70, zmin = 0)
{
  if (!is.list(z_func)) {
    stop("z_func n'est pas une liste")
  }
  FUNS <- lapply(z_func, match.fun)
  cl <- lapply(FUNS, function(FUN) match.arg(arg = names(formals(FUN)),
        choices = "z", several.ok = TRUE))

  mnh <- read_asc(file)
  cs <- attr(mnh, "cellsize")
  if (any(pas <= cs)) {
    stop("Valeur de 'pas' trop faible")
  }
  xmin <- mnh[, min(x) - cs[1]/2]
  xmax <- mnh[, max(x) + cs[1]/2]
  ymin <- mnh[, min(y) - cs[2]/2]
  ymax <- mnh[, max(y) + cs[2]/2]
  dx <- diff(c(xmin, xmax))
  dy <- diff(c(ymin, ymax))
  nb_x <- floor(dx/pas)
  nb_y <- floor(dy/pas)
  if (nb_x * nb_y == 0) {
    stop("Valeur de 'pas' trop grande")
  }
  diff_x <- dx - (nb_x * pas)
  diff_y <- dy - (nb_y * pas)
  breaks_x <- seq.int(xmin + diff_x/2, xmax - diff_x/2, by = pas)
  breaks_y <- seq.int(ymin + diff_y/2, ymax - diff_y/2, by = pas)
  mnh[z > zmax, `:=`(z, NA)]
  mnh[z < zmin, `:=`(z, 0)]
  mnh[, `:=`(xcent = cut(x, breaks_x, labels = FALSE),
        ycent = cut(y, breaks_y, labels = FALSE))]
  mnh <- mnh[!is.na(xcent) & !is.na(ycent)]
  mnh[, `:=`(id, paste(xcent, ycent, sep = "-"))]
  setkey(mnh, id, xcent, ycent)
```

```

result <- mnh[, lapply(FUNS, function(fun) fun(z = z)),
               by = .(id, x = xmin + diff_x/2 + xcent * pas - pas/2,
                     y = ymin + diff_y/2 + ycent * pas - pas/2)]
return(result)
}

```

Ecriture des résultats sous forme de rasters

On souhaite maintenant non seulement calculer H_0 pour toutes les tuiles du MNH, mais récupérer les résultats sous forme de rasters au format `.asc` pour pouvoir les visualiser. La première étape consiste donc à appliquer `mnh_calc_custom` à toutes les tuiles MNH grâce à la table qui les liste (que l'on a créée plus haut pour servir de base aux statistiques). Ensuite, on exporte la donnée d'intérêt sous forme d'une nouvelle tuile de 25 m de résolution.

Cependant, `inventR` n'offre aucune fonction permettant d'exporter les résultats sous forme de fichiers raster. Par conséquent, la data table retournée par `mnh_calc` (ou `mnh_calc_custom`) n'a pas une structure adaptée. Il faut donc ajouter deux étapes essentielles :

- Rajouter les six lignes qui servent à définir le fichier `.asc` comme une couche d'information géographique : en fait, on commence par créer et sauver un fichier qui ne contient que ces informations et porte un nom basé sur les coordonnées présentes dans celui de la tuile MNH d'origine (de la forme `Hdom_x_y.asc`). Les données proprement dites seront ajoutées dans ce fichier avec `append`. Comme toutes les tuiles MNH de notre jeu n'ont pas exactement la même étendue (il peut manquer à certaines 2 m à l'est et/ou au sud, cf. *Processus de création des tuiles MNH*), il faut calculer le nombre de lignes et de colonnes pour qu'il soit correct. Le fixer à 40 lors d'une première tentative a posé problème avec ces tuiles MNH incomplètes. Pour plus de souplesse, on demande à l'utilisateur d'entrer la résolution souhaitée en m, qui devient aussi l'argument "pas" de `mnh_calc_custom`.
- Mettre les données dans le bon ordre : dans un fichier raster `.asc`, seul l'ordre des données détermine leur position dans la grille ; chaque ligne de texte représentant une ligne de la grille, les données sont triées par ligne (y) puis par colonne (x). Or la data table renvoyée par `mnh_calc_custom` a des données triées par colonne (x) puis par ligne (y), par l'intermédiaire d'un identifiant de la forme `x_y`. De plus, cet identifiant est de type texte, donc les numéros de ligne les plus petits sont en début de liste (alors qu'ils correspondent au sud, donc aux dernières lignes du fichier `.asc`) et `2_36` est après `11_36` par exemple. Il faut donc refaire un tri des données par y décroissant puis par x (en appliquant d'abord un tri par x , puis un tri par $-y$), sinon la tuile créée est aberrante.

