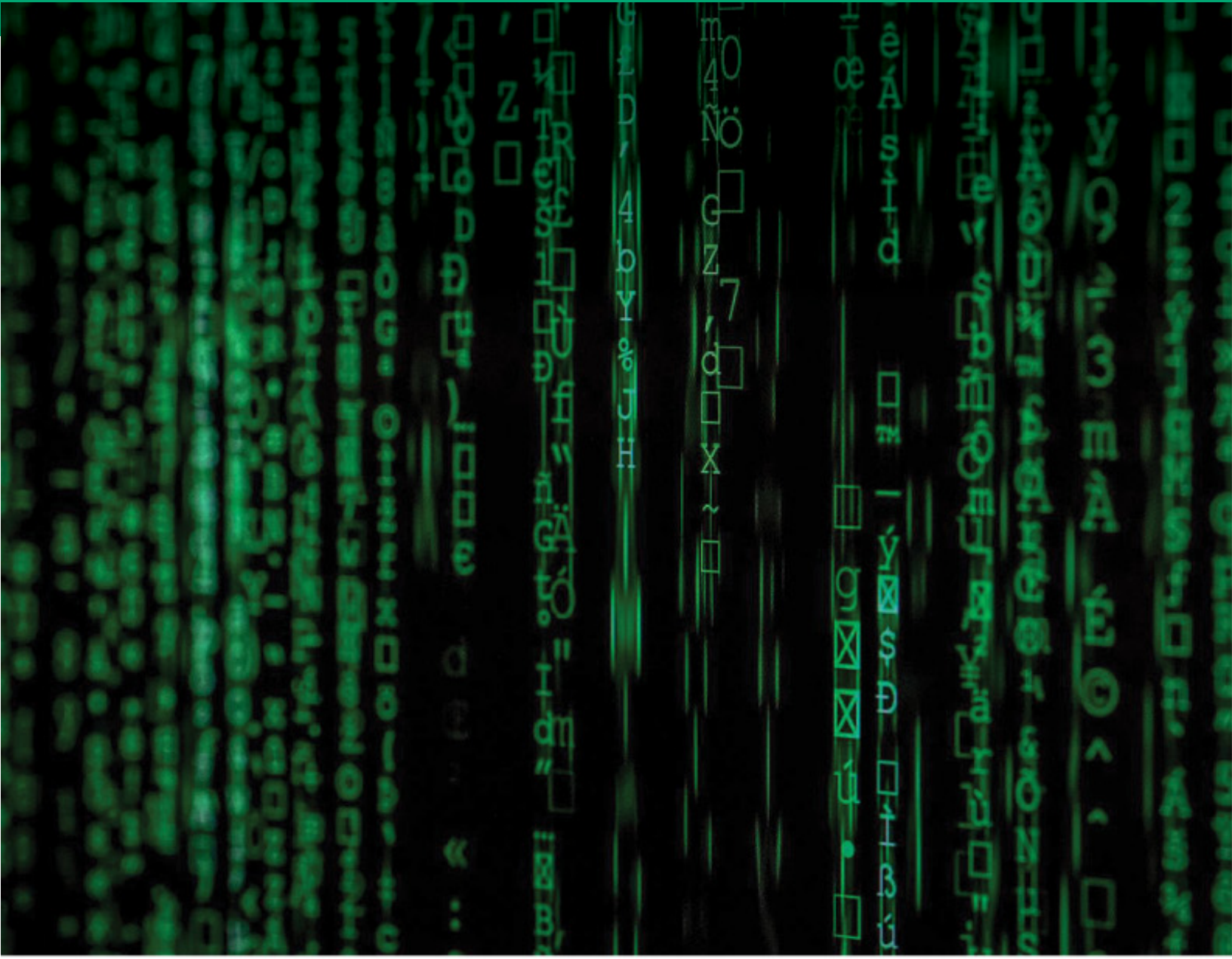


Cours 7

Version textuelle

Éléments de l’analyse de données

La version interactive de cette cour est disponible gratuitement à l’adresse suivante :

https://elearning.fao.org/?lang=fr

Ce cours offre une orientation relative aux approches et calculs typiques utilisés dans les analyses de données forestières et de sujets connexes.

À qui ce cours s'adresse-t-il?

Ce cours s'adresse principalement aux personnes impliquées dans l'analyse des données du suivi des forêts, mais peut être suivi par quiconque s'intéresse au sujet. Ce cours vise particulièrement:

1. Les techniciens forestiers responsables de la mise en œuvre des IFN de leur pays.
2. Les praticiens de l'analyse de données forestières.
3. Les équipes du suivi national des forêts.
4. Les étudiants et les chercheurs, en tant que matériel programmatique dans les écoles forestières et les cours universitaires.
5. Les jeunes et les nouvelles générations forestières.

Structure du cours

Ce cours comprend cinq leçons.

Leçon 1: Introduction à l'analyse de données
Cette leçon introduit les questions qui sont importantes pour l'analyse de données typique après la collecte et le nettoyage de données, mais doivent aussi être considérées durant tout le processus de planification et de mise en œuvre de l'inventaire.
Leçon 2: Estimation
Cette leçon propose une vue d'ensemble du processus de génération de résultats (ou estimations) à partir des données d'échantillon. Rappelons, cependant, que cette leçon n'apporte que des éléments de base – elle ne couvre pas le sujet de l'estimation statistique de manière exhaustive. Si vous êtes un expert intéressé par le sujet à un niveau plus approfondi, ou traitez des analyses de données d'IFN régulièrement, nous recommandons de compléter cette leçon avec des manuels, et/ou de discuter de vos approches avec des statisticiens des inventaires forestiers expérimentés.
Leçon 3 : Modèles statistiques dans le suivi des forêts
Cette leçon apporte des éléments concernant l'utilisation de modèles statistiques et aborde des

questions qui doivent être considérées lors de leur utilisation.

Leçon 4: Erreurs dans le suivi des forêts

Cette leçon aborde les diverses erreurs aléatoires qui surviennent le long du processus d'IFN. Elle décrit aussi la propagation d'erreur – comment les différentes sources d'erreur se propagent à l'erreur totale du résultat final.

Leçon 5: Produits typiques des analyses de données

Cette leçon traite des produits typiques des analyses de données dans le suivi des forêts et aborde les principaux produits générés par l'analyses de données des IFN.

À propos de la série

Ce cours conclut une série de huit cours individualisés couvrant divers aspects d'un IFN. Voici un aperçu de la série complète.

Cours	Apprentissages
Cours 1: Pourquoi un inventaire forestier national (IFN)?	Objectifs et but d'un IFN, et comment les IFN informent la conception de politiques et la prise de décisions dans le secteur forestier
Cours 2: Préparation d'un inventaire forestier national	La planification et le travail nécessaire pour mettre en place un IFN efficace ou un système national de suivi des forêts (SNSF).
Cours 3: Introduction à l'échantillonnage	Aspects généraux de l'échantillonnage dans les inventaires forestiers.
Cours 4: Introduction au travail de terrain	Considérations pour le travail de terrain, les variables au niveau parcellaire et les mesures au niveau de l'arbre.
Cours 5: Gestion de données dans un inventaire forestier national	Collecte d'information et gestion de données pour les IFN.
Cours 6: Assurance qualité et contrôle qualité dans un inventaire forestier national	Procédures d'AQ et de CQ dans la collecte et la gestion de données d'un inventaire forestier.

☞ **Cours 7: Éléments de l'analyse de données**	**(Vous suivez actuellement ce cours).**
Cours 8: Résultats de l'inventaire forestier national: notification et diffusion	Publication des résultats de l'IFN et importance de la notification dans le contexte des actions REDD+.

Leçon 1: Introduction à l'analyse de données

Introduction de la leçon

Dans cette leçon, nous étudierons des questions qui sont importantes pour l'analyse de données typique, non seulement après la collecte et le nettoyage de données, mais doivent aussi être considérées durant tout le processus de planification et de mise en œuvre de l'inventaire.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire l'importance de l'analyse de données dans diverses phases d'un inventaire forestier.
2. Expliquer les principes généraux de l'analyse de données.
3. Décrire le nettoyage des données et les considérations connexes.

Analyse de données dans les diverses phases d'un inventaire forestier

Bien que l'analyse de données soit pertinente dans diverses phases d'un inventaire forestier, la **phase d'analyse de données en soi** a lieu **entre la collecte de données et la publication des résultats**, autrement dit, une fois que les données ont été enregistrées, organisées et nettoyées. Le produit final de l'analyse de données cherche à satisfaire les questions posées dans l'évaluation des besoins en information (EBI).

Néanmoins, les considérations relatives à l'analyse de données sont pertinentes durant tout le processus d'inventaire car l'un des objectifs transversaux de tout inventaire forestier est de générer une base de données pertinente et fiable d'apport pour les analyses. Finalement, **la qualité des données co-détermine la qualité des produits.** Voyons maintenant le rôle de l'analyse de données durant la planification et la collecte de données.

Considération durant la phase de planification

- Il est nécessaire de s'assurer que toutes les variables (qui sont nécessaires pour produire les produits ciblés) font partie du protocole d'inventaire.
- À moins que l'on anticipe que les données seront utilisées à l'avenir pour des questions émergentes potentielles, il est recommandé d'éviter d'enregistrer des variables qui ne sont pas nécessaires dans les analyses.

- Les variables nécessaires doivent être observées et enregistrées de sorte que les exigences de précision sont remplies et des résultats sont générés pour les unités de référence cibles.
- Le plan d'échantillonnage et le plan parcellaire doivent être définis pour assurer que des estimateurs sont disponibles pour l'analyse statistique.
- Des protocoles d'assurance qualité pour les données sont nécessaires. Ils incluent:
    - l'organisation de formations adéquates pour les équipes de terrain à la fois avant et pendant la collecte de données;
    - la définition claire et transparente des normes de qualité de données; et
    - les mécanismes de contrôle adéquats.

Principes généraux de l'analyse de données

Les analyses de données dans les projets d'inventaire forestier suivent les mêmes principes que tout le processus d'inventaire - elles doivent être **méthodologiquement fondées**, **cohérentes**, **complètes** et **documentées avec transparence**. Toutes les étapes d'analyse doivent être justifiables et conformes à la conception de l'inventaire (en termes de plan d'échantillonnage, plan parcellaire et modèles utilisés).

Réponse aux besoins en information

Les analyses de données doivent traiter tous les besoins en information qui ont été formulés avant la collecte de données et, autant que possible, les questions qui ont émergé depuis.

Analyses pour l'optimisation de la conception future

Les analyses de données peuvent aussi s'étendre aux questions méthodologiques et pratiques qui appuient la planification efficace des inventaires de suivi. Cela peut inclure des études longitudinales (de suivi temporel), évaluant la consommation de temps dans différentes étapes de l'inventaire, et/ou l'évaluation du plan d'échantillonnage et du plan parcellaire dans le but d'identifier des optimisations potentielles dans la mise en œuvre des inventaires de suivi.

Contre-vérification de toutes les analyses

Lorsque l'on analyse les données pour les produits ciblés, il est important de contre-vérifier la correction de tous les résultats, y compris des résultats intermédiaires.

Le principe de base correspondant peut être formulé selon les mots de Sutherland (1996), «ne croyez jamais vos résultats», qui signifie qu'au lieu de croire, il faut être sûr et pleinement comprendre les résultats et les hypothèses sous-jacentes.

Tout doute, même mineur, doit être suivi, et cela est vrai pour les résultats qui semblent suspicieux comme pour ceux qui apparaissent entièrement plausibles et crédibles.

Documentation

Chaque étape d'analyse doit être correctement documentée dans le rapport d'inventaire, généralement dans un volume à part. Idéalement, la documentation doit contenir tous les estimateurs, les calculs étape par étape, la description des modèles, les facteurs de conversion et les systèmes d'indicateur utilisés, et doit aborder les défis de l'analyse. Pour conclure, rappelons que la documentation concordera avec la méthodologie initiale définie avec le protocole d'échantillonnage.

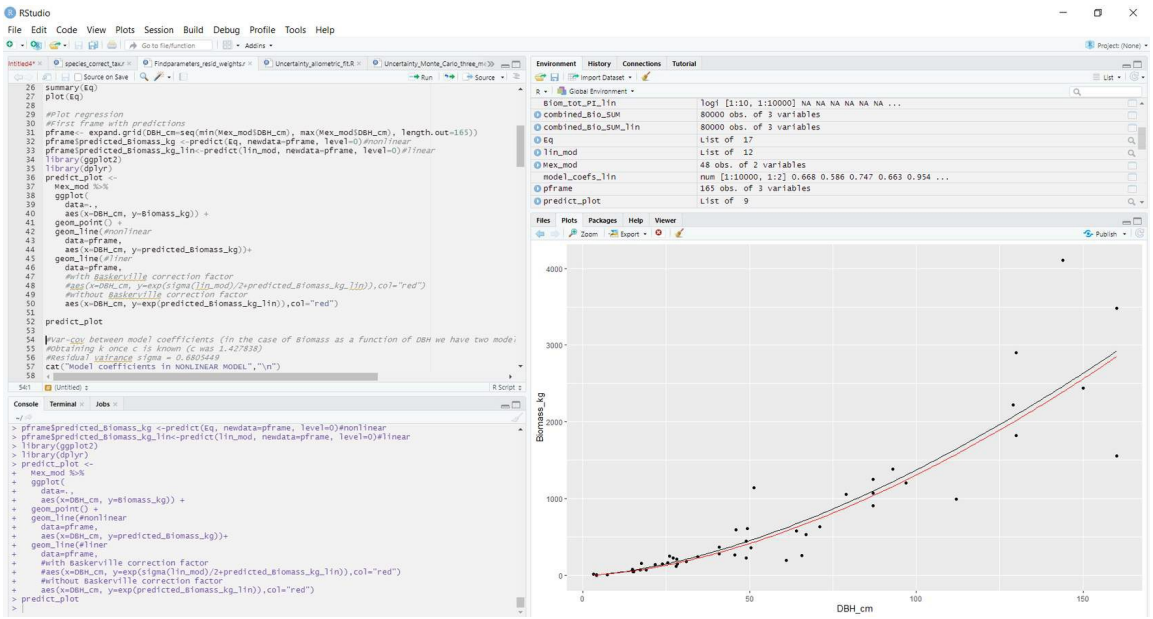
Nettoyage des données

Les analyses de données ne peuvent commencer que si les données sont cohérentes et nettoyées, et si toutes les erreurs et les incohérences identifiables ont été éliminées. C'est en définitive la qualité des données qui détermine la qualité finale des produits.

Néanmoins, des erreurs de données peuvent parfois ne pas être identifiées durant le processus courant de nettoyage, et n'apparaissent que lorsque les résultats ne semblent pas plausibles. Il est alors nécessaire de revisiter le processus de nettoyage des données pour identifier les erreurs potentielles. Il est utile de rappeler ici que l'absence de plausibilité n'est pas toujours une erreur, et que des variabilités inattendues peuvent toujours exister!

Considérations sur les logiciels pour l'analyse de données

Cette dernière décennie a connu un développement rapide de solutions logicielles – à la fois en général, et plus spécifiquement pour les IFN. Alors que les inventaires forestiers plus anciens étaient évalués par des programmes sur mesure, dans un langage de programmation ou dans un paquet logiciel statistique, actuellement la tendance consiste à mettre au point des scripts R et à utiliser des paquets R déjà existants (bibliothèques de code spécialisées dans la conduite de tâches particulières) le plus possible.



Exemple de script R généré en utilisant RStudio, un environnement de développement pour R, un langage de programmation pour l'informatique statistique est les graphiques.

Pour les petits inventaires (impliquant des ensembles de données plus réduits), les analyses peuvent aussi être mises en œuvre avec des feuilles de calculs (comme dans Microsoft Excel), des scripts R et un logiciel spécifique comme ceux mis au point par la FAO. Des programmeurs compétents sont nécessaires pour ces tâches, qui peuvent travailler en collaboration avec les experts de l'inventaire, qui définissent les produits ciblés.

Les étapes d'analyse singulières ou les flux de travail complets peuvent aussi être résolus en mettant en œuvre des estimateurs statistiques dans un modèle de données adapté utilisant un logiciel de Business Intelligence (BI, pour intelligence économique, en anglais) ou traités directement dans un système de gestion de base de données.

Le saviez-vous?

