Échantillonnage et estimation dans l'Inventaire Forestier National. Essai de reconstruction et formalisation.

Olivier Bouriaud

Échantillonnage et estimation dans l'Inventaire Forestier National. Essai de reconstruction et formalisation.

Novembre 2020 – v1

Olivier Bouriaud
Laboratoire d'Inventaire Forestier – LIF-IGN

1 Préambule

Ce document a pour objet de présenter les éléments principaux de l'échantillonnage et les méthodes d'estimation mises en œuvre dans l'inventaire forestier national (IFN) depuis sa réforme en 2004, à l'origine de la « nouvelle méthode d'inventaire », qui a vu le plan de sondage passer d'un niveau départemental désynchronisé périodique, à un niveau national systématique annuel.
L'enjeu est de reconstituer un raisonnement logique, permettant d'aboutir aux estimateurs statistiques implémentés depuis 2005, et dont le fondement est inconnu de la communauté d'inventaire à ce jour[1].
La reconstruction des méthodes d'estimation s'appuie sur de rares documents, parfois contradictoires, toujours incomplets (Hervé 2005, 2006 ; Pesty 2017 ; IFN 2017). Aucun ne documente l'approche ou la logique, ni les hypothèses à la base des estimateurs utilisés dans l'enquête. Cet essai de formalisation reflète ainsi les réflexions et efforts menés dans le but de retrouver le cheminement, les hypothèses et la logique de ces estimateurs tels qu'ils ont été initiés par Jean Wolsack puis développés et mis en œuvre par Jean-Christophe Hervé.
Les méthodes d'échantillonnage sont décrites de manière relativement succincte ici, car l'objet est de présenter les aspects de l'échantillonnage qui contraignent l'estimation, et non de réaliser une description exhaustive des éléments de l'échantillonnage qui ont pu changer durant les quinze années d'existence de la nouvelle méthode.

2 Introduction

2.1 Deux populations cibles inventoriées conjointement

L'inventaire forestier national enquête plusieurs attributs simultanément, et notablement, le fait à base d'un même dispositif : i) la couverture et l'utilisation du sol et ii) des arbres. Le premier sujet, la couverture et l'utilisation des sols, est nécessaire pour déterminer la surface de l'objet primaire d'étude qui est la forêt. La localisation et la taille du domaine d'étude forêt n'est pas connue avant enquête, et a même un caractère dynamique. La liaison très forte entre les deux attributs -forêt (type de couverture du sol) et les arbres- justifie cette superposition et le fait que l'on puisse étudier simultanément les deux.
L'inventaire n'étudie pas tous les arbres, il n'étudie que ceux qui sont dans le domaine d'étude, la forêt et autres formations boisées.

Les surfaces forestières (ou par catégories plus fines, ex. hêtraies) forment un sous-ensemble de la surface du domaine étudié. Il existe une infinité de possibilités de segmenter le territoire dans des unités (portions discrètes) de même que l'on peut tirer une infinité de points d'une surface donnée.
La population d'arbres (dans son domaine) est une population discrète, finie dont la taille n'est pas connue avant enquête. Déterminer la taille de la population d'arbres conduit, compte tenu de l'objectif de l'inventaire, à estimer par exemple le nombre d'arbres total, ou la somme des volumes des arbres.

[1] Ces estimateurs ont été formalisés dans leur forme présentée au CNIS en 2017 par Jean-Christophe Hervé, décédé le 16 avril 2017.

Conséquence du fait que ni la taille ni la localisation de la population d'arbres et de la forêt ne sont pas connues avant enquête, le domaine inventorié doit couvrir tout le territoire, et dépasse donc très certainement le **domaine d'étude**.

2.2 Définition du domaine d'intérêt échantillonné

Le domaine d'étude D a une surface connue $\mathbf{A_D}$. Dans un IFN, ce domaine correspond au pays, ou à une fraction administrative telle que le département.
La base de sondage (*sampling frame*) a une surface $\mathbf{A_F}$ (F pour Frame), et est constituée d'une grille qui contient D ($A_F > A_D$). La surface A_F n'est pas nécessairement connue. La grille définissant et contenant toutes les mailles utilisées pour construire les échantillons, elle constitue cette base de sondage.
Le domaine d'intérêt, typiquement la « forêt », a une taille inconnue $\mathbf{A_X}$ (la déterminer est même le premier objectif de l'enquête). Un échantillon est constitué à partir d'une sélection d'un nombre fini de mailles tirés dans la base de sondage. Les points peuvent être à la limite du domaine d'intérêt, mais pas à la limite de la base de sondage. La correction des points limites (tombant à la limite ou même partiellement en dehors du domaine d'étude) n'est ainsi pas nécessaire.
Un sous-domaine, par exemple la forêt dominée par le hêtre comme espèce, ou encore la forêt de structure verticale irrégulière, a une taille inconnue (la déterminer est aussi un objectif de l'enquête), est défini à partir d'une variable catégorielle, et forme une partie de la partition du domaine d'intérêt, définie par cette variable.

La forêt désigne un type de couverture du sol très spécifique faisant l'objet d'une définition internationale (FAO 2004) : terres d'une surface minimum de 0,5 ha, d'une largeur supérieure ou égale à 20 m et de taux de recouvrement des arbres supérieurs ou égal à 10%, où la vocation forestière ne peut a priori être écartée. Les vergers cultivés sont exclus.

On appelle arbre tout végétal ligneux (hors liane) dépassant 5 mètres de haut (hauteur en crête) à maturité in situ.

L'IFN a un champ d'étude excédant la forêt, couvrant en permanence non seulement la forêt mais aussi les bosquets et les landes, dont la définition est basée sur les mêmes critères que la forêt, mais avec des seuils de surface et taux de couvert différents (ex : pour les Landes, taux de recouvrement absolu des arbres inférieur à 10 %, hors arbres épars, sur une surface supérieure ou égale à 5 ares et sur une largeur supérieure ou égale à 20 mètres).
Quelques ordres de grandeur exprimés en fraction de la surface du territoire : la forêt représente 31% de la surface du territoire métropolitain (dont 90% forêt fermée et 10% forêt ouverte), les landes représentent 5%.