```

# on demande à l'utilisateur la future résolution souhaitée
res <- as.numeric(readline("Résolution souhaitée en m :"))

for (i in 1:nrow(MNH_stats)){
  xll <- (as.numeric(substr(sub(".asc", "", sub("MNH_foret_", "", MNH_stats[i, fichier])),
                            1, 4))*1000 - 0.5
  yll <- (as.numeric(substr(sub(".asc", "", sub("MNH_foret_", "", MNH_stats[i, fichier])),
                            6, 9))*1000 - 999.5
  # on définit le nom de la future tuile simplifiée
  nom <- paste0("Hdom_", substr(sub(".asc", "", sub("MNH_foret_", "",
                                                MNH_stats[i, fichier])), 1, 4), "_",
                substr(sub(".asc", "",
                          sub("MNH_foret_", "", MNH_stats[i, fichier])), 6, 9), ".asc")

  # on récupère le nombre de lignes et de colonnes de la tuile MNH
  mnh <- read_asc(MNH_stats[i, fichier])
  nc_mnh <- attr(mnh, "ncols")
  nl_mnh <- attr(mnh, "nrows")
}

```

```

# et on en déduit le nombre de lignes et de colonnes final
nc <- nc_mnh %/% res
nl <- nl_mnh %/% res

# on crée les premières lignes du fichier .asc, qui permettent d'en faire une tuile
# raster manipulable sous QGIS
prefixe <- strsplit(paste("NCOLS ", nc, "\nNROWS ", nl, "\nXLLCORNER", xll,
                        "\nYLLCORNER", yll, "\nCELLSIZE ", res, "\nNODATA_value -9999"), "\n")
# on sauve la tuile (qui ne contient encore que ces premières lignes)
fwrite(prefixe, file = paste0("E:\\Documents SGeorges/stats/", nom))

# on applique la fonction mnh_calc_custom à la tuile MNH n°i : le résultat est
# une data table indexée sur id
hdom_i <- mnh_calc_custom(file = MNH_stats[i, fichier], pas = res, z_func = list
                        (h0 = hdom_lid_1, hmoy = hmean_lid, hsd = hsd_lid, hgap = hgap_lid))
# on trie cette data table par y décroissant, puis par x pour que les données
# soient dans le bon ordre dans le fichier .asc
hdom_i <- hdom_i[order(x)]
hdom_i <- hdom_i[order(-y)]

# on met le contenu de la colonne h0 (hauteur dominante) sous forme d'une matrice
# de 40 lignes x 40 colonnes, qui correspond au format d'un raster .asc
hdom_i_mat <- matrix(hdom_i[, h0], nrow = 40, byrow = TRUE)
# on rajoute ces données à la tuile "vide" créée plus haut
write.table(hdom_i_mat, row.names = FALSE, col.names = FALSE,
            file = paste0("E:\\Documents SGeorges/stats/", nom), na = "-9999", append = TRUE)
}

```

Une fonction d'écriture de rasters basée sur le bloc de code ci-dessus a depuis été ajoutée au package. Ici, on n'a effectué les opérations que pour H_0 , mais on pourrait facilement adapter le script pour récupérer les résultats d'une autre fonction, *hgap_lid* par exemple pour étudier les trouées, voire sauver les résultats de toutes les fonctions à la fois (avec une tuile raster différente par tuile MNH et par fonction).

Classes de hauteur

Le modèle de hauteur dont on dispose à présent est beaucoup moins lourd et permet de vérifier un peu mieux la qualité des données, mais pas encore finalement de visualiser toutes les données d'un coup, ni de les réunir en une seule couche qu'on puisse mettre en base de données. De toute façon, pour l'utiliser en stratification, il faudra à un moment donné définir des strates à partir de classes de hauteur. Aussi, on le fait dès maintenant sous R.

Les statistiques effectuées sur le modèle de hauteur initial ont servi à définir manuellement huit classes de hauteur. On utilise donc de nouvelles variantes de *hdom_lid_onf* et *mnh_calc* pour lui appliquer ces classes. D'abord, on définit *hdom_class*, qui au lieu de seulement calculer H_0 renvoie le numéro de la classe en fonction de la valeur de H_0 :

```

hdom_class <- fonction (z, na.rm = TRUE)
{
  output <- as.numeric(ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 3.5, 1,
                             ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 9.5, 2,
                             ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 14.5, 3,
                             ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 18.5, 4,
                             ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 21.5, 5,

```



```

        ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 25.5, 6,
              ifelse(quantile(z, probs = 0.99, na.rm = na.rm) < 30.5, 7,
                    8)))))))))
  return(output)
}

```

On s'en sert dans la définition de la nouvelle fonction *mnh_calc_class* (par ailleurs identique à *mnh_calc_custom*) :

```

mnh_calc_class <- function (file, z_func = list(h0 = hdom_class, hmoy = hmean_lid,
                                             hsd = hsd_lid, hgap = hgap_lid), pas = 25, zmax = 70, zmin = 0)

```

Et on l'applique au MNH comme précédemment, avec les mêmes méthodes pour sauver le résultat :