Le rôle des logiciels de BI dans l'analyse de données forestières

La plupart des logiciels de BI ne couvre pas les bons estimateurs à appliquer pour produire les résultats. Par conséquent, si un logiciel de BI doit être utilisé, ces estimateurs doivent être

correctement inclus sous forme de formules. Jusqu'ici, cependant, cela n'est pas courant dans la plupart des pays. L'IFN de l'Allemagne calcule actuellement toutes les estimations en utilisant la syntaxe de SQL Server directement sur la base de données.

En tout cas, il n'existe de pas logiciel unique, mais seulement des procédures spécifiques qui peuvent être mises en œuvre pour chaque inventaire forestier reflétant exactement la conception de l'inventaire. Comme chaque inventaire nécessaire un logiciel complexe nouveau (ou au moins adapté), des vérifications rigoureuses sont nécessaires pour assurer que les résultats sont corrects. Demander à deux analystes de données de mener les mêmes analyses indépendamment, puis comparer les résultats, est une bonne idée.

Dans certains cas, ces résultats peuvent avoir l'air corrects, plausibles et cohérents, et répondent aux attentes des experts de l'inventaire, mais ils restent imparfaits. Rappelons que tous les résultats doivent être contre-vérifiés.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Bien que les considérations relatives à l'analyse de données soient pertinentes durant tout le processus d'inventaire, l'analyse de données en soi a lieu après la collecte de données et est un prérequis pour la notification des résultats.
- Les analyses de données dans les projets d'inventaire forestier suivent les mêmes principes que tout le processus d'inventaire – elles doivent être méthodologiquement fondées, cohérentes, complètes et documentées avec transparence.
- Des erreurs de données peuvent parfois ne pas être identifiées durant le processus courant de nettoyage, et n'apparaissent que lorsque les résultats ne semblent pas plausibles. Cela implique de revisiter le processus de nettoyage des données pour identifier les erreurs potentielles.
- Alors que les inventaires forestiers plus anciens étaient évalués par des programmes sur mesure, actuellement (2023) la tendance consiste à mettre au point des scripts R et à utiliser des paquets R déjà existants (bibliothèques de code spécialisées dans la conduite de tâches particulières) le plus possible.

- Il n'existe de pas logiciel unique, mais seulement des procédures spécifiques qui peuvent être mises en œuvre pour chaque inventaire forestier reflétant exactement la conception de l'inventaire.

Leçon 2: Estimation

Introduction de la leçon

La collecte de données de terrain (et certaines analyses fondées sur la télédétection) dépend fortement de l'échantillonnage statistique. L'estimation est le processus qui génère des résultats (ou estimations) à partir des données d'échantillon. En tant que telle, l'estimation est fondamentale dans les analyses de données.

Selon le plan d'échantillonnage et le plan parcellaire utilisés, l'estimation peut être simple mais aussi très complexe. Cette leçon offre une introduction à l'estimation fondée sur un échantillon – si vous souhaitez aborder ce sujet plus en profondeur, veuillez consulter des manuels ou discuter avec des techniciens des inventaires forestiers.

Plus de détails sur cette question sont abordés dans le **Cours 3: Introduction à l'échantillonnage**.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire le rôle des estimations dans les analyses de données d'IFN.
2. Énumérer les principes généraux de l'estimation statistique.
3. Expliquer les estimations ponctuelles et par intervalle.
4. Discuter du rôle des données auxiliaires dans l'estimation des inventaires forestiers.

Observations générales sur les estimations

Les IFN utilisent diverses sources de données – depuis le suivi de terrain à partir d'échantillons (soit le cœur du suivi des forêts), jusqu'à la technologie de télédétection de pointe. Tous les produits des analyses de données d'IFN sont donc des estimations, et proviennent soit:

1. d'observations des variables cibles sur les parcelles d'échantillonnage de terrain (les estimations fondées sur un calcul);
2. de l'appui de données auxiliaires, souvent issues du SIG ou de la télédétection (les estimations assistées par modèle; voir la leçon 3 de ce cours pour une explication complète); soit
3. entièrement de modèles (estimations fondées sur un modèle).

Note

Rappelons que **tous les résultats des études par échantillonnage sont des estimations**, qu'il s'agisse de moyennes, de variances, d'intervalles de confiance, de régressions ou de corrélations. Il est donc préférable d'utiliser une terminologie claire. Par exemple, au lieu de conclure *L'analyse de l'IFN a montré que la forêt couvre 43,5 pour cent du pays*, il est plus exact de dire *L'IFN estime le couvert forestier dans le pays à 43,5 pour cent*.

Les estimations sont des variables aléatoires. Cela signifie qu'elles sont contraires à des valeurs paramétriques fixes dans la population, qui sont constantes. Toutes les estimations suivent une distribution avec une valeur moyenne de cette distribution (la valeur attendue pour laquelle l'estimation ponctuelle est l'approximation à partir d'échantillons tirée de cette distribution) et un écart-type qui décrit la variabilité dans cette distribution des valeurs moyennes estimées (estimé grâce à l'estimation par intervalle).

Dans le cas hypothétique de répétition d'un IFN avec exactement le même plan mais une randomisation différente, on produirait des résultats numériques différentes des estimations, à la fois pour les estimations ponctuelles et par intervalle.

Les estimations servent à connaître la population de sorte que l'intérêt principal ne réside pas tant dans les données d'échantillon elles-mêmes, mais dans l'inférence de la vraie valeur de la population que l'échantillon permet. Plus l'erreur-type sera petite, plus l'on pourra supposer que les estimations sont – en moyenne – proches de la vraie valeur paramétrique. Dans ce cas, on considérera que l'estimation est fiable.

Puisque l'on infère la (vraie) valeur de la population à partir de l'échantillon (estimé), les termes estimations fondées sur un calcul, estimations assistées par modèle et estimations fondées sur un modèle sont fréquemment remplacés par inférence fondée sur un calcul, inférence assistée par modèle et inférence fondée sur un modèle.

Principes généraux de l'estimation statistique

Lorsque l'on réalise une estimation sur des bases statistiques (au contraire d'estimations subjectives), les calculs doivent strictement correspondre au plan d'échantillonnage et au plan parcellaire utilisés – ce qui signifie que différents experts tendront à obtenir les mêmes résultats. Les formules utilisées pour l'estimation sont les estimateurs. Il existe des cas où plus d'un estimateur alternatif est disponible, mais généralement il n'y a pas de choix.

L'une des principales caractéristiques des plans d'échantillonnage d'inventaire dans les IFN est que les estimateurs utilisés doivent être non biaisés (ou au moins approximativement non biaisé dans le cas de l'estimateur par ratio). Cela signifie que la valeur attendue de notre plan d'échantillonnage doit être identique à la valeur de la population recherchée. La valeur attendue est la valeur que donnera la moyenne dans le cas (hypothétique) où l'on répèterait l'étude par échantillonnage de nombreuses fois – avec le même plan mais une randomisation différente.

Les plans et les estimateurs que nous avons présentés dans d'autres cours de cette séries (notamment dans le **Cours 3: Introduction à l'échantillonnage**) sont presque tous non biaisés, avec trois exceptions remarquables::

1. L'estimateur par ratio est approximativement non biaisé dans certaines circonstances.
2. Il n'y a pas d'estimateur non biaisé pour la variance d'erreur dans l'échantillonnage systématique, alors qu'il y a des estimateurs non biaisés pour la moyenne.
3. Il n'y a pas d'approche non biaisée pour analyser les parcelles d'échantillonnage qui incluent les arbres k les plus proches (non couvert dans ce cours mais expliqué dans des manuels spécialisés) – un plan parcellaire souvent utilisé dans les études écologiques mais moins dans les inventaires forestiers.

Si l'un de ces éléments de conception sont utilisés dans un inventaire, les questions générées par l'utilisation de ces estimateurs biaisés doivent être traitées avec transparence. Cela est particulièrement important pour l'échantillonnage systématique car c'est l'un des plans d'échantillonnage les plus fréquemment utilisés dans les IFN. Ici, nous resterons prudents en appliquant le cadre d'estimateur d'un échantillonnage aléatoire simple (EAS) pour l'estimation de la variance d'erreur, car l'on sait que cette approche donne une estimation conservatrice et surestime toujours la vraie variance d'erreur (même si l'on ne sait pas dans quelle mesure).

Estimations ponctuelles et par intervalle: générer des estimations sur la position et la dispersion

Dans l'analyse des données des IFN, on s'intéresse principalement aux estimations ponctuelles et par intervalle. Dans cette section, nous nous pencherons sur celles-ci.

L'estimation ponctuelle informe le point sur l'axe numérique où se situe l'estimation (par ex. pour la biomasse aérienne sur une base de surface, 200 Mg/ha), et l'estimation par intervalle informe la variabilité estimée de cette estimation ponctuelle (par ex. ET% = 5 %). Si l'on utilise la terminologie des statistiques descriptives, on peut aussi dire que: l'estimation ponctuelle est une mesure de la position de l'estimation, tandis que l'estimation par intervalle est une mesure de la dispersion des estimations.

Lorsque l'on se réfère aux estimations ponctuelles, il s'agit souvent de la valeur moyenne, mais aussi de l'estimation d'un coefficient de régression (b1) ou d'un coefficient de régression (r), r, ou de la variance de la population (s^2) qui sont des estimations ponctuelles.

Les estimations ponctuelles produisent l'information centrale pour les utilisateurs des données. Les personnes non expertes concentrent principalement leur interprétation des résultats sur ces mesures de position des estimations. Il est important de réitérer que ces estimations ponctuelles ne sont pas vérité, mais seulement des estimations.

Le fait qu'aucune estimation ponctuelle ne sera identique au vrai paramètre de la population n'est pas l'expression d'un biais mais bien de la variabilité dans l'échantillonnage, ou d'une erreur d'échantillonnage. Cependant, l'écart entre cette estimation particulière (dérivée de notre étude par échantillonnage ou IFN), en termes numériques, et le vrai paramètre de la population reste inconnu.

Distribution des estimations ponctuelles

Pour permettre cette inférence probabiliste de la vraie valeur de la population, on doit connaître la distribution des estimations ponctuelles. Par exemple, pour des valeurs moyennes, on sait qu'elles suivent la loi t de Student pour les petits échantillons et la loi normale pour les plus grands échantillons; où grand, en statistiques, est généralement défini comme $n \geq 30$ (pour un n grand, la loi t de Student s'approche de la loi normale).

Lorsque l'on sait que les valeurs moyennes varient selon la loi normale autour de la valeur attendue (la vraie moyenne en cas d'estimateur non biaisé), on peut utiliser les densités de probabilité dans cette loi

normale pour estimer la probabilité que la vraie valeur se trouve dans un intervalle défini autour de la moyenne estimée qui provient de notre étude par échantillonnage (IFN).

Pour la moyenne estimée, cet intervalle est symétrique autour de la moyenne estimée et a un écart-type qui correspond à l'erreur-type. Il faut ici se souvenir que l'erreur-type est calculée à partir de la variance de la population estimée, s^2y.

Cette variance de la population estimée est une estimation elle-même (l'estimation à partir d'échantillons de la variance de la population): elle peut être considérée comme une estimation ponctuelle (la valeur de la variance de la population estimée) à laquelle est liée une estimation par intervalle (la variabilité des variances de population). On peut aussi estimer les intervalles de confiance pour la variance de la population estimée, et pour les variances estimées, l'intervalle de confiance sera asymétrique, car les variances estimées suivent la distribution F (asymétrique).

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

Les estimations par intervalle sont aussi importantes dans les IFN car elles sont une mesure de précision de l'estimation et, par conséquent, de l'incertitude, ce qui est couramment interprété comme une mesure de la fiabilité des estimations.

Méthodes d'auto-amorçage et du couteau suisse pour les estimations par intervalle dans un plan complexe

Dans certains plans d'inventaire où un estimateur est trop complexe, un rééchantillonnage est indiqué. Le rééchantillonnage est une simulation dans laquelle les données d'échantillon sont exploitées pour simuler plusieurs échantillons (sous-échantillons) et des inférences sont établies à partir des résultats correspondants pour les statistiques de l'étude par échantillonnage complète.

L'auto-amorçage (en anglais, bootstrap) est la technique la plus souvent utilisée lorsque les intervalles de confiance sont déterminés dans des plans complexes où des estimateurs non biaisés directs ne sont pas disponibles. Cela remonte à 1979, lorsque Bradley Efron a inventé et introduit cette modification de la méthode du couteau suisse (en anglais, jackknife) déjà proposée depuis longtemps par Quenouille

(1956), le fameux rééchantillonnage «un contre tous» (leave-one-out, en anglais).
Ces techniques s'appuient sur des simulations et non pas sur des hypothèses concernant les paramètres d'une distribution spécifique des estimations; elles sont ainsi appelées techniques non paramétriques.

Le principe de l'auto-amorçage suit l'idée évoquée de rééchantillonnage à partir de la taille de l'échantillon n qui a été prise. On peut procéder «avec ou sans remise».

L'auto-amorçage «sans remise» signifie qu'un grand nombre de fois un sous-échantillon de taille n est pris dans la taille de l'échantillon original, et ceux qui ont été sélectionnés ne sont pas remis dans l'ensemble dans lequel s'effectue la sélection (en d'autres termes, ils ne peuvent être sélectionnés qu'une seule fois).

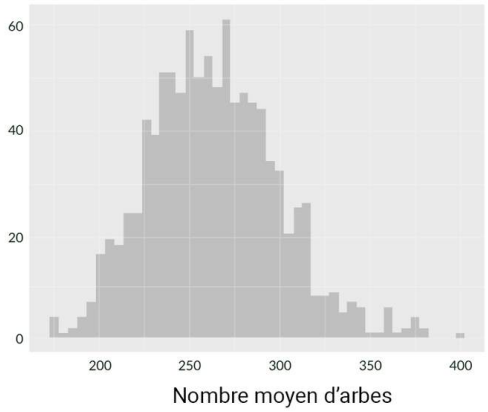
L'auto-amorçage «avec remise», cependant, signifie que la même observation peut être sélectionnée plusieurs fois dans l'échantillon d'amorçage. Ici, le terme grand signifie que cet échantillonnage est répété plusieurs milliers de fois, par exemple 10 000 fois.

Table 1
Échantillon aléatoire du nombre d'arbres dans 25 1 ha. parcelles dans une forêt

63 84 91 92 124 145 152 154 162 164 174 189 192
223 269 294 323 344 354 358 368 372 463 477 900

Table 2
Ré-échantillons d'amorçage de la Table 1 — Ré-échantillonnage

Échantillon 1	63 84 92 92 152 152 152 154 162 174 189 192 223 269 294 354 354 358 358 368 372 463 477 477 900
Échantillon 2	84 91 92 154 154 164 174 174 174 189 192 223 294 294 323 358 358 358 368 372 372 463 477 900 900
⋮ 1000 échantillons	
Échantillon 1000	63 92 145 152 152 154 154 154 154 162 164 164 189 192 192 223 223 294 344 368 368 463 477 900 900



Pour chacun de ces échantillons d'amorçage, la statistique cible (par ex. la valeur moyenne) est calculée de sorte qu'à la fin on ait, par exemple, 10 000 valeurs moyennes d'amorçage qui peuvent être représentées graphiquement sous forme de distribution. Cette distribution a une valeur moyenne (qui correspond bien sûr à la valeur moyenne de l'échantillon original de taille n) et une distribution particulière des densités de probabilité à partir de laquelle, pour chaque probabilité, des intervalles de confiance peuvent être dérivés.

Si, par exemple, les limites d'un intervalles de confiance de 95 pour cent doivent être calculées, on cherche les points de découpage où 2,5 pour cent des moyennes d'amorçage se trouvent dans la partie

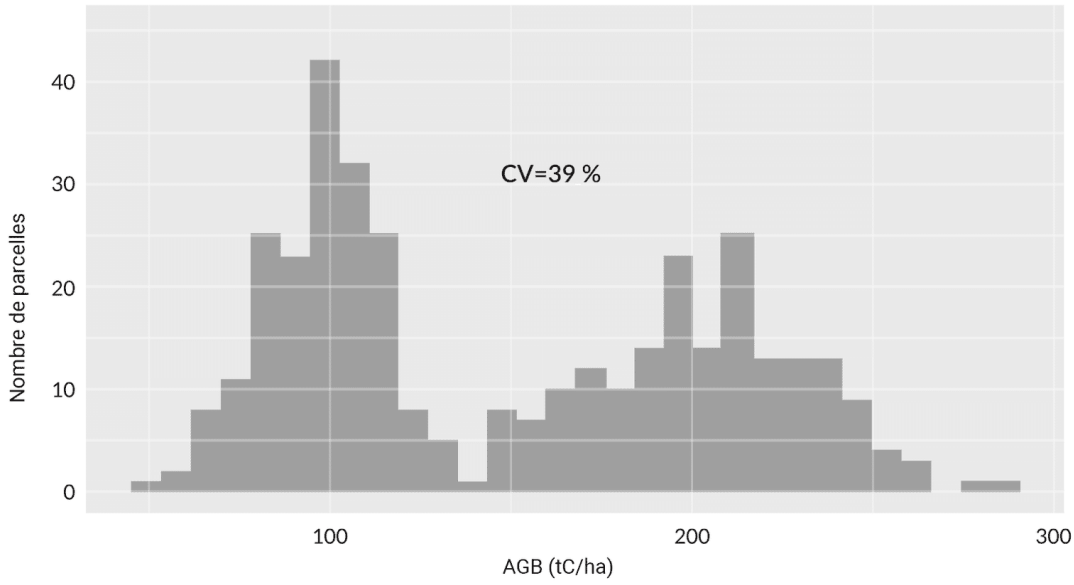
supérieure et 2,5 pour cent sont dans la partie inférieure. Les deux «points de découpage» sont alors pris comme limites supérieure et inférieure de l'intervalle de confiance de 95 pour cent.