2.3 Données cibles de l'IFN

L'IFN produit des données quantitatives assorties de termes d'erreur concernant des surfaces et des volumes (et beaucoup d'autres attributs). Les très nombreux caractères des populations étudiées constituent des clés de ventilation des estimations. Leur combinaison fournit encore plus de clés de ventilation, l'ordre de grandeur du nombre total de combinaisons des modalités étant le million.
Un exemple : type de propriété x sylvoécorégion x classe de diamètre x espèce dominante, sachant que chacune des variables a plusieurs modalités, en ordre de grandeur 3 x 100 x 10 x 100 = 300,000.

Il faut remarquer le fait que les clés de ventilation peuvent provenir de chacune des deux populations, des surfaces et des arbres : typiquement le type de propriété est une variable surfacique, la classe de diamètre et l'espèce dominante proviennent des arbres échantillonnés. Deuxième point essentiel, l'IFN ne produit pas qu'une estimation de l'état des forêts (surfaces et stock sur pied) à un moment donné et de manière cyclique, il fournit aussi une estimation des flux dans les populations : dynamique de la surface forestière[2], accroissement, recrutement, récolte et mortalité s'agissant des arbres.

2.4 Un inventaire continu, annuel, des combinaisons d'estimations annuelles

L'enquête couvre tout le territoire métropolitain de façon annuelle. Chaque année correspond à un échantillon propre, couvrant tout le territoire, appelé « campagne annuelle ». Par construction, l'inventaire combine les estimations issues de cinq échantillons successifs pour produire les chiffres de référence. Cette combinaison prend la forme d'une simple moyenne arithmétique des estimations, et des erreurs[3]. Les efforts sont donc ici orientés sur l'échantillonnage annuel et l'inférence associée.

3 L'échantillonnage

L'échantillonnage est conçu pour répondre aux objectifs de double enquête (des surfaces et des arbres), et pour permettre d'estimer à la fois un état au moment de l'inventaire, et des flux. Les échantillons annuels sont tirés à partir d'une méthode n'ayant pas subi de changement majeur depuis sa mise en place en 2004. Une grille systématique à maille carrée de 1x1 km sert de base de sondage.

3.1 Maillage et échantillonnage annuel. Propriétés interannuelles

A chaque maille de 1x1 km est associée une année d'échantillonnage, de sorte que chaque année possède une fraction constante (approximativement) de mailles, elles aussi disposées de manière systématique dans l'espace.

a)

1	6	2	7	3	8	4	9	5	10	1	6	2
9	5	10	1	6	2	7	3	8	4	9	5	10
3	8	4	9	5	10	1	6	2	7	3	8	4
6	2	7	3	8	4	9	5	10	1	6	2	7
5	10	1	6	2	7	3	8	4	9	5	10	1
8	4	9	5	10	1	6	2	7	3	8	4	9
2	7	3	8	4	9	5	10	1	6	2	7	3
10	1	6	2	7	3	8	4	9	5	10	1	6
4	9	5	10	1	6	2	7	3	8	4	9	5
7	3	8	4	9	5	10	1	6	2	7	3	8
1	6	2	7	3	8	4	9	5	10	1	6	2
9	5	10	1	6	2	7	3	8	4	9	5	10
3	8	4	9	5	10	1	6	2	7	3	8	4

b)

1	6	2	7	3	8	4	9	5	10	**1**	6	2
9	5	10	**1**	6	2	7	3	8	4	9	5	10
3	8	4	9	5	10	**1**	6	2	7	3	8	4
6	2	7	3	8	4	9	5	10	**1**	6	2	7
5	10	**1**	6	2	7	3	8	4	9	5	10	**1**
8	4	9	5	10	**1**	6	2	7	3	8	4	9
2	7	3	8	4	9	5	10	**1**	6	2	7	3
10	**1**	6	2	7	3	8	4	9	5	10	**1**	6
4	9	5	10	**1**	6	2	7	3	8	4	9	5
7	3	8	4	9	5	10	**1**	6	2	7	3	8
1	6	2	7	3	8	4	9	5	10	**1**	6	2
9	5	10	**1**	6	2	7	3	8	4	9	5	10
3	8	4	9	5	10	**1**	6	2	7	3	8	4

Figure 1. **Représentation du pavage de mailles recouvrant tout le territoire** (a), et fractionné en sous-ensembles annuels systématiques (b) : fraction annuelle de l'année 1.

[2] En réalité seulement depuis 2016, en mobilisant l'inventaire répété 2 fois sur les points
[3] Sous l'hypothèse d'indépendance entre échantillons annuels

La grille a été fractionnée en deux sous-ensembles quinquennaux, de sorte qu'un échantillon annuel, un ensemble de 5 échantillons annuels successifs, et un ensemble de 10 échantillons annuels successifs couvrent le territoire de façon encore systématique[4].

3.2 Un échantillonnage annuel en deux phases

Le tirage s'effectue exclusivement dans la base de sondage que constitue la fraction annuelle en cours dans la grille (ex. Figure 1b), c'est à dire dans la sous-population de mailles réparties de manière systématique spatialement et strictement disjointes. Par construction, la couverture des mailles représente environ un dixième de la surface du territoire. En conséquence, les 9 dixièmes du territoire ne peuvent pas être échantillonnés une année donnée.