```

res <- as.numeric(readline("Résolution souhaitée en m :"))
mini <- as.numeric(readline("Hauteur minimale prise en compte :"))
maxi <- as.numeric(readline("Hauteur maximale prise en compte :"))

for (i in 1:nrow(MNH_stats)){
  xll <- (as.numeric(substr(sub(".asc", "", sub("MNH_foret_", "", MNH_stats[i, fichier])),
                            1, 4))*1000 - 0.5
        yll <- (as.numeric(substr(sub(".asc", "", sub("MNH_foret_", "", MNH_stats[i, fichier])),
                            6, 9))*1000 - 999.5
  nom <- paste0("Hclass_", substr(sub(".asc", "", sub("MNH_foret_", "",
        MNH_stats[i, fichier])), 1, 4), "_", substr(sub(".asc", "",
        sub("MNH_foret_", "", MNH_stats[i, fichier])), 6, 9), ".asc")

  mnh <- read_asc(MNH_stats[i, fichier])
  nc_mnh <- attr(mnh, "ncols")
  nl_mnh <- attr(mnh, "nrows")
  nc <- nc_mnh %/% res
  nl <- nl_mnh %/% res

  prefixe <- strsplit(paste("NCOLS ", nc, "\nNROWS ", nl, "\nXLLCORNER", xll,
        "\nYLLCORNER", yll, "\nCELLSIZE ", res,
        "\nNODATA_value -9999"), "\n")
  fwrite(prefixe, file = paste0("E:\\Documents SGeorges/stats/", nom))

  hdom_i <- mnh_calc_class(file = MNH_stats[i, fichier], pas = res, z_func = list
        (h0 = hdom_class, hmoy = hmean_lid, hsd = hsd_lid, hgap = hgap_lid),
        zmax = maxi, zmin = mini)
  hdom_i <- hdom_i[order(x)]
  hdom_i <- hdom_i[order(-y)]