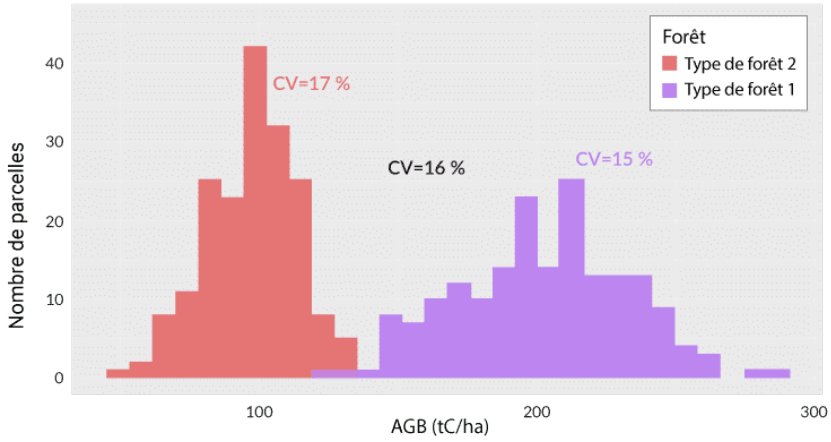
Données auxiliaires dans l'estimation des inventaires forestiers

Les données auxiliaires proviennent de l'observation de variables auxiliaires dans certains plans d'inventaire pour améliorer la précision de l'estimation des variables cibles. Les variables auxiliaires sont aussi appelées co-variables ou variables subordonnées (en latin: *auxilium* = aide, *ancilla* = serviteur). Jusqu'ici, nous avons vu des variables auxiliaires dans les estimateurs par ratio et par régression où nous avons saisi une forte corrélation entre les variables cibles et auxiliaires pour extraire et intégrer l'information de la variable auxiliaire dans l'estimation de la variable cible.

Plus généralement, lorsque l'on fait une distinction uniquement entre variables cibles et auxiliaires, on peut considérer plusieurs autres variables comme auxiliaires (appuyant les analyses). Cette observation se réfère par exemple à toutes les variables topographiques qui servent à séparer les résultats de nos variables cibles en classes – par exemple, matériel sur pied par classe d'altitude ou biomasse par classe de pente. Ici, les variables auxiliaires définissent des critères pour une post-stratification, permettant des analyses spécifiques et une évaluation des relations entre les variables cibles et auxiliaires.

Voyons maintenant deux figures sur l'amélioration des estimations d'erreur à travers la post-stratification à l'aide de données auxiliaires.





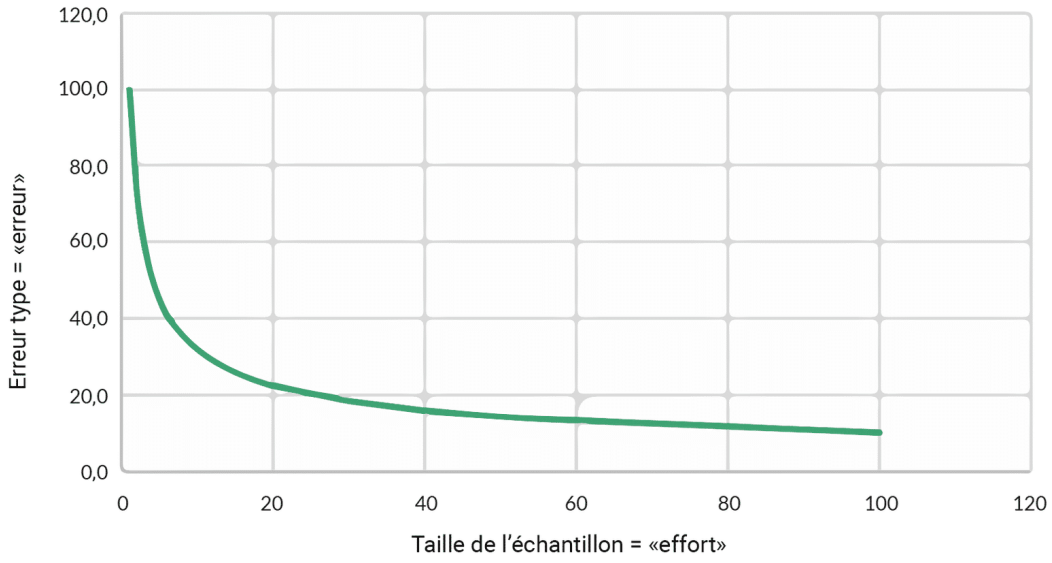
Estimations pour différentes unités de référence/sous-populations

L'unité de référence de base pour l'estimation dans les IFN est le pays entier. La taille de l'échantillon est généralement définie de manière à ce que la précision de l'estimation remplisse les attentes à l'échelle du pays. La taille de l'échantillon est généralement assez grande et les estimations seront précises, avec des erreurs-types faibles. Selon la taille de l'échantillon, les erreurs-types relatives sont dans certains cas inférieures à 1 pour cent. Mais une précision aussi élevée n'arrive que lorsque l'on prend le pays entier comme unité de référence, soit l'aire pour laquelle les résultats sont publiés.

Souvent, les estimations pour les unités infranationales – comme les provinces, les états ou les territoires – sont aussi intéressantes. Bien entendu, lorsque l'on utilise la même grille systématique de points échantillons, la taille de l'échantillon pour ces unités de références plus petites sera moindre, et les erreurs-types pour les estimations correspondantes seront plus élevées. **Plus l'unité de référence et la taille de l'échantillon sont petites, et moins les estimations seront précises.** Par exemple, dans l'IFN de l'Allemagne, alors que la surface forestière pour l'ensemble du pays présente une erreur-type relative de ET% = 0,7 pour cent avec une taille d'échantillon d'environ n = 21 000 clusters, elle est de ET% = 1,6 pour cent pour l'État fédéral de Bavière (n = 2 815) et de ET% = 25,8 pour cent pour la combinaison des États fédéraux de Hambourg et Brème avec uniquement n = 15.

Il est instructif de considérer ici la simple relation entre la taille de l'échantillon et l'erreur-type dans l'EAS, illustrée par la figure ci-dessous; où la forme de base de la relation est valables pour tous les plans d'échantillonnage: le gain marginal de précision pour les grandes taille d'échantillon, mais les petits changements de la taille de l'échantillon ont un impact beaucoup plus important sur l'erreur-type dans

les tailles d'échantillon plus petites!!



Lorsque des données de télédétection sont disponibles, des estimations pour les aires plus petites peuvent être générées avec une précision bien supérieure en utilisant ce que l'on appelle l'«estimation des petites surfaces», où les données de télédétection sont utilisées pour appuyer la génération d'estimations pour des unités de références (presque) arbitrairement petites.

Bien qu'une définition de l'estimation des petites surfaces soit fournie dans la leçon 5 de ce cours, on peut ici apportée un exemple simple: Prenons une grande surface forestière de 1 000 hectares, où l'on veut estimer la densité moyenne des arbres (nombre d'arbre par demi-hectare) pour une petite aire de seulement 10 hectares dans la forêt. Mais l'on ne dispose de données que pour 5 parcelles d'échantillonnage dans la petite surface, ce qui n'est pas suffisant pour une estimation précise de la densité des arbres.

En utilisant une estimation de petite surface avec des données de télédétection, on peut utiliser l'information de l'aire forestière plus grande pour générer une estimation plus précise de la densité des arbres dans la petite surface. On peut utiliser des données de télédétection, comme l'imagerie par satellite et le balayage de détection et estimation de la distance par la lumière (LiDAR) pour dériver de l'information supplémentaire sur la densité des arbres et d'autres caractéristiques de la forêt dans la surface forestière plus grande.

On peut alors utiliser ces données de télédétection comme appui pour générer une estimation pour la plus petite aire d'intérêt. Par exemple, les données de télédétection peuvent suggérer que la densité

moyenne des arbres dans surface forestière plus grande est de 400 arbres par demi-hectare. En utilisant une estimation de petite aire, on peut ajuster cette estimation à partir des données de la petite aire pour obtenir une estimation plus exacte pour celle-ci. Par exemple, le modèle peut estimer que la densité moyenne des arbres dans la petite aire est de 450 arbres par demi-hectare, avec une marge d'erreur plus faible que l'on ce que l'on pourrait obtenir en utilisant uniquement les données d'échantillon.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Tous les résultats des études par échantillonnage sont des estimations, qu'il s'agisse de moyennes, de variances, d'intervalles de confiance, de régressions ou de corrélations.
- Les estimations servent à connaître la population: l'intérêt principal ne réside pas tant dans les données d'échantillon elles-mêmes, mais dans l'utilisation de ces données d'échantillon pour inférer la vraie valeur de la population.
- Lorsque l'on réalise une estimation sur des bases statistiques, les calculs doivent strictement correspondre au plan d'échantillonnage et au plan parcellaire utilisés, ce qui signifie que:
    - différents experts devraient obtenir les mêmes résultats; et
    - l'on n'est pas libre de réaliser l'estimation avec une approche arbitraire.
- Dans l'analyse des échantillons d'IFN, on cherche le plus possible à utiliser des estimateurs non biaisés. Pour certains plans, ces estimateurs n'existent pas, y compris les estimateurs pour la variance d'erreur dans l'échantillonnage systématique.
- L'estimation ponctuelle est une mesure de la position, tandis que l'estimation par intervalle est une mesure de la dispersion. Cette terminologie est valable pour les variables comme pour les estimations.
- Dans certains plans d'inventaire où un estimateur est trop complexe, une simulation dans laquelle les données d'échantillon sont exploitées pour simuler plusieurs échantillons, ce que l'on appelle un «rééchantillonnage», peut être indiquée.
- Dans certaines conditions, il est également efficace d'observer des variables auxiliaires avec les variables cibles afin d'améliorer la précision de l'estimation des variables cibles.

Leçon 3: Modèles statistiques dans le suivi des forêts

Introduction de la leçon

Les modèles statistiques sont omniprésents dans le suivi des forêts.
Cette leçon apporte une vue d'ensemble et des éléments concernant l'utilisation de modèles statistiques dans les IFN et aborde des questions qui doivent être considérées lors de leur utilisation.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire le rôle des modèles statistiques dans le suivi des forêts.
2. Identifier les principales caractéristiques des modèles statistiques.
3. Démontrer comment construire un modèle de biomasse.
4. Identifier un modèle statistique adéquat pour une situation particulière.
5. Expliquer certains problèmes de publication des résultats courants dans les modèles statistiques.

Qu'est-ce qu'un modèle statistique?

Les modèles statistiques cherchent à établir une relation quantitative entre une variable réponse (ou variable dépendante) et une ou plusieurs variables explicatives (ou variables indépendantes). En d'autres termes, en ayant mesuré/observé la (ou les) variable(s) explicative(s), le modèle est utilisé pour générer une valeur pour la variable réponse. Essentiellement, un modèle statistique prédit une valeur pour une variable cible. Dans le contexte des inventaires forestiers, les modèles statistiques sont utilisés lorsque:

1. une variable cible ne peut pas être mesurée dans un inventaire forestier (par ex. la biomasse ne peut pas être mesurée par pesée. Si vous coupez tous les arbres de la forêt pour les peser, vous n'avez plus de forêt! Néanmoins, la biomasse peut être modélisée à partir de la mesure d'autres variables dites explicatives); ou
2. une mesure est exigeante en termes de temps/de coût (par ex. la hauteur prend du temps à mesurer, et elle est souvent mesurée uniquement pour un sous-ensemble d'arbres puis prédite à partir d'un modèle pour les autres arbres comme fonction du dhp).

Les modèles statistiques décrivent la relation entre les données/observations de deux variables, fréquemment observées à partir des mêmes objets (comme les arbres). Les modèles statistiques ne servent pas à établir une relation de cause à effet, ce qui serait le but de ce que l'on appelle les modèles de processus: ils visent à inclure explicitement les causes de processus biologiques afin de prédire des produits spécifiques dans différentes situations. Bien qu'une telle relation puisse exister aussi pour les modèles statistiques, ce n'est pas la question de cet exercice de modélisation, et un modèle statistique ne peut pas être interprété de la sorte.

Exemple de modèles statistiques utilisés dans le suivi des forêts

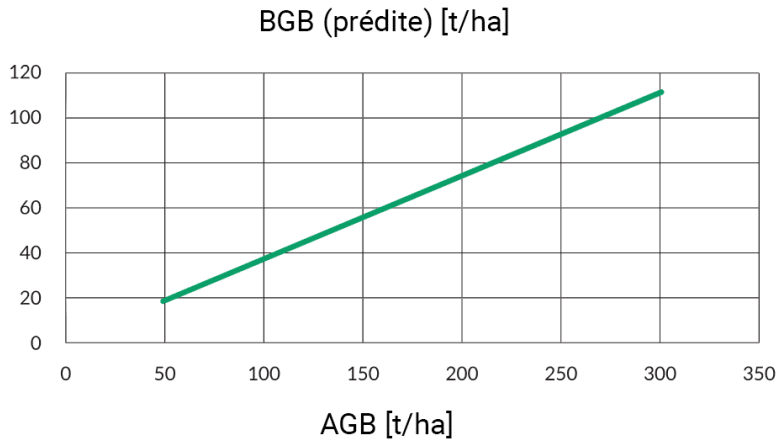
Il existe différents types de modèles – de complexité variable – utilisés dans le suivi des forêts, et dans certains cas, il est difficile de remarquer qu'un modèle a été utilisé.

Exemple 1: Déterminer la surface terrière à partir d'une mesure du diamètre

Un modèle de base, par exemple, est l'hypothèse selon laquelle la structure transversale des arbres à hauteur de poitrine sera toujours un cercle: la surface terrière des arbres est calculée selon un le modèle circulaire simple. Bien sûr, dans la réalité, aucun arbre n'est un cylindre mathématiquement parfait, et aucune surface transversale à hauteur de poitrine ne constitue un cercle mathématiquement parfait, mais l'approximation donne des résultats raisonnables jusqu'ici, et il n'y a pas de meilleure solution.

Exemple 2: Facteurs de conversion comme modèles statistiques

D'autres modèles de base sont de simples facteurs qui sont fréquemment utilisés pour convertir – par exemple – la biomasse de la tige en biomasse totale ou la biomasse aérienne (AGB) en biomasse souterraine (BGB). Les coefficients de forme – un indicateur sommaire de la forme d'un tronc – sont aussi des modèles simples, utilisés pour déterminer le volume des arbres individuels. On peut dire que ces facteurs de conversion sont juste des modèles de régression linéaires simples «réduits», où l'ordonnée à l'origine est zéro et le facteur lui-même est le coefficient de pente. La figure ci-dessous fournit un exemple.



Un simple facteur de conversion peut être considéré comme un modèle de base qui permet de prédire une variable à partir d'une variable explicative. Le GIEC (2006, table 4.4) recommande par exemple le facteur de conversion de 0,37 pour les forêts ombrophiles tropicales pour déterminer la BGB à partir de la AGB. Ce facteur de conversion se traduit par un modèle de régression linéaire simple avec un coefficient d'ordonnée à l'origine de zéro et un coefficient de pente de 0,37: BGB = 0,37 * AGB.

Remarquons que dans ce cas, le GIEC donne une source de ce facteur de conversion mais ne publie pas l'erreur-type ou d'autres mesures d'incertitude. À la place, il spécifie la plage de valeurs du facteur de conversion pour certains types de forêt.
Bien sûr, une incertitude considérable provient de ces simples facteurs de conversion, mais dans de nombreux cas, comme pour la biomasse souterraine, il serait impossible de prélever ses propres échantillons.