Lors de la construction de l'échantillon de première phase, un point unique est tiré dans chaque maille et le type de couverture du sol par photo-interprétation est observé sur chacun de ces points.
Lors de la construction de cet échantillon, le point tiré dans chaque maille a des coordonnées géographiques aléatoires au sein de la maille. La procédure de transition d'un échantillon de mailles à un échantillon de points n'est toutefois pas considérée comme un degré d'échantillonnage. Il s'agit d'une randomisation géographique (Cochran 1977, page 184, « unaligned systematic sampling » ou échantillonnage systématique non aligné), qui a pu avoir comme fin d'amoindrir les risques de coïncidence entre des formes de relief ou des limites de forêt d'une part, et les lignes directrices de la grille[5]. L'aspect de structure spatiale au sens des variables régionalisées n'est pas négligé, et pourrait avoir contribué à ce souci de découplage spatial, bien qu'aucune étude n'ait jamais été menée pour rechercher une covariance à cette échelle spatiale (typiquement, $10^{1/2} \sim 3.16$ km entre points en première phase).

L'échantillon de seconde phase est constitué par une fraction de l'échantillon de première phase, en vue d'une visite sur le terrain après introduction d'un support d'échantillonnage (la « placette »), par exemple pour la mesure des arbres (dans le jargon forestier, ces points sont « levés »). Chaque point permet d'enquêter les arbres situés dans la placette centrée sur ce point.

Tableau 1. **Taux de tirage des points de deuxième phase en fonction de la catégorie d'utilisation du sol**.

Catégorie d'utilisation du sol (résultat de la phase 1)	*Taux d'échantillonnage*
Non forêt	100%
Forêt fermée	50%
Forêt ouverte	50%
Landes	25%
Bosquets[6]	25%

Le taux d'échantillonnage des points pour la deuxième phase dépend de la modalité de la variable au point comme ce serait le cas dans une stratification, conférant un caractère **d'échantillonnage à probabilités inégales** à l'échantillonnage de deuxième phase. Les

[4] Pour plus de détails, on pourra consulter Morneau et al. 2020, Forest Ecosystems, DOI : ….
[5] Aux mêmes fins, l'échantillonnage dans l'inventaire forestier américain s'appuie sur une tessélation hexagonale du territoire.
[6] Les bosquets, définis par une surface comprise entre 0.05 et 0.5 ha, ne font en toute rigueur pas partie intégrante du domaine d'intérêt, issu de la définition internationale de la forêt, adoptée par l'inventaire en 2005. Toutefois, selon le maintien d'un usage ancien (antérieur à 2005), ils ont continué à être échantillonnés à partir de 2005. Ils représentent environ 100,000 ha.

effectifs associés à ces deux phases sont très différents, les ordres de grandeur étant de 60,000 points pour la première phase et de 6,000 à 8,000 pour la deuxième phase.

Le sous-échantillon est tiré en s'appuyant sur la maille, avec un taux variable selon la catégorie de végétation (mais fixe au sein du domaine d'échantillonnage) comme présenté dans le tableau 1. Le taux peut être ajusté et être inférieur au taux maximum une année donnée, mais cela ne change rien aux estimations.

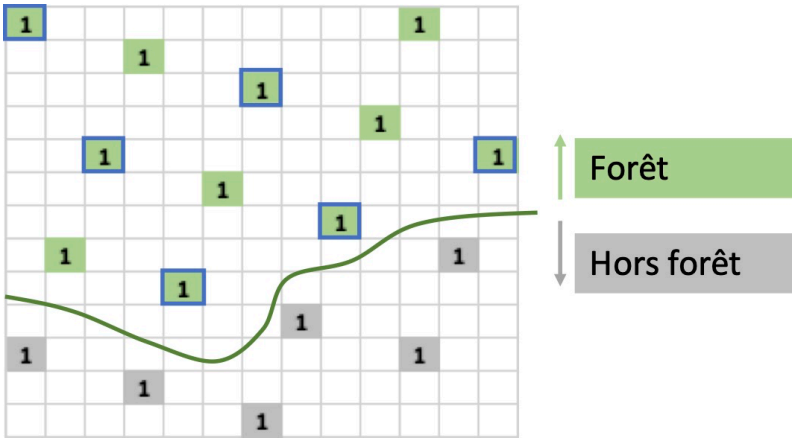

Figure 2. **Illustration du sous-échantillonnage de deuxième phase appuyé sur le pavage de maille**.
Le schéma représente une fraction annuelle donnée, dont les points photo-interprétés tombent soit en forêt (en vert) soit en dehors (en gris). Les points sous-échantillonnés pour faire partie de l'échantillon de phase 2 sont entourés en bleu (taux de 50% dans l'exemple).

Un taux d'échantillonnage de 50% signifie par exemple que la moitié des points de l'échantillon de première phase de la catégorie Forêt fermée seront intégrés de l'échantillon de deuxième phase. Ce taux est obtenu en se basant sur les propriétés homothétiques de la grille de sondage (Figure 3), qui permettent de retenir une fraction des points (ou des mailles) toujours répartie systématiquement et formant toujours des carrés : ces sous-ensembles, bien décrits dans les documents existants sont appelés « niveaux », et définissent des échantillons dont la taille successive est dans un rapport constant de 2. Passer d'un niveau à un autre impose de diviser par deux le nombre de points.

L'échantillon de seconde phase résulte donc d'un tirage aléatoire systématique dans l'échantillon de seconde phase, avec une taille qui s'en déduit selon la modalité de la variable UCS, c'est-à-dire finalement d'un échantillonnage systématique aléatoire.