  hdom_i_mat <- matrix(hdom_i[, h0], nrow = 40, byrow = TRUE)
  write.table(hdom_i_mat, row.names = FALSE, col.names = FALSE,
        file = paste0("E:\\Documents SGeorges/stats/", nom), na = "-9999",
        append = TRUE)
}

```

Le modèle résultant est cette fois suffisamment léger pour être manipulé en entier (les tuiles pèsent au plus 10 Ko, contre en moyenne 26 Ko à l'étape précédente et 16 Mo avant réduction de la résolution). Ces nouvelles tuiles ont été fusionnées en une couche unique qui a été mise en base de données.

Comparaisons entre stratifications

Sophie Georges

26 juillet 2017

Étapes

Les stratifications à tester

Après avoir mis en place les stratifications à tester dans l'application en ligne dédiée, on dispose de stratifications de phase 2 opérationnelles correspondant au croisement de quatre facteurs :

- échantillon annuel de points pour l'Alsace (2010 à 2015)
- choix d'une des trois stratifications de phase 1 créées en se basant sur cet échantillon :
 - stratification basée sur des regroupements d'essences effectués à partir de la BD Forêt V2 (sous QGIS) et codés REGRT en base de données
 - stratification simplifiée utilisant les types FAO reconstitués à partir de REGRT (auxquels il manque la catégorie "forêt ouverte", déjà regroupée vu les faibles surfaces)
 - stratification basée sur le modèle de huit classes de hauteur précédemment obtenu sous R (cf. "*Statistiques sur les MNH et constitution du modèle par classes de hauteur*") et codé HDOM_CL en base de données
- choix d'une des possibilités suivantes pour stratifier la phase 2 :
 - des classes issues de la photo-interprétation, utilisées en routine par l'inventaire (PI dans la suite du document)
 - REGRT, seul
 - REGRT croisé avec PI
 - HDOM_CL, seul
 - HDOM_CL croisé avec PI
- croisement supplémentaire ou non de ces strates de phase 2 avec les zones écologiques issues de GRECO fusionnées (Z_ECO en base), non regroupables entre elles.

On s'intéresse en fait à des données obtenues sur une fenêtre glissante de cinq ans, en regroupant les résultats de cinq stratifications consécutives (l'échantillon 2015, plus petit, a été écarté). Dans la suite du document, "stratification" correspond en fait sauf précision supplémentaire à ce groupe de cinq stratifications annuelles.

Cela correspond cependant tout de même à 30 combinaisons faisant intervenir de multiples variables. De plus, on souhaite calculer le facteur d'efficacité RE sur chaque composante de la variance. R permet toutes ces opérations de manière relativement simple, grâce à la manipulation des données sous forme de **data tables** et au package **inventR**, qui comporte justement une fonction dédiée pour appeler les stratifications et récupérer les données statistiques sous forme de *data tables* : *ocre_nm*.

Les différentes étapes

L'ensemble des scripts étant très long et répétitif, ne seront inclus que des exemples des tableaux créés. Les étapes de travail ont été effectuées dans l'ordre suivant : tout d'abord, toutes les stratifications ont été importées dans R, sans ventilation des résultats (c'est-à-dire un seul total global par stratification et par variable). À partir des *data tables* résultantes ont été construits des tableaux reprenant ces données, avec certaines stratifications qualifiées de référence et servant de point de comparaison. Ces références permettent de calculer RE comme suit :

$$RE = V(Y_{ref})/V(Y_{str})$$

où $V(Y)$ est l'estimation de la variance de la moyenne de la grandeur Y , str est la stratification testée et ref est la stratification de référence.

De plus, la variance se décompose en trois composantes :

- V_{SD} la variance introduite par l'incertitude sur la surface du sous-domaine
- V_{ST} la variance introduite par les erreurs d'échantillonnage sur les surfaces de strates estimées à partir du nombre de points
- V_{HT} la variance due à la variabilité intra-strate de la grandeur considérée, non définie si celle-ci est la surface du sous-domaine (celle-ci n'a qu'une seule valeur en tout, et non une par point). On calcule de même RE pour chaque composante.

L'importation et le calcul de données par *ocre_nm* ont été effectués sur cinq grandeurs différentes : surface du sous-domaine (c'est-à-dire de toute la forêt lorsqu'on ne ventile pas les résultats), volume de bois, surface terrière (la surface cumulée de la section des troncs), nombre de tiges et production de bois en volume. Cependant, les tableaux de comparaison et le calcul des RE n'ont finalement concerné que la surface et le volume de bois. Les résultats sont ensuite rendus sous forme de graphes pour être interprétés (présenté à la fin, car le code est pratiquement le même qu'à l'étape suivante).

À partir des premiers résultats, seules trois des stratifications initiales ont été conservées : la plus simple, utilisant les "types FAO" en phase 1 et PI en phase 2, comme première référence (la "référence simple"), la stratification utilisant REGRT seul à la fois en phase 1 et en phase 2, et celle utilisant HDOM_CL seul à la fois en phase 1 et en phase 2. Ces deux dernières ont pour particularité que, comme les surfaces des strates sont déjà connues en phase 2, V_{ST} est nulle ou presque et donc non étudiée. On introduit également comme référence supplémentaire la stratification utilisée par défaut par l'inventaire ("référence inventaire").

Cette fois, on ventile les données par un critère, soit au niveau placette, soit au niveau arbre (auquel cas la surface n'est pas disponible), soit les deux. On construit comme précédemment des tableaux de comparaison où RE est calculé par rapport à l'une des références, et on génère des graphes de la même manière. Cette fois, en plus de la surface et du volume, on étudie la production en volume.

Comparaisons systématiques sans ventilation

Importation des données

Tout d'abord, il faut importer dans R toutes les données correspondant aux stratifications. Tout est géré par la fonction *ocre_nm*, permettant de choisir les stratifications par leur nom, les années concernées (les deux doivent être cohérents et les stratifications avoir toutes la même structure) et les variables sur lesquelles faire le calcul. Pour faciliter cette étape, les stratifications annuelles ont reçu un nom similaire lors de leur création, de la forme *ALSACE_201x_FRE2* par exemple pour les stratifications avec la "nomenclature FAO" en phase 1 et REGRT en phase 2. Ce qui donne lors de l'importation :

```
fao_regrt <- ocre_nm(ans = 2010:2014,
  # pour toutes les années de 2010 à 2014 incluses
  vars=c("ST", "V", "GTOT", "NT", "PV"),
  # on importe les grandeurs surface, volume, surface terrière, nombre de
  # tiges et production en volume
  stratif = paste0("ALSACE_", 2010:2014, "_FRE2"),
  # pour les stratifications dont le nom correspond à ce patron
  detvar = TRUE, # avec le détail des composantes de la variance
  test = TRUE) # les stratifications sont sur le serveur de test
```

Le calcul prend environ 5 secondes par stratification, les résultats étant stockés définitivement dans la *data table* résultante.

Tableaux de comparaison

Une fois tous les résultats issus des stratifications disponibles sous forme de *data tables*, on peut librement faire des opérations sur celles-ci pour reformer des tableaux comparant, par exemple, les versions avec et sans Z_ECO de stratifications identiques par ailleurs : on calcule le RE en utilisant comme référence les versions avec Z_ECO (il vaudra 1 pour elles). Voici le début du code pour effectuer cette comparaison sur les volumes :

```
# pour initialiser la data table avec les 30 lignes nécessaires
lignes_z <- data.table(matrix("", ncol = 1, nrow = 30))