Exemple 3: Modèles de régression courants

Les modèles de régression courants utilisés dans les inventaires forestiers incluent:

1. la prévision de la hauteur à partir du *dhp* (courbes de hauteurs); et
2. la prévision du volume/ de la biomasse / du carbone à partir du dhp, ou du dhp et de la hauteur, ou du dhp, de la hauteur et d'un diamètre supérieur (fonctions de volume, de biomasse ou de carbone, respectivement). La figure ci-dessous fournit un exemple.

Bien sûr, d'autres modèles sont utilisés pour des objectifs spécifiques, par exemple dans les inventaires de souches, prédire le dhp à partir du diamètre de souche (et de la hauteur de souche) ou pour les arbres à contreforts, prédire le dhp à partir du diamètre au-dessus des racines en contreforts. La figure ci-dessous fournit un exemple.

Pour les fonctions de biomasse, les termes de **fonctions de biomasse allométriques** ou **modèles de biomasse allométriques** sont souvent utilisés. Le terme allométrique provient du grec ancien et du latin, où «ἄλλος» (allos) signifie autre et metric signifie mesure. Allométrique signifie donc que l'on déterminera la biomasse à partir d'autres variables. Suivant ce sens original, le terme est essentiellement redondant lorsqu'il spécifie un modèle, car il décrit simplement ce que tout modèle fait: produire la valeur d'une variable à partir de mesures de valeurs d'autres variables.

Les types de modèle énumérés jusqu'ici sont couramment générés par des études de recherche préalables à l'inventaire. Pour des inventaires forestiers spécifiques, des modèles de biomasse sont souvent tirés de la littérature après vérification de leur adéquation à l'inventaire spécifique (cela sera abordé plus loin dans cette leçon).

Néanmoins, il y a des modèles qui sont générés avec et à partir des données de l'inventaire lui-même: un exemple typique est une courbe de hauteurs utilisées pour prédire les hauteurs totales des arbres. Les mesures de hauteur prennent du temps sont donc chères, donc les hauteurs sont seulement mesurées pour un sous-ensemble (bien défini) d'arbres échantillons. Un modèle est alors construit à partir de ces mesures, qui est utilisé pour prédire les hauteurs des arbres non mesurés. Dans ce cas, les mesures de hauteur montreront une plus grande variabilité que les hauteurs prédites, car ces hauteurs prédites représentent des valeurs moyennes pour un classe de dhp donnée.

Exemple 4: Modèles calculés à partir de variables cibles et auxiliaires

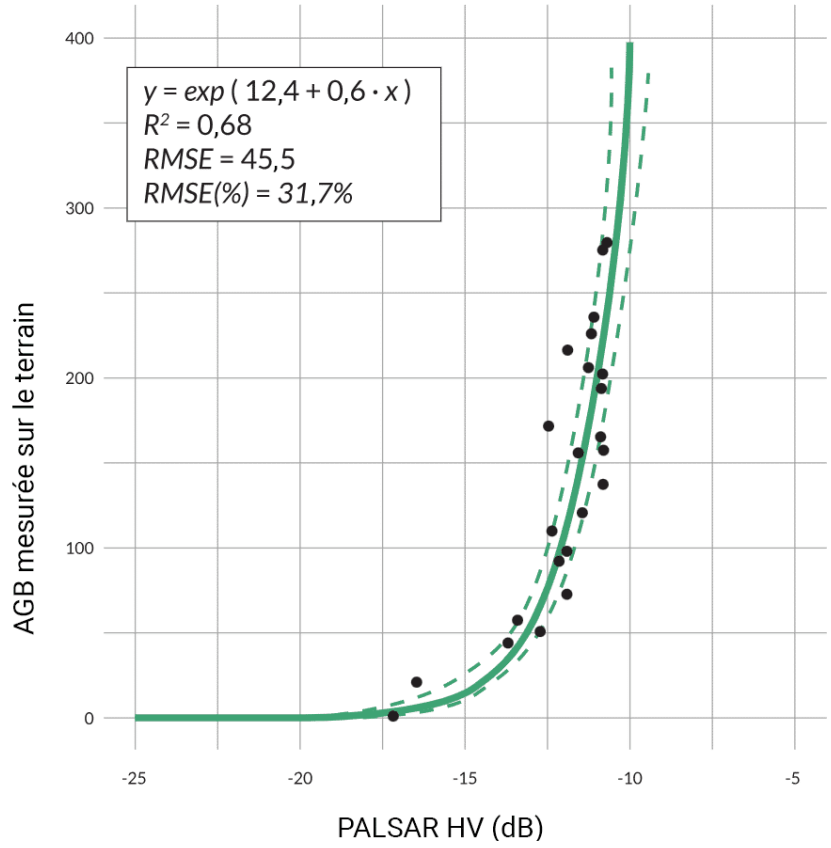
Un autre exemple de modèles construits pendant la mise en œuvre de l'inventaire est donné lorsque des variables auxiliaires sont observées pour leur utilisation avec des estimateurs par ratio et par régression, ou dans un échantillonnage double avec un estimateur par ratio ou par régression: alors, un modèle (un ratio simple ou une régression) est calculé à partir des parcelles d'échantillonnage où les deux variables, cible et auxiliaire, ont été enregistrées. Le modèle appuie l'estimation et permet, s'il existe une corrélation suffisante entre la variable cible et les variables auxiliaires, d'obtenir une estimation plus précise de la variable cible. Autrement dit, le modèle assiste le processus

d'estimation, et l'estimateur utilisé est donc aussi appelé un estimateur assisté par modèle.

Bien que le processus d'estimation soit ici appuyé par le modèle, le caractère non biaisé de l'estimation provient toujours de la randomisation de la sélection de l'échantillon, qui assure que l'échantillon est représentatif de la population. Cependant, remarquons que l'estimateur par ratio est uniquement non biaisé lorsque le modèle simple est valable. Cela signifie que le modèle doit saisir exactement la relation entre les variables cible et auxiliaire, et qu'il n'y a pas d'autres facteurs qui puissent affecter le processus d'estimation. Si le modèle simple n'est pas suffisant, l'estimateur par ratio pourra être biaisé et les résultats de l'estimation peuvent être inexacts.

L'estimation et l'inférence peuvent être entièrement fondées sur des modèles, et les approches correspondantes sont appelées estimation fondée sur un modèle ou inférence fondée sur un modèle. Dans ce cas, le caractère non biaisé de l'inférence est entièrement dépendant de la validité du modèle.

Un exemple typique ici est la modélisation de la relation entre la biomasse observée sur le terrain et les données de télédétection. Lorsqu'un tel modèle a été établi, il est possible de prédire la biomasse forestière pour chaque pixel. Avec cette prédiction, on est en position non seulement de produire une estimation de la biomasse pour l'ensemble de la région inventoriée (en additionnant les prédictions de biomasse par pixel), mais aussi de générer une carte de la biomasse.



Modélisation de la relation entre les mesures de biomasse sur le terrain dans des parcelles d'1 ha et le signal de rétrodiffusion par satellite du radar à antenne synthétique ALOS PALSAR (polarisation RAS - HV), avec un ajustement de régression exponentiel. Modifié à partir de Huang et al. (2015).

Note

À partir de la liste relativement longue de modèles utilisés dans le suivi des forêts, il apparaît clairement qu'ils jouent ici un rôle crucial: cela est principalement dû au fait que les forêts sont des objets complexes à suivre, et diverses variables d'intérêt ne peuvent pas être directement observées. Afin de rendre un système de suivi opérationnel, il est essentiel de travailler avec des valeurs prédites à partir de modèles. Il est néanmoins important de distinguer clairement les valeurs qui sont enregistrées par des observations/mesures immédiates de celles qui sont prédites à partir de

modèles.

Les deux points importants à rappeler ici sont les suivants:

- les observations immédiates ne comportent qu'une seule source d'erreur: l'erreur de mesure, tandis que les prédictions à partir de modèle comportent des erreurs de mesure (des variables explicatives nécessaires) et des erreurs de modèle; et
- les prédictions à partir de modèle présentent une variabilité inférieure aux observations immédiates: pour le même ensemble de valeurs d'une variable explicative, la même valeur prédite sera toujours générée, alors qu'en réalité, par exemple, différents arbres avec le même dhp peuvent avoir des biomasses assez différentes.

Principales caractéristiques des modèles statistiques

Voyons maintenant les principales caractéristiques des modèles statistiques.

Les modèles prédisent des valeurs moyennes, et non pas de vraies valeurs.

Il est important de comprendre qu'avec un ensemble de variables explicatives donné, les modèles ne produisent pas la vraie valeur d'une variable réponse. Les prédictions doivent être comprises comme des valeurs moyennes. En utilisant un modèle pour prédire, par exemple, la hauteur totale des arbres à partir des mesures du dhp, on assigne des hauteurs moyennes aux arbres échantillons: tous les arbres avec un dhp spécifique, disons 40 cm, auront la même hauteur prédite.

Cela ne correspond évidemment pas à la vraie situation où les arbres avec le même dhp varient en hauteur: la variance des hauteurs prédites des arbres échantillons est donc toujours inférieure à celle des hauteurs totales si elles avaient été mesurées.

Les modèles statistiques s'appuient sur des observations des échantillons, et les coefficients de modèles sont eux-mêmes des estimations

Tous les modèles sont des estimations en soi – il se fondent sur les observations d'un ensemble (échantillon) d'arbres et ne représentent pas le (vrai) modèle paramétrique, mais seulement une approximation (estimation) de celui-ci. Si différentes équipes de terrain prennent un échantillon aléatoire de 100 arbres – chacune dans la même région inventoriée mais avec différentes randomisations – pour calculer le modèle de biomasse en utilisant chacune le même modèle

mathématique, elles obtiendront toutes différents coefficients de modèle.
Comme avec l'échantillonnage courant pour les valeurs moyennes, la précision de l'estimation sera généralement meilleure lorsque la taille de l'échantillon est plus grande: pour chaque valeur individuelle réponse ou valeur moyenne, des intervalles de confiance peuvent être déterminés. Pour cela, les mesures de variabilité du modèle doivent bien sûr être connues.

Les modèles statistiques sont caractérisés par des mesures statistiques

Comme pour l'estimation fondée sur un échantillon des valeurs moyennes, il est aussi important pour les modèles estimés d'accompagner les estimations ponctuelles par des estimations par intervalle. Les estimations ponctuelles des modèles de régression sont les coefficients de régression estimés. Pour chaque coefficient de régression estimé, une erreur-type peut être estimée, et plus elle est petite, plus l'estimation sera précise.
La signification est une caractéristique importante de la régression: si, dans l'exemple d'une régression linéaire simple, le coefficient de pente n'est pas statistiquement significativement différent de zéro, on dira que la régression n'est pas significative. Cela signifie que la ligne de régression est parallèle à l'axe x et, par conséquent, que la valeur prédite pour la variable cible y sera la même pour toutes les variables explicatives de x. La valeur prédite est la valeur moyenne pour y. Alors au lieu de calculer une régression qui détermine une valeur moyenne spécifique par classe de valeurs de x, on peut utiliser la moyenne générale, comme valeur prédite pour toutes les classes de x.

Les modèles statistiques emploient des points de données variables

Une autre statistique importante qui caractérise la précision de la prédiction à partir du modèle de régression est la variabilité des points de données utilisés pour construire le modèle autour de la ligne de régression. Tout comme les valeurs d'une variable singulière varient autour de la valeur moyenne, les point de données dans un modèle de régression varient autour de la ligne de régression – où la ligne de régression représente une **valeur moyenne changeante** qui prend différentes valeurs pour différentes classes de x. On appelle cela la variabilité de la variance des résidus. Lorsque cette variance des résidus est faible, les point de données sont très rapprochés autours de la ligne de régression, et on peut supposer que les prédictions du modèle sont assez précises.
Rappelons qu'il est important de connaître non seulement les coefficients de régression afin de pouvoir calculer les valeurs prédites, mais aussi les statistiques de variance du modèle pour pouvoir évaluer sa qualité. La précision de l'estimation d'un modèle de régression linéaire simple, par exemple, est

supérieure autour de la valeur moyenne des variables explicatives et devient plus faible vers les limites de la plage de valeurs explicatives incluses; au-delà de cette plage, le modèle ne doit pas être utilisé, et si c'est le cas, la précision sera faible.

Les modèles statistiques sont valables pour des «conditions spécifiques»

Les données de base utilisées pour construire un modèle co-définissent la validité du modèle pour un objectif d'inventaire spécifique. Lorsque l'on parle de données de base, on se réfère aux facteurs moyens comme:

Région géographique	Dans le meilleur cas, les données de base proviennent de la région géographique où l'inventaire à lieu. Si ce n'est pas le cas, on doit s'assurer que le modèle est adéquat (voir la section sur l'identification de modèles statistiques adéquats dans cette leçon).
Espèces d'arbre	Certains modèles sont spécifiques à une espèce ou un groupe d'espèces; la possibilité de les appliquer à d'autres espèces doit être vérifiée.
Plage de variables d'entrée	Les modèles doivent généralement être utilisés uniquement pour des valeurs des variables d'entrée qui sont couvertes par les données de base. Il peut être risqué d'utiliser par exemple un modèle de biomasse qui a été construit pour des valeurs entre 30 et 150 cm pour des arbres en-deçà de 30 cm – une telle extrapolation peut produire des valeurs invraisemblables car la fonction du modèle n'est pas essentiellement définie en dehors de la plage de données de base.

Comment construire son propre modèle de biomasse

Construire son propre modèle de biomasse n'est pas une tâche habituelle dans un projet d'IFN. On peut recourir aux modèles qui ont déjà été publié auparavant. Parfois, ces modèles ont été construits dans le cadre de recherche universitaire ou de rapports techniques et dont difficiles à trouver. Mais il est probable que des modèles adéquats pour toutes les situations existent.

La FAO, avec le CIRAD, propose une base de données qui héberge des modèles de biomasse dans

l'***Initiative GlobAllomeTree*** (en anglais), qui peut servir de référence pour rechercher ou construire des modèles de biomasse. La consultation d'un manuel approfondi sur la mise au point d'équations allométriques est recommandée aux personnes intéressées (Picard et al. 2012).

Dans cette section, nous aborderons brièvement les étapes à suivre afin de construire votre propre modèle de biomasse. Ce processus est valable de même pour tout autre modèle statistique. Par exemple, si l'on souhaite estimer la biomasse des arbres exploités illégalement à partir d'un inventaire de souches, on appliquera des modèles de biomasse normaux, et pour cela on devra prédire le dhp à partir du diamètre de souche. On pourra vouloir construire ici son propre modèle.

Note

Il existe un grand nombre de modèles publiés pour différentes espèces d'arbre qui permettent de prédire le dhp à partir du diamètre de souche. Il s'agit soit de facteurs simples ou de modèles de régression, certains incluant aussi la hauteur de souche (par ex. Pond and Froese, 2014).

Remarquons que l'on ne trouve pas de définition du diamètre de souche ni de la hauteur de souche dans ces publications, pas plus que des indications pour savoir comment les mesurer. Néanmoins, le diamètre de souche peut être très irrégulier (car les arbres ont souvent des contreforts à de très faibles hauteurs) et des définitions claires et sans équivoque seraient nécessaires.
Cela illustre l'importance des définitions sans ambigüité, non seulement dans le suivi des forêts, mais aussi pour les variables d'entrée pour les modèles statistiques.

Étape 1: Commencez par des définitions

Suivant les bonnes pratiques de suivi des forêts qui recommandent d'avoir une terminologie, des définitions et des mesures claires et documentées avec transparence, il faut commencer par des définitions. Les définitions s'appliquent à la population dans laquelle les arbres échantillons font être sélectionnés, y compris:

- la définition géographique de l'aire de provenance des arbres échantillons;
- une définition des espèces ou groupes d'espèces; et
- une définition de la plage de dimensions (généralement une plage de dhp) pour laquelle le

modèle sera valable.

En outre, la biomasse doit être définie en termes de compartiments de biomasse (tige, grandes branches, petits rameaux, masse souterraine, feuilles, etc.) qui doivent être considérés ainsi que les diamètres minimum.

Étape 2: Déterminez le nombre d'arbres échantillons à abattre

Le nombre d'arbres échantillons à abattre doit être déterminé; cela est généralement fait selon les ressources disponibles. L'abattage et la pesée des arbres étant coûteux, le nombre d'arbres est souvent assez limité, même s'il est bon de travailler avec de grands nombres d'arbres échantillons pour que les prédictions du modèle soient plus précises.