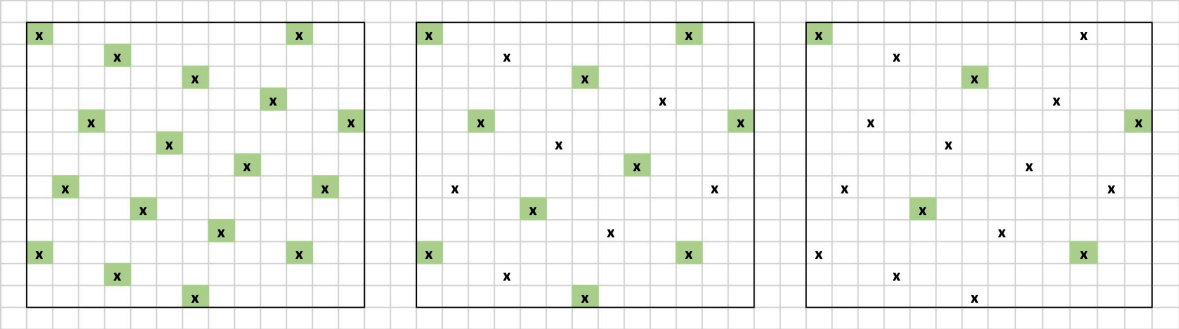

Figure 3. **Niveaux d'échantillonnage emboités** : à gauche, le niveau 1 est l'ensemble des mailles de la fraction annuelle, au milieu le niveau 2 représente un sous-échantillon de 50% de l'effectif de points, le niveau 3 à droite divise à nouveau par deux le nombre de points tout en gardant les propriétés d'équidistance et de couverture spatialement systématique.

3.3 Définition de « post-strates »

Jusqu'ici on pourrait avoir un échantillonnage assez classique et donc des estimateurs classiques en deux phases pour stratification. Mais il est sans doute apparu rapidement que les tirages annuels de l'échantillon de deuxième phase résultaient dans un nombre de points localement assez faible, et que cela poserait des problèmes d'estimations et surtout d'inférence (calcul des erreurs). Il a alors été décidé de procéder à des regroupements des catégories utilisées pour les tirages (forêt fermée, forêt ouverte, etc.). Le détail de l'algorithme ne sera pas présenté ici, car il n'a pas en soi de fondement statistique et il semble même que cet algorithme, qui avait pour mission principale de constituer des strates contenant au moins 10 points de deuxième phase, ne converge pas dans un très grand nombre de situations.
Ce qui importe pour les estimations c'est que le regroupement résulte dans l'assemblage, dans une même post-strate, de points de deuxième phase ayant été tirés avec des taux différents. En découplant le tirage de l'estimation, on constitue des strates regroupant des points ayant des poids inégaux, par exemple en formant une strate groupant forêt ouverte (taux de 50%) et bosquets (taux de 25%).

Ainsi, l'échantillonnage est un échantillonnage à probabilités inégales. Cette caractéristique est majeure et différentie tout à fait l'échantillonnage et les estimations de l'IFN Français de ceux des autres inventaires dans le monde, parce que le taux de tirage de la deuxième phase n'est pas constant au sein des post-strates.
A priori, il s'agit d'un avantage et la possibilité de faire des regroupements donne à l'enquête une capacité d'adaptation intéressante même si cela complique le développement des estimateurs. Des strates regroupant des points de poids égaux est un cas particulier de cette formalisation plus générale.

4 Estimations : post-stratification, estimateurs

4.1 Échantillonnage d'une population continue et articulation entre la population de surfaces et celle d'arbres

Tirer un point dans une maille, ou plus généralement un point dans un domaine, c'est choisir un point dans une population infinie. Le taux d'échantillonnage n'est pas calculable. On lui substitue alors l'estimation de la probabilité de tirage définie par l'inverse de la surface du domaine dans lequel il est tiré.

Soit D la projection plane du domaine inventorié sur R^2 de surface connue A_D. La probabilité d'inclusion d'un point donné *x* dans le plan D est $f(x) = 1/A_D$, f(x) représentant la fonction de probabilité affectée à chaque point. Tous les points ont la même probabilité de tirage, comme c'est le cas ici pour une région donnée (propriété intéressante du tirage systématique !), de sorte que f(x) = constante.

Densité spatiale moyenne
Le calcul de valeurs totales implique l'utilisation d'une **densité spatiale moyenne** du paramètre étudié. La difficulté principale repose sur le fait que la moyenne est estimée sur un échantillon pour lequel on ne peut pas calculer le taux d'échantillonnage.

On s'intéresse à un paramètre de la population *y* continu dans l'espace à deux dimensions R^2 de surface connue A_D. Si on mesurait la valeur de y en tout point x, la densité moyenne de y pourrait s'écrire

$$\bar{y} = \frac{1}{A_D} \iint_D y(x).dx$$

et le total de *y* sur A_D serait $\tau_y = \iint_D \ y(x).dx$

que l'on peut écrire aussi à base de la fonction de densité de y (notée ρ) : $\tau_y = \int_{A_D} \rho(x)dx$

Le passage d'une population continue à une population discrète se fait par le principe dit de dualité, qui remplace ou approche la densité moyenne calculée par intégrale par une estimation basée sur un échantillon discret et fini.
La densité moyenne est alors estimée à base d'un échantillon de taille n, telle que

$$\tilde{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

En d'autres termes on approche la densité spatiale moyenne par un calcul de moyenne empirique basée sur un nombre de points tirés dans le domaine.
Mandallaz (1991) le présente dans le théorème 3.1.3.1, disant que « calculer une somme d'éléments discrets d'une population finie est équivalent à intégrer une fonction sur un domaine, c'est à dire sur un continuum ».

Dans l'approche mise en place, les points des échantillons sont plutôt « injectés » que réellement « tirés » dans D, en ce sens qu'on ignore volontairement la population infinie de laquelle ils sont issus, et la structure de l'échantillon, qui est systématique spatialement, et qui interdit tout à fait la superposition de points conduisant ainsi à la nullité de toutes les probabilités conjointes de tirage.