# on initialise ensuite la table vide avec tous ses champs
V_z_eco <- data.table(phase_1 = lignes_z[, V1], phase_2 = '', nbpoints = 0, tot = 0,
                     CV = 0, var_tot = 0, RE_tot = 0, vsd = 0, RE_vsd = 0, vst = 0,
                     RE_vst = 0, vht = 0, RE_vht = 0)

V_z_eco[1:10, phase_1 := "FAO"]
V_z_eco[1, phase_2 := "photo-interpretation avec Z_ECO"]
V_z_eco[1, nbpoints := as.numeric(fao_pi_z[1, nbpoints])] # sert à contrôler le résultat
# on récupère les données de la stratification pertinente
V_z_eco[1, tot := fao_pi_z[1, tot]]
# et on effectue de premiers calculs dessus
V_z_eco[1, CV := sqrt(fao_pi_z[1, var_tot])/fao_pi_z[1, tot]]
V_z_eco[1, var_tot := fao_pi_z[1, var_tot]]
# dans ce tableau de comparaison, cette stratification-ci est une référence, donc tous les
# RE calculés pour la ligne n°1 valent 1
V_z_eco[1, RE_tot := fao_pi_z[1, var_tot]/fao_pi_z[1, var_tot]]
V_z_eco[1, vsd := fao_pi_z[1, vsd]]
V_z_eco[1, RE_vsd := fao_pi_z[1, vsd]/fao_pi_z[1, vsd]]
V_z_eco[1, vst := fao_pi_z[1, vst]]
V_z_eco[1, RE_vst := fao_pi_z[1, vst]/fao_pi_z[1, vst]]
# cette ligne et la suivante sont omises quand la grandeur est une surface
V_z_eco[1, vht := fao_pi_z[1, vht]]
V_z_eco[1, RE_vht := fao_pi_z[1, vht]/fao_pi_z[1, vht]]

V_z_eco[2, phase_2 := "photo-interpretation sans Z_ECO"]
V_z_eco[2, nbpoints := as.numeric(fao_pi[1, nbpoints])]
V_z_eco[2, tot := fao_pi[1, tot]]
V_z_eco[2, CV := sqrt(fao_pi[1, var_tot])/fao_pi[1, tot]]
V_z_eco[2, var_tot := fao_pi[1, var_tot]]
# calculé en fonction de la stratification de la ligne n°1 (sa référence dans ce tableau)
V_z_eco[2, RE_tot := fao_pi_z[1, var_tot]/fao_pi[1, var_tot]]
V_z_eco[2, vsd := fao_pi[1, vsd]]
V_z_eco[2, RE_vsd := fao_pi_z[1, vsd]/fao_pi[1, vsd]]
V_z_eco[2, vst := fao_pi[1, vst]]
V_z_eco[2, RE_vst := fao_pi_z[1, vst]/fao_pi[1, vst]]
V_z_eco[2, vht := fao_pi[1, vht]]
V_z_eco[2, RE_vht := fao_pi_z[1, vht]/fao_pi[1, vht]]
```

Et ainsi de suite pour les 28 lignes suivantes. Le tableau *V_z_eco* permet ainsi par la suite de représenter sur un graphe le RE total ou par composante calculé pour les stratifications sans Z_ECO par rapport à leurs homologues avec Z_ECO. On fait ainsi varier un seul paramètre. C'est ce qui a permis d'écarter les stratifications avec Z_ECO dans les étapes suivantes : le RE des stratifications testées (sans Z_ECO) dans *V_z_eco* est supérieur à 1, ce qui signifie qu'on réduit la variance quand on enlève Z_ECO.

On construit de même S_{z_eco} , et d'autres tableaux qui font varier un autre paramètre (par la suite, les tableaux portant sur la même comparaison mais des grandeurs différentes, comme S_{z_eco} et V_{z_eco} , ont été réunis pour tout faire figurer sur un seul graphique). Étant donné le nombre de ces tableaux de comparaison, une *data table* a été construite pour en faire la liste.

Pour chaque variable (surface ou volume), on effectue donc les comparaisons suivantes :

- PI seule avec Z_ECO (référence) comparée aux autres phases 2 avec Z_ECO, à phase 1 identique
- même chose sans Z_ECO
- à phases 1 et 2 identiques, stratifications avec (référence) et sans Z_ECO
- PI seule avec Z_ECO comparée à PI associée à REGRT ou HDOM_CL (avec Z_ECO aussi), à phase 1 identique
- même chose sans Z_ECO
- inversement, REGRT ou HDOM_CL seuls comparés à PI associée à REGRT ou HDOM_CL (avec et sans Z_ECO), à phase 1 identique
- REGRT seul ou avec PI comparé à HDOM_CL avec ou sans PI (avec et sans Z_ECO), à phase 1 identique
- “types FAO” en phase 1 comparés à REGRT ou HDOM_CL en phase 1, avec Z_ECO, à phase 2 identique
- même chose sans Z_ECO
- REGRT en phase 1 comparé à HDOM_CL (avec et sans Z_ECO), à phase 2 identique.

Tous ces tableaux, étant des *data tables*, sont générés très vite.

Comparaisons ciblées avec ventilation

Importation des données ventilées

On utilise toujours *ocre_nm*, mais cette fois on n'a besoin que de quatre stratifications : les trois qui ont été retenues à l'issue des comparaisons systématiques, plus la stratification de référence de l'inventaire. Par contre, on rajoute un paramètre supplémentaire qui est la ventilation souhaitée. Celle-ci se fait d'après des paramètres déjà définis en base de données par une donnée et une unité, et permet de calculer les variables pour des sous-domaines comme “forêt privée” ou “essence principale chêne rouvre”. Il existe deux types de ventilation des données : par placette d'observation ou par arbre. Les ventilations par arbre ne permettent pas de calculer des surfaces. Dans certains cas, les mêmes catégories sont disponibles pour les deux types de ventilation. Dans l'exemple suivant, on ventile les données en deux sous-domaines : essence principale feuillue, et essence principale conifère. Si l'on veut ventiler au niveau placette :