Il y a une grande différence entre l'échantillonnage d'un grand arbre et d'un petit arbre, car l'abattage, la découpe et la pesée des grands arbres sera beaucoup plus chère, de manière disproportionnée. Cela entraîne une situation où les études de biomasse ont beaucoup de petits arbres et seuls quelques grands arbres.

C'est une sorte de dilemme car la variabilité dans la biomasse des petits arbres est plus faible que la variabilité des grands arbres, et la conclusion serait que l'on a besoin de grands arbres pour établir une meilleure fondation pour un modèle où la variabilité est supérieure (et où fréquemment la majeure portion de la biomasse du peuplement forestier réside dans les grands arbres).

Étape 3: Sélectionnez les arbres échantillons

Si la forme générale des modèles de biomasse est assez bien connue et suit des lois physiques, le but est d'essayer de déterminer la forme de la fonction de biomasse pour l'ensemble de la plage de valeurs d'entrée (en général la plage de dhp). Cela signifie que l'on doit essayer de sélectionner des arbres échantillons pour toutes les classes de diamètre – et (théoriquement) plus d'arbres échantillons là où la variabilité de la variable cible (biomasse) est supérieure.

Rappel à la réalité

La sélection des arbres échantillons est souvent dominées par des considérations pratiques comme l'accessibilité, les permis de coupe, entre autres. C'est un nouvel exemple dans le contexte du suivi des forêts où la théorie se confronte à la pratique ou, autrement dit, où la science rencontre la vie

réelle.

Étape 4: Mesurez les arbres échantillons sur pied

Les arbres échantillons doivent d'abord être mesurés sur pied, comme lorsque l'on mesure les variables explicatives dans un inventaire forestier par défaut. Par exemple, la hauteur totale peut difficilement être déterminée après l'abattage, puisque ce que l'on mesure alors sera la longueur de l'arbre. En outre, le dhp est mieux mesuré sur un arbre sur pied, car après l'abattage il sera plus difficile de déterminer la hauteur de poitrine.

Étape 5: Abattez les arbres échantillons

Une fois mesurés sur pied, les arbres échantillons doivent être soigneusement abattus, car il est nécessaire de s'assurer que toutes les parties importantes de la biomasse peuvent être attribuées à l'échantillon. Puis, les compartiments de biomasse pertinents doivent être séparés et pesés.

Rappel à la réalité

Lorsque l'on mesure et abat les arbres échantillons, il est courant que les erreurs de mesure ne soient pas considérées ou factorisées dans le modèle pour déterminer les mesures d'incertitude. Les mesures des arbres échantillons sur pied et abattus sont simplement prises comme des vraies valeurs. Nous savons néanmoins que cela n'est pas le cas et que les erreurs de mesure peuvent jouer un rôle, particulièrement lorsque le nombre d'arbres échantillons est faible. Dans les IFN, on a généralement de grandes tailles d'échantillon et un grand nombre d'arbres échantillons enregistrés; alors, on peut supposer que les erreurs de mesure aléatoires ont un poids relativement faible car le grand nombre d'observations permettra de conserver une variance d'erreur faible.

Étape 6: Gérez les données des arbres échantillons

Les données des arbres échantillons devront être gérées dans une base de données où tous les résultats par compartiment sont stockés et finalement additionnés aux valeurs cibles pour chaque arbre échantillon particulier. Enfin, pour constituer les «entrées» des analyses futures, une liste est produite avec les données par arbre individuel de la variable cible et des variables explicatives. Cette liste est la

matrice d'entrée pour l'estimation et les coefficients de modèle.

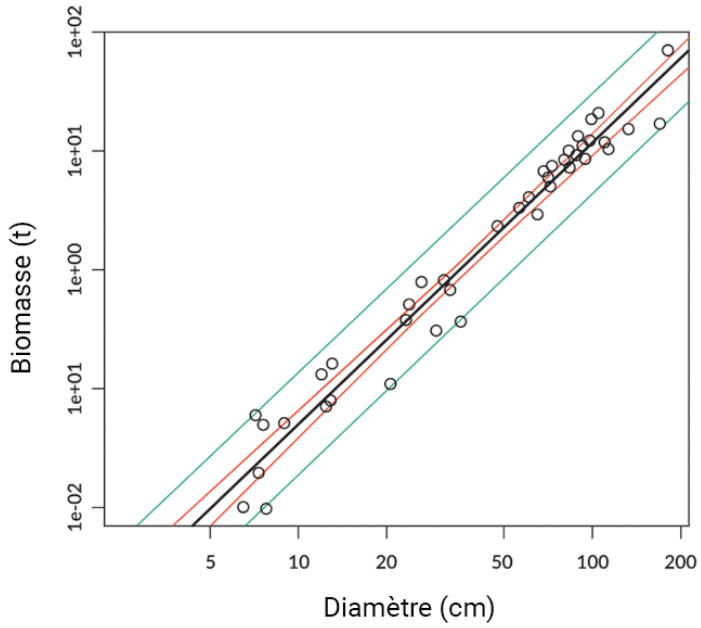
Étape 7: Identifiez un modèle mathématique adéquat pour l'ensemble de données

Les étapes suivantes font appel aux statistiques appliquées: un modèle mathématique doit être identifié qui sera capable de bien s'adapter à l'ensemble de données. Pour les modèles de biomasse, comme pour d'autres modèles, des modèles mathématiques typiques sont connus. Il est courant de comparer la performance de différentes modèles mathématiques et de choisir celui qui permettra les prédictions les plus précises.

Une approche habituelle consiste à créer un modèle en utilisant une sélection aléatoire de 75 pour cent des arbres échantillons, puis d'évaluer la performance du modèle en utilisant les 25 pour cent restants des arbres échantillons. Ici, on considère les 25 pour cent d'arbres de vérification comme des arbres sélectionnés indépendamment pour la validation du modèle. Il est important de faire la distinction entre:

- l'incertitude du modèle, qui provient de l'analyse des 75 pour cent des arbres échantillons utilisés pour construire le modèle; il va sans dire que le modèle est bien adapté à cet ensemble de données puisque l'ensemble de données est la base du modèle; et
- l'incertitude de la prédiction, qui est généralement plus importante et se réfère à la prédiction pour les arbres qui n'ont pas été utilisés dans la construction du modèle. Considérons ici que les arbres échantillons utilisés pour construire le modèle sont un échantillon de la population d'intérêt, et, bien entendu, notre échantillon n'est pas capable de saisir toute la variabilité présente dans la population et n'en est qu'une estimation, comme on le voit dans la figure ci-dessous.

Ainsi, tous les arbres enregistrés dans un inventaire appartiennent à cet ensemble d'arbres qui n'ont pas été utilisés pour construire le modèle. Par conséquent, l'incertitude de la prédiction est un point important dans la mise au point du modèle. Picard et al. (2012) apporte des éclairages sur ces questions.



Une fois satisfait de l'incertitude du modèle et de l'incertitude de la prédiction, on peut revenir à l'ensemble de données complet des arbres échantillons et estimer les coefficients de modèle finaux à partir de 100 pour cent des arbres échantillons..

Note

Un certain nombre de questions statistiques doivent être prises en compte dans la construction du modèle statistique de sorte qu'il est recommandé de consulter un statisticien modélisateur lorsque l'on construit et choisir un modèle spécifique.

Les questions statistiques incluent:

1. la corrélation avec les variables explicatives (et ce que l'on appelle la colinéarité); et
2. une caractéristique typique de la biomasse (et des modèles de volume et de carbone): la variabilité de la biomasse qui est faible pour les petits arbres et élevée pour les grands arbres (ce que l'on appelle l'hétéroscédasticité).

Cette dernière affecte l'estimation des variances et des intervalles de confiance pour les prédictions des valeurs moyennes et des valeurs individuelles – par exemple, pour les modèles de biomasse

uniquement fonction du dhp, les intervalles de confiance seront plus étroits pour les petits arbres et s'élargiront à mesure que le dhp augmente

Étape 8: Documentez le modèle

Une fois le modèle final décidé, un rapport complet et transparent du modèle est la dernière étape, où non seulement le modèle et ses coefficients doivent être documentés, mais aussi les caractéristiques d'incertitude, y compris le coefficient de détermination, l'erreur-type de la régression et l'incertitude de la prédiction, et d'autres incertitudes possibles.
On peut aussi souhaiter rendre l'ensemble de données original disponible pour le public, car ces ensembles de données peuvent être très utiles une fois combinés avec de nouveaux ensembles de données pour générer des modèles plus précis, plus adaptés localement ou plus généralisables.

Identification de modèles statistiques adéquats

De nombreux modèles statistiques sont disponibles pour le suivi des forêts. Le GIEC, par exemple, propose une longue liste de différents facteurs de conversion et fonctions de biomasse (voir, par ex., les Lignes directrices du ***GIEC 2006 Publications - IPCC-TFI*** (iges.or.jp) ou leurs valeurs actualisées en 2019 ***Publications - IPCC-TFI*** (iges.or.jp) (en anglais)).

Dans de nombreux cas, le modèle à utiliser est clair dès le début, car il a déjà été utilisé auparavant dans le même contexte géographique ou thématique, ou il est connu pour sa bonne performance dans les circonstances du projet d'inventaire spécifique. Dans les IFN qui portent sur des vastes surfaces, et incluent de nombreuses espèces et groupes d'espèces, l'application de différents modèles selon le groupe d'espèces, et/ou la région géographique et/ou les conditions de terrain peut être adéquate.

La première étape de la sélection de modèle est de vérifier quels modèles ont déjà été utilisés auparavant, la provenance des données des arbres échantillons utilisés pour la construction du modèle, et d'évaluer les statistiques d'incertitude des modèles; généralement, plus il y a eu d'arbres échantillons traités, plus le modèle sera exact. Parfois une décision doit être prise concernant, notamment, l'utilisation de divers modèles de biomasse spécifiques aux espèces, ou d'un modèle général pour toutes les espèces. Dans les IFN où les tailles d'échantillon sont couramment importantes, la recommandation de la recherche récente est de s'intéresser au nombre d'arbres échantillons utilisés dans la construction du modèle plutôt qu'aux espèces spécifiques pour lequel il a été élaboré.

Cela signifie qu'un modèle général pour toutes les espèces fondé sur un grand nombre d'arbres échantillons est généralement préféré à l'utilisation de plusieurs modèles spécifiques aux espèces construits à partir d'un faible nombre d'arbres chacun.

Astuces rapides!

Certaines règles générales ont été identifiées concernant la mise au point de modèles de biomasse (McRoberts and Westfall, 2014): le nombre d'arbres échantillons doit être d'au moins 100, et les modèles doivent avoir un coefficient de détermination supérieur à 0,95. Ainsi, dans les IFN, les erreurs de modèle pour la biomasse sont relativement faibles par rapport à l'erreur-type. Cependant, il est important de souligner que le rôle moindre des erreurs de modèle dans les IFN avec de grandes tailles d'échantillon concerne les erreurs aléatoires uniquement, car les erreurs systématiques et les biais vont bien sûr se propager et donner des estimations finales biaisées!

Dans les situations où le choix du modèle n'est pas évident dès le départ, la tâche consiste à vérifier l'adéquation et comparer la performance de différents modèles. Cela peut uniquement être réalisé avec un nombre d'arbres échantillons suffisamment grand, et ce sera coûteux, particulièrement pour les modèles de biomasse, car déterminer/mesurer la biomasse d'arbres échantillons est toujours cher.

La CCNUCC (2011) propose un guide de base pratique sur la manière de mener des tests d'adéquation des modèles pour la biomasse aérienne des forêts. Une description exhaustive et scientifique de cette analyse de l'adéquation des modèles en général peut être trouvée dans Pérez-Cruzado et al. (2015).

Questions de publication des résultats dans les modèles statistiques

Il existe essentiellement deux types de questions de publication des résultats lorsque l'on parle de modèles statistiques dans le suivi des forêts:

- *Publier le modèle en lui-même et permettre aux utilisateurs potentiels de comprendre pleinement la provenance et les caractéristiques du modèle*

Il est ici important de documenter tous les détails de la construction du modèle avec transparence et exhaustivité, comme cela a déjà été dit: d'où proviennent les données des arbres échantillons en termes

de région géographique, d'approche d'échantillonnage et de possibles restrictions, combien d'arbres échantillons ont été utilisés et quelle était leur distribution dans la plage de variables explicatives. Essentiellement, tous les détails de la construction du modèle qui sont nécessaires pour qu'un utilisateur potentiel puisse le comprendre et saisir sa provenance doivent être notifiés. Cela inclut aussi les mesures statistiques de l'exactitude du modèle, même si cela n'est pas une pratique courante par défaut dans les IFN où les erreurs de modèle sont publiées et propagées à l'erreur totale (voir aussi le point ci-dessous). Néanmoins, les mesures d'incertitude sont importantes lorsqu'un utilisateur potentiel compare différents modèles; il pourra alors tendre à préférer le modèle le plus exact.

- *Publier les prédictions d'un modèle dans le contexte de la mise en œuvre d'un IFN*

Ici, les prédictions du modèle sont traitées dans les projets d'inventaire forestier comme des observations normales, l'incertitude du modèle correspondante n'est pas souvent notifiée, car il a été déterminé par des études empiriques et théoriques que sa contribution à l'erreur finale est mineure. Cependant, documenter les modèles que l'on a utilisés et donner leur source et leurs caractéristiques dans le rapport d'inventaire est une bonne pratique.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Les modèles statistiques cherchent à établir une relation quantitative entre une variable réponse et une ou plusieurs variables explicatives.
- Les modèles statistiques ne servent pas à établir une relation de cause à effet – les modèles de processus, qui visent à inclure explicitement les causes inférées par des processus biologiques, abordent cette question.
- Il existe différents types de modèles – de complexité variable – utilisés dans le suivi des forêts, et dans certains cas, il est difficile de remarquer qu'un modèle a été utilisé.
- Dans de nombreux cas, le modèle à utiliser est clair dès le début, car il a déjà été utilisé auparavant dans le même contexte géographique ou thématique, ou il est connu pour sa bonne performance dans les circonstances du projet d'inventaire spécifique.
- Dans les IFN qui portent sur des vastes surfaces, et incluent de nombreuses espèces et groupes d'espèces, l'application de différents modèles selon le groupe d'espèces, et/ou la région géographique et/ou les conditions de terrain peut être adéquate.

Leçon 4: Erreurs dans le suivi des forêts

Introduction de la leçon

Cette leçon aborde les types et les rôles des erreurs aléatoires qui surviennent le long du processus d'IFN. Elle décrit aussi la propagation d'erreur – comment les sources d'erreur se propagent à l'erreur totale du résultat final.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Définir le terme «erreur» dans les études par échantillonnage empiriques.
2. Décrire pourquoi les considérations des erreurs sont importantes dans le suivi des forêts.
3. Expliquer la relation entre erreur et effort.
4. Comprendre les types et les rôles des erreurs dans le suivi des forêts.
5. Expliquer comment traiter les erreurs dans le suivi des forêts.

Observations générales sur les erreurs dans les inventaires forestiers

Définition du terme erreur dans les études empiriques

Les inventaires forestiers sont des études par échantillonnage empiriques, où il est plus exact de se référer aux erreurs comme la variabilité résiduelle et non pas comme des fautes. Alors que les fautes peuvent être évitées avec un travail rigoureux et axé en permanence sur la qualité, les erreurs sont inévitables – on ne peut qu'essayer de les réduire. Les erreurs auxquelles nous faisons référence ici sont de caractère aléatoire et tendent à suivre la loi des erreurs de Gauss, également appelée loi normale.

Comme on l'a mentionné, et contrairement aux erreurs aléatoires, les erreurs systématiques ou les biais peuvent généralement être évités, car ils sont fondés sur de mauvais calibrages, l'utilisation d'estimateurs biaisés, ou d'autres applications erronées des approches de génération de données. Puisque les erreurs aléatoires sont omniprésentes, on doit essayer de faire travailler les équipes de terrain et les autres parties qui génèrent des données de manière rigoureuse et cohérente, pour contribuer à réduire les sources d'erreur correspondantes. Généralement, dans l'échantillonnage statistique, on se réfère aux erreurs systématiques comme définissant l'exactitude, tandis que les erreurs aléatoires définissent la précision.

Le saviez-vous?