Illustration du raisonnement :
On constitue l'échantillon Y_1 de taille N_1 dans D, l'indice 1 signifiant qu'on se positionne dans la première phase. L'objet de l'échantillonnage est d'estimer les proportions de *k* catégories de surfaces (utilisation et couverture du sol) dans D.
Soit $y_{1i}(l)$ la valeur de l'attribut *y* pour un point quelconque *i* de N_1, tel que $y_{1i}(l) = 1$ dans *l*, 0 sinon. Typiquement *l* est l'indicateur de la présence de la forêt et y_{1i} représente l'issue d'une photo-interprétation. Alors on peut définir le vecteur des proportions empiriques *F* dans D (Fattorini et al. 2009) : $F = {}^t[F_1,\ldots, F_k]$ où pour tout $l \in [1, \ldots, k]$,

$$F_l = \frac{1}{N_1}\sum_{i=1}^{N_1} y_{1i}(l)$$

Puisque $\frac{1}{N_1}\sum_{i=1}^{N_1} y_{1i} = 1$, $\frac{F_l}{1} = \frac{F_l\, A_D}{A_D} = \frac{A_l}{A_D}$

Le vecteur F est un **estimateur de la proportion de la surface** des *k* catégories (modalités de UCS) tel que

$$\hat{F}_l = \frac{A_l}{A_D}$$

Les proportions de points tombant dans chaque catégorie sont donc des estimateurs des proportions des surfaces de chacune de ces catégories.

Selon le principe d'échantillonnage dans une population continue, le poids des points tirés dans D est constant et fixe. On note que, dans le contexte d'un tirage de points sur une grille, le nombre de points tombant dans le domaine pour une taille donnée de grille est une variable aléatoire d'espérance finie et bornée. Autrement dit le poids à un tirage quelconque n'est égal au poids théorique qu'en espérance (Valentine et al. 2009). On remarquera alors que, tous les points étant tirés à probabilité constante donc égale, la valeur du poids n'est plus nécessaire elle-même !

Cette approche du calcul des paramètres est valable pour les variables cibles liées aux surfaces et aux arbres. Pour les surfaces, la moyenne est une proportion d'une catégorie donnée. Pour les arbres, la moyenne est un volume, une surface terrière ou un nombre d'arbres etc., par catégories.

Par exemple en notant y le volume moyen, suivant ce principe le volume total dans D est

$$\tau_y = A_D \sum_{i=1}^{N} \frac{y_i}{N}$$

Du fait de l'échantillonnage en deux phases et de la stratification, les moyennes sont calculées sur des strates de calcul de taille inconnue (dans les exemples précédents, A_D est connu), déterminées lors de la première phase. Le principe fondamental (total = surface x moyenne spatiale) reste inchangé.

4.2 Du rôle de la placette dans la seconde phase

Tout se passe comme si on avait constitué un échantillon de première phase, systématique spatialement, en dehors de toute considération de coordination spatiale liée à la grille.

De cet échantillon systématique est tiré un autre échantillon de points, transformés en placettes sur le terrain dont le centre est centré sur le point.
La transformation des points en placettes (i.e. des entités de surface non nulle) est nécessaire à l'échantillonnage des arbres mais impose une certaine réflexion dans le cadre de l'estimation.

Les populations que l'on manipule sont de fait :
- une population d'arbres, de taille inconnue, discrète, finie
- une population de placettes, de taille inconnue, infinie
L'articulation entre la population de placettes et celles des arbres est décrite dans le cadre de la méthode de « partage des poids ».

4.3 La méthode de partage des poids

Proposée par Guillaume Chauvet dans le cadre de l'échantillonnage de l'IFN, la méthode de transfert des poids donne un cadre théorique à une opération réalisée implicitement par les inventoristes forestiers en reconnaissant explicitement l'existence de deux populations. Le principe de la méthode est de transférer les poids de tirage d'unités d'une population pour lesquels ils sont connus à des unités pour lesquels ils ne le sont pas, mais qui sont associées aux unités de la première population.
L'association entre les unités des deux populations doit être parfaitement connue, et les unités dont le cadre d'échantillonnage n'est pas connu doivent être associées à au moins une des unités de la population pour laquelle le cadre est connu.
Dans le cas de l'IFN, les unités cibles, les arbres, n'ont pas un cadre d'échantillonnage connu, vu que l'on ne connaît ni leur nombre ni leur localisation. La population de placettes devient donc celle que l'on peut échantillonner de manière probabiliste et devient de fait la population d'unités parentes.
Cette méthode permet ainsi de transférer le poids d'échantillonnage des placettes aux arbres, et le poids des arbres dans les placettes (concentriques) sera pris en compte dans l'estimation de valeurs totales niveau placette. Un article est en cours de rédaction sur ce sujet (G. Chauvet, O. Bouriaud, P. Brion).

5 Estimation des totaux

Par la suite, on considèrera que tous les paramètres mesurés sur les unités d'échantillonnage peuvent être rapportés à un point (valable pour les deux phases). Les estimateurs ne « manipulent » pas de mailles ni des placettes, seulement des points.

Les estimations sont réalisées non pas à l'échelle du territoire en entier, mais au niveau départemental. Les totaux et les erreurs résultent de la somme des valeurs départementales (la variance est cumulative dans le contexte d'un découpage spatial complémentaire sans chevauchement).
Ainsi le raisonnement présenté dans ce chapitre porte sur un domaine d'étude correspondant à un département, sans nuire d'ailleurs à la généricité du propos. La surface du domaine est connue sans erreur et ne fait pas l'objet d'un échantillonnage. On notera donc A_D cette surface, avec une coïncidence entre de D comme domaine et D comme département.

5.1 Total d'une variable liée à des proportions de surface

La variable étudiée est ici par exemple la proportion de forêt dans la surface totale, ou tout autre sous-domaine, estimé sur chaque point de deuxième phase à l'aide d'une variable catégorielle à deux ou plusieurs modalités.
L'estimation du total fait appel à un estimateur en deux phases pour post-stratification, la première phase étant utilisée pour estimer la taille des strates (post-strates), la deuxième phase apportant l'estimation des moyennes de la variable dans chacune des strates.
Compte tenu du principe de dualité (assimilation d'une population infinie à une population discrète), l'estimation de la moyenne de la variable *y* dans chaque strate est basée sur l'échantillon de deuxième phase. Mais les points de deuxième phase ont une probabilité d'inclusion inégale, selon l'issue de la photo-interprétation. Ainsi, à chaque point correspond une probabilité d'inclusion donnée, égale à l'inverse de son poids – lequel est noté *w*. Les documents existants et les bases de données utilisent plutôt le poids que la probabilité d'inclusion dans la formalisation et le calcul des estimations. Pour tenir compte des poids variables des points, la moyenne est calculée comme une moyenne pondérée pour tout strate donnée *h* :

$$\hat{\bar{m}}_{hy} = \sum_{i=1}^{n_{2h}} \frac{w_i y_i}{w_i}$$

n_{2h} représente le nombre de points de deuxième phase dans la strate *h*.