```
# "nomenclature FAO" en phase 1 et photo-interprétation seule en phase 2 (1e référence)
ref_fc <- ocre_nm(ans = 2010:2014, vent_p = "ESSPRE_FR:ESSD1",
  # ventilation par placette, donnée ESSPRE_FR (essence principale
  # feuillue ou conifère), unité ESSD1 (classement des essences en
  # feuillus et conifères)
  vars=c("S", "V", "GTOT", "NT", "PV"),
  stratif = paste0("ALSACE_", 2010:2014, "_FPI2"),
  detvar = TRUE, test = TRUE)

# stratification par défaut de l'inventaire (2e référence)
ref2_fc <- ocre_nm(ans = 2010:2014, vent_p = c("DEP:RA:42", "ESSPRE_FR:ESSD1"),
  # cette stratification étant définie France entière, il faut en plus la
```

```

# ventiler par unité administrative (DEP), région administrative à 22
# régions (RA) Alsace (42)
vars = c("S", "V", "GTOT", "NT", "PV"),
detvar = TRUE)

# regroupements d'essences seuls en phase 1 et en phase 2
r_fc <- ocre_nm(ans = 2010:2014, vent_p = "ESSPRE_FR:ESSD1", vars = c("S", "V", "GTOT",
"NT", "PV"), stratif = paste0("ALSACE_", 2010:2014, "_RRE2"),
detvar = TRUE, test = TRUE)

# classes de hauteur seules en phase 1 et en phase 2
h_fc <- ocre_nm(ans = 2010:2014, vent_p = "ESSPRE_FR:ESSD1", vars = c("S", "V", "GTOT",
"NT", "PV"), stratif = paste0("ALSACE_", 2010:2014, "_HHC2"),
detvar = TRUE, test = TRUE)

```

Et pour obtenir les mêmes sous-domaines avec une ventilation au niveau arbre :

```

# "nomenclature FAO" en phase 1 et photo-interprétation seule en phase 2 (1e référence)
ref_fc_a <- ocre_nm(ans = 2010:2014, vent_a = "FR", # noter que FR suffit
vars=c("V", "GTOT", "NT", "PV"), # pas de surface cette fois
stratif = paste0("ALSACE_", 2010:2014, "_FPI2"),
detvar = TRUE, test = TRUE)

# stratification par défaut de l'inventaire (2e référence)
ref2_fc_a <- ocre_nm(ans = 2010:2014, vent_p = "DEP:RA:42", vent_a = "FR", vars=c("V",
"GTOT", "NT", "PV"), detvar = TRUE)

# regroupements d'essences seuls en phase 1 et en phase 2
r_fc_a <- ocre_nm(ans = 2010:2014, vent_a = "FR", vars=c("V", "GTOT", "NT", "PV"),
stratif = paste0("ALSACE_", 2010:2014, "_RRE2"),
detvar = TRUE, test = TRUE)

# classes de hauteur seules en phase 1 et en phase 2
h_fc_a <- ocre_nm(ans = 2010:2014, vent_a = "FR", vars=c("V", "GTOT", "NT", "PV"),
stratif = paste0("ALSACE_", 2010:2014, "_HHC2"),
detvar = TRUE, test = TRUE)

```

Tableaux de comparaison

Comme lors des comparaisons systématiques, on crée des tableaux de comparaison où l'on calcule le RE global et par composante. Pour chaque ventilation et chaque variable, on en fait en fait un pour chaque référence. Certains ont pu être rédigés de manière plus compacte en faisant appel à des boucles pour lire les tables issues de *ocre_nm*, en utilisant des expressions complexes pour associer la bonne ligne de ces dernières à la ligne correspondante de la nouvelle data table. Par exemple, avec les catégories d'exploitabilité :