La précision et l'exactitude sont deux termes centraux dans l'échantillonnage statistique, et elles sont déterminées par les erreurs systématiques (biais) et aléatoires. Souvent, le terme d'incertitude est utilisé dans la notification des erreurs, car c'est un terme moins technique et plus accessible. Mais il est aussi moins clairement défini. Par conséquent, lorsque l'on utilise le terme incertitude dans le contexte de l'échantillonnage statistique, il est bon de préciser ce qu'il signifie spécifiquement.

Pourquoi les considérations des erreurs sont-elles importantes dans le suivi des forêts?

La présence et la magnitude des erreurs sont des facteurs contributifs importants de la crédibilité des résultats d'inventaire. Si l'erreur est de 50 pour cent, on aura moins confiance dans les résultats que pour une erreur de 1 pour cent. Il est par conséquent impératif de notifier les erreurs pour tous les résultats, de quantifier les erreurs qui peuvent l'être et de traiter les erreurs qui ne peuvent pas être directement quantifiées.

Lors de la planification du plan d'inventaire, l'objectif est toujours d'utiliser les ressources disponibles pour optimiser la précision de l'estimation et éviter les erreurs systématiques. Il est donc important de comprendre les rôles des différents éléments de conception de l'inventaire et comment les sources d'erreur correspondantes contribuent aux erreurs totales.

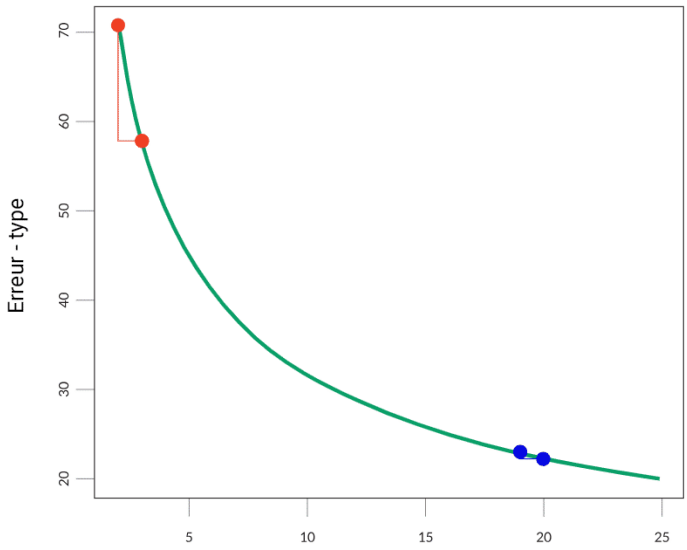
Si, par exemple, des ressources supplémentaires sont disponibles qui peuvent être utilisées pour améliorer le plan d'inventaire, on aura besoin d'identifier comment allouer ces ressources de sorte qu'elles contribuent au mieux à améliorer la précision = réduire l'incertitude = réduire l'erreur totale pour la ou les variables cibles centrales

La relation entre erreurs et efforts

L'erreur-type est souvent présentée comme une fonction de l'effort. L'effort est défini par la taille de l'échantillon qui peut être établie. L'augmentation marginale de la précision s'amenuise à mesure que la précise augmente. Cela signifie que, pour le même effort, l'on aura une hausse réduite de la précision si

l'on part d'un niveau élevé de précision. Ici, l'erreur se réfère à l'erreur-type et l'effort se réfère à la taille de l'échantillon qui peut être établie. Ainsi, pour optimiser la conception d'un inventaire forestier pour réduire l'erreur, il peut être pertinent de regarder en premier lieu les sources d'erreur relativement importantes – même si leur impact sur l'erreur finale n'apparaît pas comme le plus lourd à première vue.

Pour donner un exemple: dans la figure ci-dessous, on suppose un EAS et un écart-type estimé de s = 100. L'augmentation de la taille de l'échantillon de 1, passant de n = 2 à n = 3, réduit l'erreur-type estimée de TE = 70,7 à TE = 57,7, ce qui représente environ 18 pour cent. Si l'on investit les mêmes ressources supplémentaires en termes absolus et l'on augmente la taille de l'échantillon de 1 passant de n = 19 à n = 20, l'augmentation de la précision = la réduction de l'erreur-type sera de TE = 22,9 à TE = 22,4, ce qui représente une réduction relativement de seulement 2,2 pour cent.



La relation illustrée dans cette figure est valable pour l'erreur-type dans les plans d'EAS. Cependant, on peut supposer que des relations similaires sont valables pour d'autres sources d'erreur, par exemple, augmenter les efforts de formation d'un montant absolu fixe pour de meilleures observations des variables aura généralement plus d'effet pour les variables dont l'erreur de mesure est relativement importante (comme la mesure de hauteur totale) et aura peu d'effet sur des variables dont l'erreur de mesure est déjà assez faible (comme la mesure du dhp).

Qu'est-ce qu'une bonne précision dans les IFN – quel niveau d'erreur doit être ciblé?

Il n'existe pas de règle généralement valide pour la précision cible dans les IFN. L'approche courante est

que les ressources (budget) sont définies, et l'inventaire est conçu de manière à obtenir la meilleure précision possible dans la limite de ces ressources. La précision obtenue est une fonction du plan d'échantillonnage et, en particulier, de la taille de l'échantillon.
Dans certains IFN, la taille de l'échantillon présente un ordre de grandeur de 10 000, de sorte que la précision de l'estimation pour le pays entier est élevée – dans certains cas l'erreur-type relative est inférieure à 1 pour cent – mais dans d'autres IFN, la taille de l'échantillon peut avoir un ordre de grandeur deux rangs en-dessous, produisant des erreurs-types relatives supérieures à 10 pour cent. Néanmoins, lorsque les estimations sont produites pour des sous-populations plus petites, où la taille de l'échantillon est bien moindre, l'erreur-type peut croître et atteindre des niveaux pour lesquels la fiabilité des résultats est compromise.

Types d'erreur dans le suivi des forêts et leur rôle

Les systèmes de suivi des forêts sont complexes, et de nombreuses personnes sont impliquées dans leur mise en œuvre. Cela a le potentiel d'augmenter les diverses sources d'erreur qui doivent être observées dans la planification et la mise en œuvre de ces systèmes. Il y a trois types d'erreur qui arrivent dans le suivi des forêts et jouent un rôle important, mais ont une importance variable selon la conception de l'inventaire:

① erreur de mesure;

② erreur de modèle; et

③ erreur d'échantillonnage.

Dans cette section, nous verrons plus en profondeur ce que chacune d'entre elles signifie.

Erreur de mesure

Lorsqu'une observation est réalisée, cette observation est sujette à une variabilité résiduelle. Lorsque, par exemple, le dhp est mesuré avec un pied à coulisse très précis, disons au 5e de décimale d'un millimètre, les mesures répétées – réalisées avec diligence et correction – donneraient pratiquement toutes des mesures différentes. La variabilité de ces mesures indique l'existence d'erreurs de mesure.

Ces erreurs de mesure peuvent arriver au-delà des variables quantitatives. Les variables catégorielles qui sont observées dans, disons, 10 classes différentes, peuvent présenter de mauvaises affectations: cette erreur de mesure sera alors aussi appelée erreur de classification. Ou, si l'on observe des variables

nominales, comme les espèces d'arbre, des confusions/mauvaises identifications, également considérées comme des erreurs de mesure, peuvent arriver. Il est important de réaliser que ces erreurs de mesure arriveront dans l'observation de toute variable.

Un cas intéressant est la mesure de la hauteur totale avec le principe trigonométrique. En fait, la hauteur n'est pas mesurée mais calculée à partir de trois mesures: la distance horizontale de l'arbre, et les mesures d'angle avec le faîte de l'arbre et le pied de la tige. Selon le dispositif de mesure, la distance horizontale peut aussi se fonder sur deux mesures: la distance de pente et l'angle de pente. Toutes ces mesures individuelles sont porteuses de leurs erreurs de mesure spécifiques et l'erreur dans la hauteur provient finalement de la propagation des erreurs dans toutes ces mesures individuelles.

Astuces rapides!

Dans le suivi des forêts, il y a généralement peu d'information disponible sur les erreurs de mesure, et les mesures sont utilisées dans les calculs comme si elles étaient exemptes d'erreur. Néanmoins, pour les grands nombres d'arbres échantillons, comme c'est souvent le cas dans les IFN, on peut justifier d'ignorer les erreurs de mesure aléatoires et de ne pas les notifier explicitement car elles seront très faibles en comparaison avec l'erreur d'échantillonnage.

Mais si vous êtes intéressé, de petites études de recherche peuvent être établies où différentes équipes de terrain réalisent toutes les observations sur un lot de parcelles d'échantillonnage. La variance des mesures pour les différentes variables peut alors être considérée comme une approximation des erreurs de mesure/d'observation; cela peut comprendre les mesures de dhp et de hauteur, l'identification des espèces d'arbre et le nombre d'arbres échantillons trouvés par parcelle.

Erreur de modèle

Les modèles sont très fréquemment utilisés dans le suivi des forêts pour établir des relations qui permettent de prédire des variables qui ne peuvent pas être mesurées, ou très laborieusement (voir aussi la leçon 3 de ce cours). La lecture à partir d'un modèle n'est bien entendu pas la vraie valeur de l'objet d'intérêt. Lorsque, par exemple, le volume de la tige est lu à partir d'un modèle de volume comme une fonction du dhp, on considère que le volume de la tige de l'arbre est identique au volume de tous les arbres dans cette classe de dhp – alors que le vrai volume d'un arbre particulier s'en

écartera. C'est ce que l'on appelle l'erreur de modèle.

Dans le suivi des forêts, les erreurs de modèle ne sont pas souvent notifiées ni factorisées dans les estimations par intervalle. Les prédictions qui sont réalisées à partir des modèles sont prises comme des vraies valeurs, ou exemptes d'erreur. Dans les modèles pour les variables des arbres, comme les modèles de biomasse, de volume ou de hauteur, tant que les modèles sont utilisés sur un relativement grand nombre d'arbres échantillons, et tant que le nombres d'arbres échantillons dans l'inventaire est important, il peut être justifié de ne pas notifier les erreurs de modèle, car elles seront très faibles en comparaison avec l'erreur d'échantillonnage.

Erreur d'échantillonnage

L'erreur d'échantillonnage provient du fait que l'on n'observe pas tous les éléments d'une population mais uniquement un échantillon. Par conséquent, tous les résultats sont des estimations et s'écartent de la vraie valeur d'une manière qui est décrite par l'erreur-type.

Dans le suivi des forêts, l'erreur-type est normalement estimée de manière non biaisée pour la plupart des plans d'échantillonnage, excepté pour l'échantillonnage systématique, où l'on doit recourir à des approximations ou des estimations conservatrices de la variance d'erreur/l'erreur-type dans l'application d'un cadre d'EAS.

Les rôles de ces erreurs

Les trois types d'erreur décrits ci-dessus existent dans toutes les études de suivi des forêts. Beaucoup de recherche a été menée au cours de la dernière décennie, spécialement dans le contexte de l'estimation de la biomasse forestière, pour trouver quelles sont les contributions relatives de ces erreurs à l'estimation finale de biomasse. Nous nous concentrons ici sur le suivi national des forêts, une grande taille d'échantillon étant parmi les principales caractéristiques des IFN, et donc un grand nombre d'arbres échantillons.

Un résultat très pertinent a déjà été publié dans l'un des premiers articles sur la propagation des sources d'erreur à l'erreur finale: Gertner et Köhl (1992) ont introduit le terme de **bilan d'erreurs** et trouvé que, dans l'IFN de a Suisse, le poids le plus élevé était de loin porté par l'erreur d'échantillonnage avec environ 98 pour cent de l'erreur totale. Les erreurs de modèle et les erreurs de mesure n'étaient responsables que du petit pourcentage restant. Cela est une découverte importante car c'est généralement uniquement l'erreur-type induite par l'échantillonnage qui est notifiée dans les inventaires forestiers, tandis que les autres sources ne sont pas souvent quantifiées ni notifiées.

Ce résultat est valable pour les IFN avec de grandes tailles d'échantillon. Dans les études d'inventaire plus petites, avec de petites tailles d'échantillon, le poids relatif des erreurs de mesure et des erreurs de modèle peut être considérablement plus élevé.

Comment traiter les erreurs dans le suivi des forêts

Si les erreurs systématiques peuvent être évitées grâce à une planification et une mise en œuvre rigoureuses, les erreurs aléatoires arrivent toujours. Ainsi, le but est de réduire ces erreurs aléatoires (= la variabilité résiduelle).

Produire des estimations qui portent des erreurs raisonnablement faibles est l'un des principaux objectifs du suivi des forêts. On utilise «raisonnablement faible» car toute hausse de la précision coûtera des ressources et sera particulièrement coûteuse lorsque la précision est déjà élevée. Dans les IFN, l'erreur d'échantillonnage est le composant de la variance d'erreur le plus important et choisir un plan d'échantillonnage adéquat et une taille d'échantillon appropriée sont les leviers disponibles pour ajuster l'**erreur-type** (soit l'**erreur d'échantillonnage**).

Un travail, une formation et un contrôle périodique rigoureux peuvent contribuer à réduire les **erreurs de mesure**. Le plus important est d'éviter ici les erreurs systématiques causées par exemple par un mauvais calibrage, et de maintenir les équipes de terrain motivées afin qu'elles conservent l'ambition de produire de bonnes données; les longues périodes de terrain sont fatigantes et peuvent facilement mener à une détérioration de la motivation, et, par conséquent, de la qualité des données.

Concernant les **erreurs de modèle** néanmoins, tout dépend de l'étroitesse de la relation statistique entre la variable cible et les variables explicatives et du choix du modèle et des caractéristiques qualitatives de celui-ci, qui sont co-déterminées par le nombre d'observations sur lequel s'appuie le modèle. En général, plus le modèle sera construit sur un grand nombre de données, plus il pourra être considéré fiable.

Les projets d'inventaire forestier et les programmes de suivi des forêts sont complexes. Les trois types d'erreur abordés ici peuvent arriver à chaque étape de ces systèmes, et l'intérêt principal porte finalement sur l'erreur totale des variables cibles. Il apparaît clairement que toutes les erreurs qui sont entrées à différentes étapes du processus auront un impact sur l'erreur totale de la variable cible, car leur impact se propagera à travers les différentes étapes du processus d'inventaire jusqu'à la variable cible finale.

Dans ce contexte de propagation d'erreur, il faut tenir compte de deux points:

1. **Comment le mécanisme de propagation d'erreur fonctionne, et comment l'erreur totale est déterminée.** Cela est important pour notifier l'erreur totale comme une indication de précision et de fiabilité générale des résultats.
2. **Dans quelle mesure les différentes erreurs contribuent à l'erreur totale.** Cela est important dans le contexte d'optimisation de la conception de l'inventaire pour les inventaires de suivi: on cherchera à réduire les erreurs (à un coût acceptable) qui auront l'impact le plus important sur l'erreur des variables cibles. Basic principles of error propagation.

Principes de base de la propagation d'erreur

Les principes de base de la propagation d'erreur sont simples et dépendent de la relation entre les différentes variables d'entrée et leurs erreurs.

Taylor (1997) propose une bonne introduction compréhensible à la propagation d'erreur, qui couvre les règles de combinaison des variables aléatoires selon l'opération utilisée pour les combiner. Les lectures introductives plus courtes, qui incluent les sommes, les produits, les rations et d'autres opérations, peuvent être trouvées dans les références suivantes (en anglais):

Guide de la propagation d'incertitude et l'analyse d'erreur: Laboratoire de physique introductif de Stony Brook

Propagation d'incertitude dans les opérations mathématiques

Dans Taylor (1997), une approche analytique de la propagation d'erreur est élaborée, qui peut être directement appliquée aux fonctions des variables aléatoires. Cependant, si une propagation d'erreur doit avoir lieu dans un plan d'inventaire complexe, où de nombreuses sources d'erreur différentes doivent être prises en compte, une telle propagation d'erreur analytique devient extrêmement difficile. Alors, une étude par simulation (également appelée simulation de Monte-Carlo), peut être plus appropriée.

Pour conduire cette étude, on doit avoir de l'information disponible sur les différents composants d'erreur. Généralement, des distributions normales de ces erreurs sont supposées. La variable cible (par ex. biomasse) est alors calculée à partir de toutes les

données d'entrée, où un écart aléatoire par rapport à l'erreur normalement distribuée est déterminé pour chaque estimation ponctuelle. Cela est répété de nombreuses fois – disons, 10 000 fois – et la variance des valeurs finales de la variable cible est alors la variable cible totale propagée. Des exemples instructifs peuvent être consultés dans Molto *et al.* (2013), McRoberts et Westfall (2016).