Le total de la variable *x* dans le domaine résulte de la sommation des totaux sur chaque strate, car celles-ci sont par construction complémentaires :

$$\hat{\tau}_y = \sum_{h=1}^{H} A_h \hat{\bar{m}}_{hy} = \sum_{h=1}^{H} A_h \sum_{i=1}^{n_{2h}} \frac{w_i y_i}{w_i}$$

où H est le nombre total de post-strates, A_h la surface de chaque post-strate. Cette surface n'est pas connue avant échantillonnage, elle émane des mesures faites sur la première phase, et est donc estimée. Comme présenté plus haut, cette estimation de surfaces se base sur la proportion de points de phase 1 tombant dans chacune des post-strates, les proportions de points étant des estimateurs des proportions de surfaces. En notant n_{1T} le nombre total de points de phase 1, et n_{1h} le nombre de points tombant dans la strate *h*, on a :

$$\hat{\tau}_y = A_D \sum_{h=1}^{H} \frac{n_{1h}}{n_{1T}} \sum_{i=1}^{n_{2h}} \frac{w_i y_i}{w_i}$$

Cette formule est tout à fait centrale dans l'IFN, et résume la philosophie de l'estimation en deux phases avec post-stratification avec poids variables en deuxième phase.

5.2 Total d'une variable liée aux arbres

Cette fois *y* est un attribut de la population des arbres, par exemple le volume. Le volume de tous les arbres de la placette est sommé et rapporté à l'hectare, le transformant en une variable continue spatialement.
En comparaison de l'estimation de proportions de surfaces, les estimations reposent sur une étape supplémentaire, sorte d'héritage de « l'ancienne méthode » qui avait trois phases. Ces trois phases correspondaient à 1) la photo-interprétation, 2) la phase dite de reconnaissance dans laquelle les points terrain sont classés dans des catégories d'utilisation et couverture du sol

(d'où les possibles reclassements), 3) les points terrain levés, chaque phase ayant un nombre décroissant de points.

Dans la « nouvelle méthode » à base de l'échantillon de deuxième phase (s_2), on estime successivement :
- la proportion de la surface couverte par une catégorie spécifique *k*, observée exclusivement sur l'échantillon de deuxième phase et qui définit un sous-domaine d'intérêt la proportion de la surface couverte par une catégorie spécifique *k*, observée exclusivement sur l'échantillon de deuxième phase et qui définit un sous-domaine d'intérêt (forme de troisième phase, puisque selon la partition opérable à partir de ce facteur, un sous-ensemble de l'échantillon de phase 2 sert de support à l'estimation). Un tel sous-domaine est par exemple, soit la forêt, soit une catégorie spécifique comme la hêtraie (au sens de la surface localement dominée par une essence, sans référence à la population d'arbres, il s'agit ici de taux de couverts).
- la moyenne de l'attribut *y* dans le sous-domaine (et seulement dans le sous-domaine).
Deux moyennes sont ainsi calculées : une proportion moyenne (indicée 2 pour 2ème phase et la distinguer de la proportion donnant la taille de la strate), et la moyenne pondérée de l'attribut :

$$\hat{\tau}_y = A_D \sum_{h=1}^{H} \frac{n_{1h}}{n_{1T}} \left\{ \left(\sum_{i=1}^{n_{2h}} \frac{w_i\, I_i(k)}{w_i} \right) \left(\sum_{i \in (s_{2h} \cap\, k)} \frac{w_i\, y_i}{w_i} \right) \right\} = A_D \sum_{h=1}^{H} \frac{n_{1h}}{n_{1T}}\, \bar{P}_{2hk}\, \bar{M}_{hk}(y)$$

où

$$\bar{P}_{2hk} = \sum_{i=1}^{n_{2h}} \frac{w_i\, I_i(k)}{w_i}$$

est la proportion moyenne du sous-domaine *k* dans la strate *h*, calculée comme la moyenne pondérée de l'indicatrice d'appartenance au sous-domaine k, $I_i(k)$; et

$$\bar{M}_{hk}(y) = \sum_{i \in (s_{2h} \cap\, k)} \frac{w_i\, y_i}{w_i}$$

est la moyenne de *y* dans le sous-domaine *k* et la post-strate *h*, estimée exclusivement sur le sous-ensemble de points de deuxième phase appartenant à *k*. Cette moyenne pourrait être indicée 2 aussi car basée exclusivement sur les points de deuxième phase, mais cela alourdi les notations.
On vérifie que $s_{2h} \cap k \subseteq s_{2h}$ avec égalité possible dans certaines situations probablement assez rares. Plus le domaine est spécifique, plus l'écart d'effectif de points augmente.

En quoi est-ce différent des autres inventaires ?
Dans les autres inventaires, la moyenne de l'attribut est estimée sur tout l'échantillon de deuxième phase, qu'il soit dans le sous-domaine ou non. Ici on ne considère que les points tombant dans le sous-domaine (défini par la catégorie *k*). L'effectif de points utilisé pour la moyenne est donc plus petit, mais il ne contient plus les valeurs nulles propres aux échantillons incluant des points en-dehors du sous-domaine.
Il s'agit de fait d'une sorte de post-stratification dont on éliminerait une strate sachant que la valeur de l'attribut y est toujours nulle.