```

# ventilation par exploitabilité (niveau placette), 1e référence
# tableaux de comparaison : volumes
lignes_exp <- data.table(matrix("", ncol = 1, nrow = 12))
V_exp <- data.table(stratification = lignes_exp[, V1], exploite = '', nbpoints = 0,
tot = 0, CV = 0, var_tot = 0, RE_tot = 0, vsd = 0, RE_vsd = 0,
vst = 0, RE_vst = 0, vht = 0, RE_vht = 0, pourc_vsd = 0,
pourc_vst = 0, pourc_vht = 0)
# en plus, on calcule la proportion de chaque composante dans l'ensemble de la variance

```

```

for (i in 1:nrow(V_exp)){
  if (((i - 1) %% 3) %% 5 == 0){
    # on définit un indicateur qui servira à associer les lignes (trouvé empiriquement)
    j = ((i - 1) %% 3)^2 - 4 + (6 - ((i - 1) %% 3)^2 + 1) %% 5
  } else {
    j = ((i - 1) %% 3)^2 + 1 + (6 - ((i - 1) %% 3)^2 + 1) %% 5
  }
  if (i %% 3 == 1) { # lignes pour lesquelles on veut les données de la référence
    V_exp[i, stratification := "réfêrence"]
    V_exp[i, exploit := ref_exp[j, expl.expl]]
    V_exp[i, nbpoints := as.numeric(ref_exp[j, nbpoints])]
    V_exp[i, tot := ref_exp[j, tot]]
    V_exp[i, CV := sqrt(ref_exp[j, var_tot])/ref_exp[j, tot]]
    V_exp[i, var_tot := ref_exp[j, var_tot]]
    V_exp[i, vsd := ref_exp[j, vsd]]
    V_exp[i, vst := ref_exp[j, vst]]
    V_exp[i, vht := ref_exp[j, vht]]
  }
  else if (i %% 3 == 2) {
    V_exp[i, stratification := "REGRT"]
    V_exp[i, exploit := r_exp[j, expl.expl]]
    V_exp[i, nbpoints := as.numeric(r_exp[j, nbpoints])]
    V_exp[i, tot := r_exp[j, tot]]
    V_exp[i, CV := sqrt(r_exp[j, var_tot])/r_exp[j, tot]]
    V_exp[i, var_tot := r_exp[j, var_tot]]
    V_exp[i, vsd := r_exp[j, vsd]]
    V_exp[i, vst := r_exp[j, vst]]
    V_exp[i, vht := r_exp[j, vht]]
  }
  else {
    V_exp[i, stratification := "HDOM_CL"]
    V_exp[i, exploit := h_exp[j, expl.expl]]
    V_exp[i, nbpoints := as.numeric(h_exp[j, nbpoints])]
    V_exp[i, tot := h_exp[j, tot]]
    V_exp[i, CV := sqrt(h_exp[j, var_tot])/h_exp[j, tot]]
    V_exp[i, var_tot := h_exp[j, var_tot]]
    V_exp[i, vsd := h_exp[j, vsd]]
    V_exp[i, vst := h_exp[j, vst]]
    V_exp[i, vht := h_exp[j, vht]]
  }
  # cette partie est valable pour toutes les lignes
  V_exp[i, pourc_vsd := (V_exp[i, vsd]/V_exp[i, var_tot])*100]
  V_exp[i, pourc_vst := (V_exp[i, vst]/V_exp[i, var_tot])*100]
  V_exp[i, pourc_vht := (V_exp[i, vht]/V_exp[i, var_tot])*100]
}

# on peut maintenant calculer les RE
for (i in 1:nrow(V_exp)){
  if (i %% 3 == 1) {
    V_exp[i, RE_tot := V_exp[i, var_tot]/V_exp[i, var_tot]]
    V_exp[i, RE_vsd := V_exp[i, vsd]/V_exp[i, vsd]]
    V_exp[i, RE_vst := V_exp[i, vst]/V_exp[i, vst]]
    V_exp[i, RE_vht := V_exp[i, vht]/V_exp[i, vht]]
  }
}

```

```

}
if (i %% 3 == 2) {
  V_exp[i, RE_tot := V_exp[i - 1, var_tot]/V_exp[i, var_tot]]
  V_exp[i, RE_vsd := V_exp[i - 1, vsd]/V_exp[i, vsd]]
  V_exp[i, RE_vst := V_exp[i - 1, vst]/V_exp[i, vst]]
  V_exp[i, RE_vht := V_exp[i - 1, vht]/V_exp[i, vht]]
}
if (i %% 3 == 0) {
  V_exp[i, RE_tot := V_exp[i - 2, var_tot]/V_exp[i, var_tot]]
  V_exp[i, RE_vsd := V_exp[i - 2, vsd]/V_exp[i, vsd]]
  V_exp[i, RE_vst := V_exp[i - 2, vst]/V_exp[i, vst]]
  V_exp[i, RE_vht := V_exp[i - 2, vht]/V_exp[i, vht]]
}
}