Les calculs de la propagation d'erreur par simulation et analytiques permettent une évaluation du poids que les différents composants d'erreur ont sur l'erreur de l'estimation finale, de sorte que ces exercices de propagation d'erreur sont très instructifs pour optimiser la conception de l'inventaire.

McRoberts et Westfall (2016) proposent un exemple intéressant de la manière dont une étude par simulation peut être menée lorsque l'intérêt porte sur la propagation de diverses sources d'erreur à la variable cible finale dans le suivi des forêts. Si une estimation de la biomasse est la variable cible, les deux auteurs ont intégré les sources d'erreur suivantes dans leur simulation:

- ***Si une estimation de la biomasse est la variable cible:***

Les deux auteurs ont intégré les sources d'erreur suivantes dans leur simulation: variabilité des estimations des paramètres du modèle, (β), dans le modèle allométrique; variabilité des mesures de dhp; variabilité des mesures de hauteur (et autres variables d'entrée du modèle); variabilité résiduelle (ce qui est prédit par le modèle n'est pas la vraie biomasse); agrégation de la biomasse des arbres individuels à la biomasse parcellaire; et estimation de la biomasse totale pour l'aire étudiée à partir d'un échantillon de n parcelles (avec un plan d'échantillonnage défini).

- ***La simulation est alors menée comme suit:***

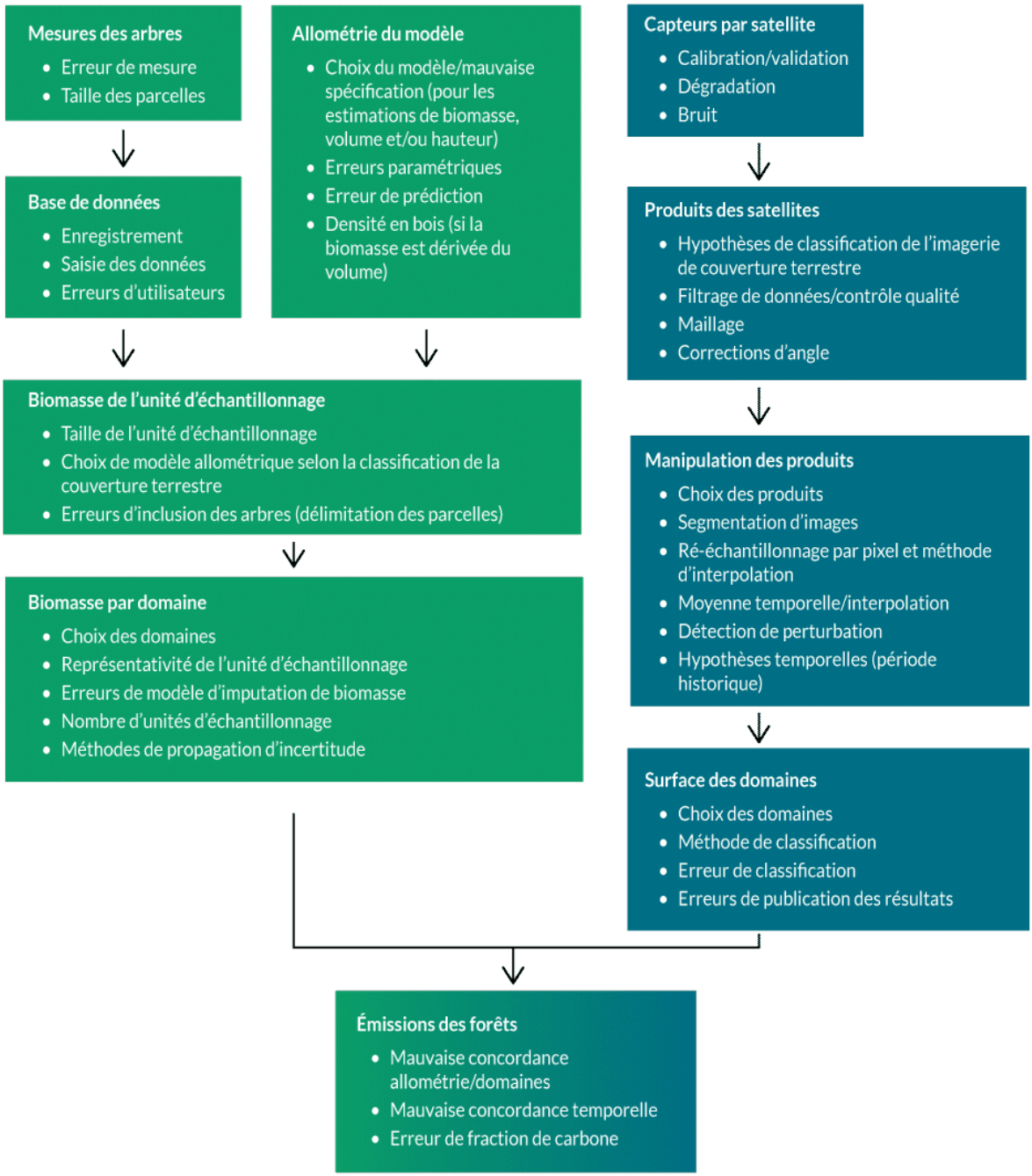
Pour chaque simulation, un ensemble de valeurs d'entrée est déterminé à partir de variables ci-dessus, où les composants d'erreur choisis aléatoirement à partir des erreurs distribuées normalement sont ajoutés à l'estimation ponctuelle. La biomasse totale est alors déterminée pour cette configuration particulière. La simulation est répétée de nombreuses (*m*) fois, chaque fois avec des valeurs d'entrée qui sont déterminées à partir des estimations ponctuelles plus un composant d'erreur aléatoire. La variabilité des *m* valeurs de la biomasse totale est alors une approximation empirique de l'erreur totale

propagée.

Commentaires de clôture sur la propagation d'erreur

Une liste complète des sources d'erreur potentielles apparaissant typiquement dans la chaîne de traitement du calcul des facteurs d'émission, des données d'activité et du carbone total est notifiée par les pays. La liste contient des concepts possibles qui sont trop fastidieux et hors de propos pour ces cours.

Veuillez remarquer que les parcours en vert sont définis exclusivement par les données d'inventaire. Les parcours en bleu sont suivis par le traitement de données obtenues par satellite. Les encadrés en bleu-vert proviennent de la combinaison des facteurs d'émission à partir d'inventaires et des estimations des données d'activité à partir de satellites.



Flux d'erreur dans les facteurs d'émission à partir d'inventaires (en vert) et les données d'activité à partir de satellites (en bleu)

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Les inventaires forestiers dont des études par échantillonnage empiriques, et lorsque l'on parle d'erreurs dans les études par échantillonnage empiriques, on se réfère à la variabilité résiduelle et non pas aux fautes.
- La présence et la magnitude des erreurs sont des facteurs contributifs importants de la crédibilité des résultats d'inventaire.
- Il n'existe pas de règle généralement valide pour la précision cible dans les IFN – la précision est décidée par l'allocation de ressources (budget) la plus efficace et la conception de l'inventaire.
- Il y a trois types d'erreur qui arrivent dans le suivi des forêts et jouent un rôle important, mais ont une importance variable selon la conception de l'inventaire: l'erreur de mesure, l'erreur de modèle et l'erreur d'échantillonnage.

Leçon 5: Produits typiques des analyses de données

Introduction de la leçon

L'objectif principal des analyses de données dans les IFN est de transformer les données des IFN en information significative pour les parties prenantes et les parties intéressées.

Cette leçon traite des principaux produits générés par les IFN.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire les produits potentiels des analyses de données d'IFN.
2. Identifier le rôle des analyses de données pour la publication des résultats des inventaires forestiers.

Produits des analyses de données du suivi des forêts: observations générales

Les analyses de données génèrent des résultats d'IFN qui répondent finalement aux attentes établies dans l'évaluation des besoins en information. Dans cette section, nous allons brièvement traiter les types généraux de produit dont la production est attendue à partir des données d'IFN.

Avoir une idée claire des produits potentiels des analyses de données d'IFN durant la phase de planification et l'EBI est toujours préférable. Il est aussi instructif – dans la phase d'EBI – de présenter tous les produits potentiels des analyses de données d'IFN et de les réduire pour obtenir des attentes réalistes.

Astuces rapides!

Il n'est pas nécessaire de tenir compte de toutes les implications analytiques pendant la phase d'EBI – c'est le problème des analystes de données. Néanmoins, il est certainement utile d'avoir des experts expérimentés dans l'analyse de données d'IFN présents dans l'évaluation des besoins en information afin d'éviter les attentes complètement irréalistes.

S'attendre à ce que les données d'IFN puissent être directement utilisées à des fins de planification

forestière aux niveaux des districts ou même des peuplements en est un exemple typique; ici, un expert expérimenté des inventaires devra clarifier les possibilités et les limitations des ensembles de données d'IFN.

Types de produit

Considérant le vaste éventail de résultats et de produits qui peuvent être générés, les IFN produisent un des ensembles de données exhaustifs qui incluent diverses options d'analyse. En général, présenter les résultats en utilisant deux stratégies distinctes est une bonne pratique:

1. l'une pour les parties prenantes et les décideurs (qui doit être technico-scientifique et répondre aux besoins exprimés dans l'EBI); et
2. l'une pour le public général (qui doit résumer les principaux résultats dans un langage accessible mais précis).

Concernant les produits technico-scientifiques, on peut diviser l'information en quatre catégories:

1) statistiques normalisées;
2) cartes;
3) optimisation de la conception de l'IFN; et
4) utilisation dans le secteur universitaire.

Statistiques normalisées

Les résultats des analyses de données d'IFN ne peuvent pas englober une liste complète des résultats normalisés. Il faut donc se limiter à ceux généralement produits dans les IFN, ainsi qu'aux produits spécifiques supplémentaires provenant des besoins en information particuliers d'un pays spécifique.

Par exemple, dans un pays avec un couvert forestier réduit, il peut être extrêmement pertinent de produire des résultats sur les arbres hors forêt (AHF) – tandis que dans les pays avec un important couvert forestier, cette ressource peut avoir une importance moindre (bien sûr, analyser des données des AHF n'est possible que si l'évaluation des AHF a été intégrée à la conception de l'inventaire).

Les unités de référence de base pour les analyses sont typiquement le pays entier et les unités infranationales – provinces, états, ou éco-zones définies. Dans la plupart des cas, la taille de l'échantillon des IFN ne permet pas d'aller plus loin et de produire des estimations pour des unités plus petites, sauf

si des techniques d'analyse scientifique spéciales, comme les estimations des petites surfaces, sont utilisées comme approche avancée.

Le saviez-vous?

Qu'est-ce que l'estimation des petites surfaces?

Les IFN utilisent généralement des échantillons systématiques avec une taille de grille en plage de kilomètres. C'est pourquoi des estimations raisonnables ne peuvent pas être produites pour des unités relativement petites (par ex. quelques km^2) car la taille de l'échantillon y sera trop petite.

Néanmoins, la recherche en cours s'intéresse à la manière de saisir l'information de grandes surfaces produite par l'échantillonnage des IFN et l'utiliser pour produire aussi des résultats pour des unités géographiques plus petites. Cela peut être réalisé en reliant les observations de terrain à des données de télédétection de couverture intégrale, qui sont alors utilisées comme données auxiliaires pour établir des modèles qui permettent de prédire les variables cibles pour chaque pixel, autrement dit pour toute l'aire inventoriée. Les données des variables cibles sont alors disponibles non seulement pour les points observés sur le terrain, mais aussi pour toute position (pixel) dans la région inventoriée. Cette approche, qui produit des résultats à partir d'échantillon brut de terrain de grande surface pour toute unité de petite surface dans la région inventoriée, est appelée **estimation des petites surfaces**.

Bien entendu, l'incertitude de la prédiction pour les petites surfaces dépend exclusivement de la qualité du modèle qui est dérivé et utilisé, en reliant les observations de terrain à grande échelle et les données de télédétection; et cela dépend, entre autres facteurs, de la résolutions spatiale et spectrale des données de télédétection, du plan parcellaire de l'inventaire de terrain, et de la qualité du co-enregistrement des données de terrain et de télédétection. .

Les **estimations des surfaces** incluent: la surface forestière, les surfaces ou pourcentage de surface pour des espèces d'arbre particulières, les surfaces par type de gestion, par statut de protection, par état de dégradation, par propriété, ou par caractéristiques topographiques. Les résultats peuvent alors être produits pour chacune de ces unités devant être publiées, par exemple la surface forestière par unité

infranationale, les surfaces des types de forêt par pays et par unité infranationale, la surface forestière à différentes altitudes, etc. La présentation se fait couramment sous forme de tableau à double entrée, avec, par exemple, la surface de 5 types de forêt dans 10 unités infranationales.

		Spécification des forêts				
État	**Mesure**	**Région boisée exploitable**	**Temporairement déboisée**	**Zone boisée**	**Terre forestière déboisée**	**Forêt**
Baden-Württemberg	[ha] Prim ET [%]	1330625 4600 1,2	1301 13 27,7	1331926 4601 1,2	39922 372 5,2	1371847 4620 1,2
Bayern	[ha] Prim ET [%]	2534232 2795 1,6	3796 11 33,7	2538028 2797 1,6	67535 194 7,7	2605563 2815 1,6
Brandenburg + Berlin	[ha] Prim ET [%]	1096101 907 2,7	2369 6 40,8	1098470 907 2,7	32378 67 12,8	1130847 909 2,7
Baden-Württemberg	[ha] Prim ET [%]	1330625 4600 1,2	1301 13 27,7	1331926 4601 1,2	39922 372 5,2	1371847 4620 1,2
Hessen	[ha] Prim ET [%]	845792 706 2,9	7598 19 22,8	853390 706 2,9	40790 91 10,8	894180 715 2,9
Mecklenburg-Vorpommern	[ha] Prim ET [%]	538651 2038 2,1	2186 19 24,0	540836 2041 2,1	17286 148 8,8	558123 2055 2,0
Niedersachsen	[ha] Prim ET [%]	1158459 1552 2,4	2985 12 30,5	1161444 1555 2,4	43147 135 9,2	1204591 1571 2,4
Nordrhein-Westfalen	[ha] Prim ET [%]	880082 861 3,1	3997 10 31,6	884059 863 3,1	25452 59 13,3	909511 867 3,1
Rheinland-Pfalz	[ha] Prim ET [%]	812818 2828 1,5	2290 22 21,7	815108 2831 1,5	24688 236 6,5	839796 2848 1,4
Saarland	[ha] Prim ET [%]	101459 100 8,0	783 2 70,5	102242 100 8,0	392 1 100,0	102 634 100 8.0

Sachsen	[ha] Prim ET [%]	51758 943 2,9	2392 12 28,8	520249 946 2,9	12956 56 14,5	533206 951 2,9
Sachsen-Anhalt	[ha] Prim ET [%]	493920 1829 2,2	9067 79 12,1	502987 1845 2,1	29494 264 6,3	532481 1884 2,1
Schleswig-Holstein	[ha] Prim ET [%]	168426 775 3,8	199 2 70,7	168626 776 3,8	4787 45 15,2	173412 778 3,7
Thüringen	[ha] Prim ET [%]	520944 895 2,7	2799 14 26,6	523743 902 2,7	25345 118 9,4	549088 912 2,6
Hamburg + Bremen	[ha] Prim ET [%]	13054 15 25,6	- - -	13054 15 25,6	791 2 70,4	13846 15 25,8
Allemagne (tous les États)	[ha] Prim ET [%]	11012420 20844 0,7	41742 221 8,0	11054162 20885 0,7	364962 1778 2,9	11419124 21040 0,7

Exemple de tableau à double entrée présentant la surface forestière par état fédéral («Land») en Allemagne divisée en différentes catégories de terres forestières. Ce tableau a été produit à l'aide de ***l'outil en ligne de l'IFN allemand*** *(en anglais); non seulement la surface estimée est donnée, mais aussi l'erreur-type relative estimée ET% et le nombre de clusters = unités d'échantillonnage primaires (qui correspondent à la taille de l'échantillon par unité infranationale) qui correspondent à la combinaison entre l'état fédéral et le type de terre forestière. Il est clairement visible ici que la précision de l'estimation est une fonction de la taille de l'échantillon.*

Rappelons que lorsque l'on divise des surfaces, tous les critères de division doivent être clairement définis de sorte que les résultats puissent être correctement interprétés en fonction de ces définitions: on doit définir clairement la «forêt» en opposition à la la non-forêt, et établir des critères précis pour distinguer dans la forêt les différentes classes de «dégradation», «type de forêt», «type de gestion», etc.