Le cas des estimations dont la ventilation fait intervenir exclusivement des variables mesurées sur les arbres, ou des sous-domaines définis à la fois par un critère de surface (la hêtraie) et un

critère d'arbre (par exemple diamètre > 50 cm), est un cas particulier de ces derniers estimateurs qui ne change rien à la logique du calcul.
En effet, pour ces situations, il faut envisager deux catégories séparément pour le sous-domaine (ex. k_1) et pour le calcul de la moyenne (k_2) dans *h* et dans k_1. On joue donc sur la moyenne, qui n'est calculée que sur la condition k_2 :

$$\bar{P}_{2hk_1} = \sum_{i=1}^{n_{2h}} \frac{w_i\, I_i(k_1)}{w_i}$$

$$\bar{M}_{hk_2}(y) = \sum_{i \in (s_{2h} \cap k_2)} \frac{w_i\, y_i}{w_i} = \sum_{i=1}^{n_{2h}} \frac{w_i\, y_i I_i(k_2)}{w_i}$$

Dans les notes fournies au CNIS (IFN, 2017), on trouve l'écriture suivante des estimateurs, strictement équivalente (pour le cas général où $k_1 = k_2 = k$) :

$$T_f(y) = \sum_k S_k P_{fk} M_{fk}(y)$$

où S_k est la surface de la strate *k*, le sous-domaine est noté *f*, et P_{fk} est la proportion de sous-domaine *f* dans la strate *k*, M_{fk} la moyenne dans *f* et *k*. L'utilisation de l'indice *k* pour une strate est toutefois assez délicat, la littérature préférant largement utiliser *h* pour une strate, *k* pour une catégorie et *f* pour un taux d'échantillonnage. S est traditionnellement une variance (ou un écart-type), pas une surface qui est plutôt notée A (aire, area).

On peut remarquer que cette méthode d'estimation donne aux points de deuxième phase une surface d'extension spécifique et que cette vision est sans doute à l'origine des estimateurs eux-mêmes. En effet, dans un domaine donné, le total d'une variable donnée *y* se calcule comme le produit de la surface du domaine (A_D) par la moyenne de la densité spatiale dans le domaine (pour l'exemple, prenons un échantillon S de taille N) :

$\hat{\tau}_y = A_D\, \bar{y}_D = \sum_{i=1}^{N} b_i\, y_i$ avec $\bar{y}_D = \frac{1}{N} \sum_{i=1}^{N} y_i$ et $b_i = \frac{A_D}{N}$

Cette expression est équivalente à affecter à chaque point *i* une surface d'extension, notée b_i. Cette surface peut être prise comme étant la taille de la maille de la grille servant de support aux points : pour une grille kilométrique chaque point représente 1 km^2, donc on pourrait calculer le total comme le produit de la moyenne par 1 km^2. Cette surface serait constante dans le cas de la grille, car systématique.
Dans le cas de l'IFN, cette surface dépend du niveau, et l'on pourrait écrire de manière générique :

$$\forall\, i, b_i = \frac{A_D}{2^{niveau(i)}} = A_D\, 2^{-w_i}$$

constant pour un niveau de grille donné.

En espérance ce n'est pas faux, mais on se rend bien compte que, à niveau constant, le nombre de points tombant dans le domaine varie en fonction de la position de la grille (du point d'ancrage de la grille) et des formes des limites du domaine. Ainsi, comme dit plus haut, le nombre de points tombant dans le domaine est une variable aléatoire elle-même. En conséquence, la surface représentée par les points du domaine, ou leur facteur d'expansion, varie également de manière aléatoire.
En tenant compte de la réalisation du tirage et du poids possiblement variable des points, on peut alors écrire :

$$b = \frac{A_D}{\sum_{i \in S} w_i}$$

d'où l'estimation du total :

$$\hat{\tau}_y = \frac{A_D}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \, y_i = A_D \sum_{i=1}^{N} \frac{w_i \, y_i}{w_i}$$

Ceci permet d'introduire un aspect très important de l'estimation : le nombre de points tombant dans un échantillon est une variable aléatoire, dont la réalisation est le premier résultat du tirage. Plus que le positionnement des points (i.e. aux aspects systématiques spatialement de la grille) on s'intéresse à leur nombre dans le domaine d'estimation et à la somme de leurs poids. Ce caractère aléatoire du nombre de points tombant dans le domaine justifie le conditionnement utilisé pour la construction des estimateurs de variance, ce qui fait la transition avec la section suivante.

6 Estimation des erreurs, estimation de la variance

Les erreurs associées aux estimations sont des erreurs d'échantillonnage. Elles sont établies dans le cadre stricte d'une approche à base d'échantillonnage (*design based*). Cette approche signifie que les valeurs des paramètres observées aux points sont considérées comme fixes, et que seules les probabilités de tirage des points comptent dans la variance de l'estimation. Tous les inventaires forestiers reposent sur ce principe, qui n'est donc pas du tout spécifique à l'inventaire français, pour deux raisons simples : une, cette approche est la seule permettant une additivité sur les sous-domaines et un raisonnement strictement identique quel que soit le paramètre étudié, deux, elle reflète une vision historique du sondage.