# on remplace les codes par des noms de catégories explicites, pour les graphes
V_exp[, exploit := ifelse(V_exp[, exploit] == '0', 'très facile',
                          ifelse(V_exp[, exploit] == '1', 'facile',
                                  ifelse(V_exp[, exploit] == '2', 'moyenne',
                                          'difficile')))]

```

Graphes

On utilise ici le package **ggplot2**, qui permet de sauvegarder le graphe et ses options sous forme de liste pour l'afficher et l'exporter n'importe quand ultérieurement. Il est complété par le package **cowplot** qui rajoute certaines options graphiques, comme les étiquettes obliques. Par exemple, si l'on veut représenter le RE global des stratifications avec HDOM_CL en phase 2 par rapport à celles avec REGRT, sans ventilation (tableau V_r2vh2) :

```

vr2vh2_tot <- ggplot(V_r2vh2[!grep("REGRT", phase_2)]) +
  # [!grep("REGRT", phase_2)] permet d'éliminer du graphe les stratifications dont le nom
  # contient "REGRT" : ce sont les références (RE = 1), elles alourdiraient le graphe
  aes(x = factor(phase_2,
                 levels = unique(phase_2)), # les phases 2 sont en abscisse
       y = RE_tot,                          # les étiquettes sont aussi à prendre dans cette colonne
       col = phase_1) +                     # le RE global est en ordonnée
  xlab('Stratification') +                 # les couleurs des points dépendent de la phase 1
  geom_point(size = 2) +                   # nom de l'axe des x
  geom_hline(yintercept = 1) +             # taille des points (on peut aussi changer de symbole)
  theme_bw() +                             # la ligne horizontale à 1 est utile pour lire les RE
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
  # un des thèmes visuels proposés par cowplot
  # on incline le texte de 60° pour que les longues étiquettes restent lisibles

```

Et de même pour les données ventilées. Le graphe est un peu différent si l'on s'intéresse au CV, ici avec les surfaces estimées pour le détail des essences principales et la deuxième référence :

```

sess2_CV <- ggplot(S_ess2) + # on affiche aussi les références
  aes(x = factor(essence, levels = unique(essence)), y = CV, col = stratification) +
  xlab("Essence principale") +
  geom_point(size = 2) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
# pas de ligne à y = 1

```


Il est souvent intéressant d'obtenir un seul graphe avec différentes données côte à côte, par exemple ici le RE pour la V_{SD} pour une même ventilation sur la surface, le volume et la production. Pour ce faire, on utilise l'argument *facet* à partir d'un tableau combinant toutes les données :

```
# on rajoute aux tableaux la colonne var indiquant la grandeur qu'ils traitent
S_fc2[, var := "Surface"]
V_fc2[, var := "Volume"]
PV_fc2[, var := "Production en volume"]
# et on les assemble en un seul tableau (fill = TRUE assure que les colonnes VHT
# qui n'existent pas dans S_fc2 soient présentes, vides là où var = "Surface")
Comp_fc2 <- rbindlist(list(S_fc2, V_fc2, PV_fc2), fill = TRUE)
# on force l'ordre dans lequel seront présentées les données dans le graphe
Comp_fc2$var<-factor(Comp_fc2$var, levels=c("Surface", "Volume", "Production en volume"))

fc2_vsd <- ggplot(Comp_fc2[stratification != "réf rence"]) +
  aes(x = factor(essence, levels=unique(essence)), y = RE_vsd, col = stratification) +
  geom_point(size = 2) +
  # on peut indiquer manuellement une  chelle de couleurs
  scale_color_manual(values=c("#E69F00", "blue", "slategray")) +
  geom_hline(yintercept = 1) +
  theme_bw() +
  # expression permet ici de mettre des caract res (SD) en indice ([])
  labs(x = "Essence fr quente", y = expression('RE pour la V'[SD]),
        col = "Stratification\n(phases 1 et 2)") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

On exporte ensuite facilement le r sultat au format PNG (par exemple), sans m me avoir besoin de le visualiser dans R. La taille et la r solution du graphe sont enti rement param trables, offrant un meilleur rendu que les options d'export de RStudio. Ici, pour le graphe ci-dessus, qui est particuli rement large :

```
ggsave("4_stratifs/essence/ess2_CV_surf.png", plot = sess2_CV, width = 8,
        height = 5, dpi = 200)
```

4. Diagramme de Gantt