En outre, il est important de prendre en compte que toutes les catégories ne peuvent pas être identifiée sur le terrain ou grâce à l'imagerie de télédétection. Dans certains cas, ces catégories devront être tirées

de documents officiels. Par exemple, la propriété et le statut de protection doivent être extraits de cartes cadastrales et de cartes d'aires protégées, respectivement.

Les **estimations des caractéristiques par surface**, incluent: le volume/ la biomasse/ les stocks de carbone par hectare, le nombre d'arbres par hectare, le nombre de grands arbres, la densité de régénération, les stocks de bois mort dans différentes classes dimensionnelles, etc.

District	Biomasse (million de tonnes)	MdE (%)	Carbone (million de tonnes)	MdE (%)
Bumthang	80 ± 16	20	37 ± 7	20
Chhukha	91 ± 21	23	43 ± 10	23
Dagana	50 ± 8	15	24 ± 4	15
Gasa	7±2	31	3 ± 1	31
Haa	57 ± 10	18	27 ± 5	18
Lhuntse	77 ± 19	24	36 ±9	24
Mongar	95 ± 18	19	45 ±9	19
Paro	30 ± 8	27	14 ±4	28
Pemagatshel	18 ± 4	21	8±2	21
Punakha	54 ± 9	36	25 ± 9	36
Samdrup Jongkhar	72 ± 14	20	34 ± 7	20
Samtse	20 ± 5	22	10 ± 2	22
Sarpang	37 ± 6	17	18 ±3	17
Thimphu	48 ± 25	51	23 ± 12	52
Trashigang	96 ± 16	16	45 ± 7	16
Trashiyangtse	41 ± 20	48	19 ± 9	48
Trongsa	52 ± 11	21	25 ±5	21
Tsirang	29 ± 11	39	14 ± 5	39
Wangduephodrang	91 ± 16	18	43 ± 8	18
Zhemgang	56 ± 7	13	26 ±4	13

De nouveau, on peut illustrer cela avec un exemple de table simple donnant la masse totale en biomasse et en carbone pour le Bhoutan (DFPS. 2019). L'erreur-type relative (calculée à partir de

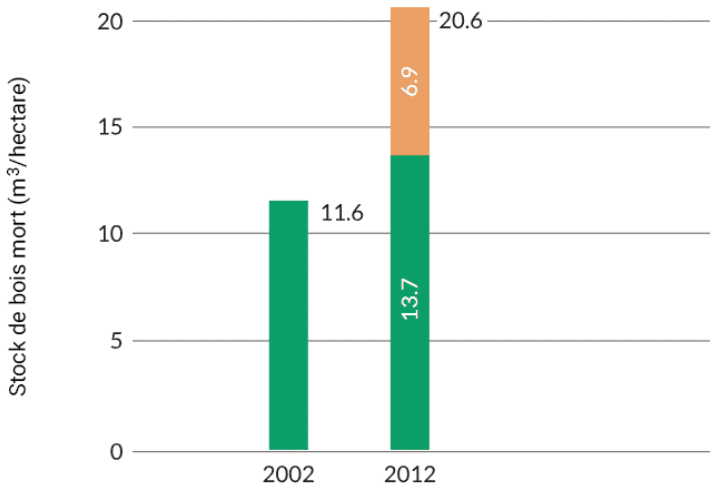
l'intervalle de confiance) est aussi donnée pour chaque district.

Les estimations des changements des variables cibles, lorsque les analyses se réfèrent à des inventaires répétés. Lorsque l'on analyse ces résultats, il est important d'observer si les définitions ont changé. Il peut arriver que l'analyse montre un changement, mais que celui-ci puisse en partie être attribué à des modifications des définitions.

On donne ici un exemple des changements des estimations de stocks de bois mort dans l'IFN allemand entre 2002 et 2012.. L'analyse a montré un changement très important des estimations de stocks de bois mort. Une partie de ce changement de magnitude inattendue des estimations était due à l'adaptation du diamètre minimum des pièces de bois mort enregistrées, passant de 20 cm à la norme internationale du GIEC de 10 cm.

Dans ce cas, on peut facilement analyser quelles portion du changement peut être attribuée à la modification de la définition, car toute l'information requise se trouve dans les données (pour appliquer l'ancienne définition, il suffisait d'écarter les pièces de bois mort avec un diamètre inférieur à 20 cm). Cela serait plus difficile avec l'analyse d'autres modifications de définitions, comme le changement du couvert arboré minimum dans la définition de la forêt.

Il est important que l'analyse établisse clairement quels sont les composants des changements: dans ce cas, avec l'ancienne définition, le changement serait une augmentation de 2,1 m³/ha passant de 11,6 m³ à 13,7 m³, mais le graphique représente un changement quatre fois plus important (de 9 m³/ha, passant de 11,6 m³ à 20,6 m³) du fait de la modification vers une définition plus inclusive (BMEL 2014)..



Si l'analyse produit aussi des **estimations des arbres hors forêt**, cela devra être indiqué sous forme de différents types d'utilisation des terres non forestières, présentant essentiellement les résultats par surface, mais pour les types d'utilisation des terres non forestières.

Il est important de répéter encore une fois que, pour toutes les estimations ponctuelles, l'analyse doit aussi produire des estimations par intervalle (erreur-type ou intervalles de confiance) afin que l'analyse montre le statut d'estimation ainsi que l'incertitude de cette estimation.

Les intervalles de confiance sont bien sûr également importants pour les estimations de changement. Si la valeur zéro est contenue dans les intervalles de confiance supérieur et inférieur, on peut supposer que les changements ne sont pas significatifs. Ici, bien entendu, lorsque l'on dérive des états de signification statistique, l'interprétation des intervalles de confiance doit prendre en compte que l'échantillonnage systématique est généralement utilisé dans les IFN.

Cartes

Les cartes sont fréquemment utilisées pour présenter les résultats d'IFN et, souvent, pour les non-experts, elles sont plus convaincantes que les statistiques. Des cartes continues de couverture intégrale peuvent uniquement être produites si l'imagerie de télédétection pour toute la surface a été analysée et les modèles respectifs ont été mis au point.

Les cartes de terres forestières/non forestières sont un produit de base qui, accompagné des cartes de biomasse, présente un grand intérêt. Comme pour tous les autres produits d'inventaire forestier, les analystes d'inventaire doivent souligner que les cartes peuvent être inexactes, comme tout produit d'études empiriques. Ainsi, ces incertitudes doivent être correctement documentées et notifiées avec les cartes.

Si des données de télédétection ne sont pas utilisées dans un IFN, les cartes peuvent uniquement être générées à la résolution spatiale de la grille d'échantillon systématique utilisée. Ces cartes ne peuvent pas produire une représentation continue d'une variable cible mais seulement donner une idée préliminaire des distributions spatiales à une échelle assez rustique. Une information est typiquement donnée par point échantillon, et ils sont souvent à une distance de plusieurs kilomètres. La figure ci-dessous en montre un exemple.

Utilisation des données d'IFN pour une optimisation future de la conception de l'IFN

L'analyse des données d'IFN peut aussi servir à l'optimisation de la conception de l'inventaire pour la

planification des futurs inventaires. Bien entendu, on doit être prudent lorsque l'on change la conception de l'IFN entre des cycles successifs car la cohérence est nécessaire dans une série temporelle. Mais une adaptation de la conception menée de temps à autre peut accroître l'efficacité, et, souvent, de nouvelles variables cibles doivent être intégrées afin que les IFN puissent répondre de manière significative à de nouvelles questions émergentes.

En outre, la gestion du temps peut être optimisée par l'introduction de nouveaux dispositifs de mesure. Les mesures de contrôle peuvent être réorganisées et les exactitudes cibles redéfinies. Il faut aussi remarquer que les nouvelles technologies peuvent impliquer des réductions de l'effectif des équipes de terrain.

Dans ce contexte, il peut être intéressant de vérifier si une réduction du nombre de sous-parcelles dans un plan parcellaire en cluster mènerait à une réduction significative de la précision de l'estimation. Il peut s'avérer que pour certaines variables cibles, un nombre inférieur de sous-parcelles serait suffisant. Cette analyse d'optimisation peut facilement être mise en œuvre en réalisant l'analyse des données d'un nombre inférieur de sous-parcelles par cluster, en réduisant ainsi la taille des parcelles par unité d'échantillonnage sélectionnée.

Utilisation des données d'IFN dans le secteur universitaire

La tâche principale (et par défaut) de l'analyse des données d'IFN est de générer les résultats centraux sollicités par les parties prenantes et les décideurs dans l'évaluation des besoins en information. Cependant, les données des IFN sont aussi une source importante pour d'autres usages, comme la recherche et l'enseignement universitaire; parfois, les IFN sont les seuls projets qui génèrent une de données scientifiquement valables sur les forêts ou même les paysages dans un pays entier.

De nombreux sujets peuvent être analysés à partir des données des IFN, y compris des questions méthodologiques (comme les évaluations d'optimisation des conceptions d'inventaire ou l'application ou l'adaptation de modèles) et des questions thématiques (questions de rendement, comparaison des compositions et positions d'espèces, leviers du recul des forêts, etc.). Ces types d'analyses ne font pas partie du travail de base de l'équipe d'inventaire, mais nécessitent que les bases de données soient mises à disposition des chercheurs.

La recherche universitaire qui utilise les données des IFN contribue aussi à éduquer les futurs experts des IFN de sorte que les planificateurs et les analystes des IFN doivent proactivement encourager

l'utilisation des données d'IFN à l'université, dans la recherche et l'enseignement (Liang and Gamarra 2020). En particulier dans les systèmes de suivi des forêts à long terme, des séries temporelles précieuses sont disponibles et permettent un suivi du développement des forêts et de la durabilité des politiques forestières nationales.

Des études de recherche utilisent des **données d'IFN répétés pour actualiser des tables de production** (par ex. Staupendahl et Schmidt 2016) ou pour **identifier l'aptitude des espèces d'arbre face au changement climatique** (par ex. Prasad *et al.* 2020).

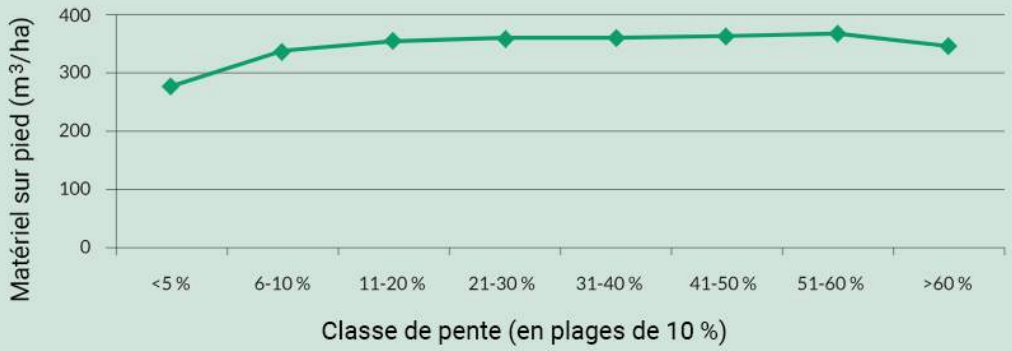
Deux exemples d'études de recherche spécifiques sont l'**estimation de la longueur de la limite des forêts** à partir des données de l'IFN de l'Allemagne à différentes échelles permises par le plan parcellaire dans Kleinn *et al*. 2011; et la **comparaison des stocks d'arbres hors forêt** (AHF) à partir de 12 IFN appuyés par la FAO dans le Sud global (Schnell *et al.* 2015).

Les chercheurs peuvent être intéressés par les ensembles de données sur une grande surface et à long terme des programmes de suivi national des forêts lorsqu'ils souhaitent examiner des questions méthodologiques ou thématiques spécifiques. Par exemple, les échantillons systématiques de parcelles en cluster contiennent de l'information sur la fragmentation des paysages: si les forêts sont plus fragmentées, il y aura plus d'intersections avec les parcelles en cluster et un nombre réduit de parcelles en cluster sera pleinement contenu à l'intérieur ou à l'extérieur de la forêt, sans intersections. À partir du nombre et de la proportion de clusters qui intersectent, on peut dériver des estimations de l'état général de fragmentation – comme le travail présenté par Kleinn (2000) pour une grande surface d'inventaire au Costa Rica – ou des estimations de la longueur de la limite de la forêt pour la région entière ou des unités de référence infranationales, comme l'ont fait Kleinn *et al.* (2011) pour l'Allemagne.

Le saviez-vous?

Lors d'un cours donné par l'un des auteurs, un étudiant a soulevé la question suivante: le matériel sur pied serait-il supérieur dans les classes de pente supérieure car les couronnes des arbres obtiendraient plus de lumière et l'aire de surface serait supérieure à celle d'un plan. Le professeur avait accès à la base de données de l'IFN de l'Allemagne, qui permet des analyses flexibles en

combinant des variables. En reliant le matériel sur pied comme variable de réponse et les classes de pente comme catégories, le graphique ci-dessous a pu être rapidement produit, permettant d'apporter une réponse préliminaire à cette question.



Matériel sur pied (m³/ha) par rapport aux classes de pente (%) – un graphique rapidement produit par l'analyse de données de l'IFN de l'Allemagne répondant à la question d'un étudiant: Le matériel sur pied tend-il à être supérieur sur des pentes plus raides (au moins jusqu'à une limite de pente supérieure) car l'aire de surface du terrain est plus grande?

Principales caractéristiques des produits des analyses de données

Les principaux éléments qui caractérisent les analyses de données sont essentiellement les mêmes que ceux de la publication des résultats, qui ont été formulés en termes généraux comme les principes directeurs du Cadre de transparence renforcé (CCNUCC 2020): **transparence**, **exactitude**, **cohérence**, **exhaustivité**, **comparabilité –** et exposés dans une documentation complète. C'est essentiellement la combinaison de ces caractères qui rend les analyses de données des IFN crédibles pour les parties prenantes et les utilisateurs des données.

Il faut toujours garder à l'esprit que les résultats d'IFN entrent dans le domaine des politiques liées aux forêts dans un pays et qu'il y a différents intérêts en jeu: toutes les parties prenantes ne seront pas contentes des résultats pour de multiples raisons.

Il est donc de la plus haute importance que:

- l'analyse de données soit «blindée» et correcte et puisse être défendue sur la base de la documentation complète et transparente; et

- 'interprétation des conclusions soit compatible avec les résultats des analyses. L'interprétation par différents acteurs peut varier à partir des mêmes résultats et statistiques, selon les idées et valeurs particulières de différents groupes d'intérêt – mais cela n'est alors plus à la portée de l'analyste de données.

Le rôle des analyses de données pour la publication des résultats des inventaires forestiers

Il devrait être désormais clair que l'analyse de données précède la publication des résultats. L'analyse de données a lieu entre la collecte de données/la gestion de données et la publication des résultats. Ainsi, pendant les analyses de données, il est important de s'intéresser à ces deux étapes: d'où proviennent les données (collecte de données/gestion de données) et comment les résultats de l'analyse seront utilisés et traités (élaboration des rapports).

Il est donc impératif que l'analyse de données et la publication des résultats soit étroitement interconnectées et, si différents experts travaillent dans ces domaines, ils doivent travailler ensemble. En conséquence, les résultats préliminaires et immédiats peuvent être discutés, interprétés et comparés, de manière à détecter très tôt les incohérences potentielles. Cela est particulièrement utile car ces incohérences peuvent très bien être l'expression de résultats inattendus ou, plus simplement, être causés par des fautes de calcul ou des fautes dans la collecte de données.

On sous-estime parfois le temps nécessaire aux analyses de données, avec toute la contre-vérification nécessaire de la qualité des données et la cohérence des résultats, pour finalement répondre aux attentes exprimées dans l'EBI.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- IL est utile d'avoir une idée claire des produits potentiels des analyses de données d'IFN durant la phase de planification et l'évaluation des besoins en information (EBI).
- L'information de l'analyse de données forestières peut être utilisées pour générer des statistiques normalisées, appuyer l'élaboration de cartes à partir de la télédétection, servir également à optimiser la conception des futurs IFN, et être utilisé dans la recherche et le secteur universitaire.

- Les cartes constituent une présentation courante et convaincante des résultats d'IFN; pour beaucoup, elles sont souvent plus facilement accessibles et convaincantes que les statistiques.
- L'analyses de données doit s'intéresser à la fois à la source de données (appartenant à la collecte de données/gestion de données) et à la manière dont des produits de l'analyse seront utilisés et traités (élaboration des rapports)