6.1 Variance de la surface d'un sous-domaine

Pour commencer la présentation des estimateurs de variance, on se place dans un domaine D de surface connue A_D. On rappelle ici la forme du total de la surface du sous-domaine *k* dans D :

$$\hat{\tau}_k = A_D \sum_{h=1}^{H} \frac{n_{1h}}{n_{1T}} \bar{P}_{2hk}$$

qui fait intervenir deux termes :
- le premier est lié à la première phase permet l'estimation de la taille de la (post-)strate *h* dans D, que l'on peut noter $P_{1h} = \frac{n_{1h}}{n_{1T}}$
- le deuxième est la proportion moyenne du sous-domaine *k* dans les strates et dans D, issue de la deuxième phase

La variance du total s'écrit alors :

$$var(\hat{\tau}_k) = var\left(A_D \sum_{h=1}^{H} P_{1h}\, \bar{P}_{2hk}\right) \qquad (1)$$

Pour sa démonstration, l'estimateur de variance peut se simplifier en factorisant la surface du domaine, laquelle interviendra à la fin dans l'estimation de la variance au carré :

$$var\left(\frac{\hat{\tau}_k}{A_D}\right) = \frac{1}{A_D^2} var(\hat{\tau}_k) = var\left(\sum_{h=1}^{H} P_{1h}\, \bar{P}_{2hk}\right)$$

S'agissant d'une somme sur les H strates du domaine,

$$var\left(\sum_{h=1}^{H} P_{1h}\, \bar{P}_{2hk}\right) = \sum_{h=1}^{H} var(P_{1h}\bar{P}_{2hk}) + \sum_{h=1}^{H}\sum_{l\neq h}^{H} cov(P_{1h}\bar{P}_{2hk}, P_{1l}\bar{P}_{2lk}) \qquad (2)$$

Pour le premier terme, somme des variances propres aux strates, on conditionne par rapport au nombre de points de première phase et de deuxième phase tombés dans la strate *h*, et d'après le théorème de la variance totale on a :

$$\begin{aligned} var(P_{1h}\bar{P}_{2hk}|\, n_{1h}) &= var\{E(P_{1h}\bar{P}_{2hk}|\, n_{1h})\} + E\{var(P_{1h}\bar{P}_{2hk}|\, n_{1h})\} \\ &= var\{P_{1h}\, E(\bar{P}_{2hk}|\, n_{1h})\} + E\{var(P_{1h})\, var(\bar{P}_{2hk})\} \\ &= P_{1h}^2\, var(\bar{P}_{2hk}) + var(P_{1h})\, var(\bar{P}_{2hk}) \end{aligned}$$

$$= \left(P_{1h}^2 + var(P_{1h})\right) var(\bar{P}_{2hk})$$

La variance de la proportion moyenne de *k* dans la strate *h* peut s'estimer comme (variance d'une proportion) :

$$\widehat{var}(\bar{P}_{2hk}) = \frac{\bar{P}_{2hk}(1-\bar{P}_{2hk})}{n_{2h}-1}$$

D'où finalement

$$\widehat{Var}(P_{1h}\bar{P}_{2hk}|\, n_{1h}) = \left(P_{1h}^2 + var(P_{1h})\right)\left(\frac{\bar{P}_{2hk}(1-\bar{P}_{2hk})}{n_{2h}-1}\right)$$

Mais les points de deuxième phase n'ont pas tous le même poids, et $\bar{P}_{2hk}$ est en fait une moyenne pondérée. On substitue alors aux effectifs de points de deuxième phase n_{2h} un terme qui tient compte de l'existence de poids variables entre unités d'échantillonnage dans le domaine, selon une quantité synthétique dénommée « effectifs équivalents d'équiprobables » et enveloppée d'un certain mystère.

6.2 Les effectifs équivalents d'équiprobables

Leur définition dans les notes de spécification des estimateurs est *in extenso* la suivante :
« Pour une partie quelconque de l'échantillon, on appelle *effectif équivalent d'équiprobables*, et on note généralement neq, la quantité : neq = {somme (wi)}^{2} / somme (wi^2)
pour des points *i* appartenant à l'échantillon considéré [lequel est toujours de deuxième phase]. » (Hervé 2006, 2007)

On pourrait appeler plus explicitement neq des « poids quadratiques moyens inverses ».
Ces quantités trouvent leur origine dans l'utilisation de poids variables dans l'échantillon de deuxième phase (S_2) et n'ont pour objet que de simplifier la notation et l'implémentation des calculs de variance.

Soit la $\bar{m}_x$ la moyenne pondérée d'un attribut *x* dans S_2, où w_i est le poids de l'unité *i* dans S_2, et soit s^2 l'estimation sur S_2 de la variance de *x* :

$$E(s^2) = \frac{1}{w_i}\sum_{i\in S_2} w_i E(x_i - \bar{m}_x)^2 = s^2 - 2\,s^2\frac{\sum_i\ w_i^2}{\left(\sum_i\ w_i\right)^2} + var(\bar{m}_x)$$

or, $var(\bar{m}_x) = \frac{\sum_i\ w_i^2}{\left(\sum_i\ w_i\right)^2}\ var(x)$, soit encore

$$\widehat{var(\bar{m}_x)} = \frac{\sum_i\ w_i^2}{\left(\sum_i\ w_i\right)^2} s^2 = \frac{s^2}{neq}$$

Cela permet donc d'écrire simplement la variance de la moyenne en fonction de la variance empirique s^2.

d'où $E(s^2) = s^2\left(1 - \frac{\sum_i\ w_i^2}{\left(\sum_i\ w_i\right)^2}\right) = s^2\left(1 - \frac{1}{neq}\right) = s^2\left(\frac{neq-1}{neq}\right)$

Le deuxième terme de l'équation (2) représente la covariance entre strates et provient du fait que les effectifs de première phase dans chaque strate ne peuvent pas être considérés comme indépendants (par leur lien au nombre total, la somme des proportions des strates valant 1 par construction).

$$cov(P_{1h}\bar{P}_{2hk}, P_{1l}\bar{P}_{2lk}) = \bar{P}_{2hk}\,\bar{P}_{2lk}\,cov(P_{1h}, P_{1l})$$

or, $cov(A_h, A_l) = cov(A_D P_{1h},\ A_D P_{1l}) = -\frac{P_{1h}\,P_{1l}}{n_1} A_D^2$
ou en ne conservant que les proportions, $\widehat{cov}(P_{1h}, P_{1l}) = -\frac{P_{1h}\,P_{1l}}{n_1 - 1}$

Ce résultat est classique, il s'agit du terme de covariance entre deux proportions quelconques de deux modalités (h, l) d'une loi multinomiale.
Un exemple de démonstration de l'estimation de la covariance est fourni chez par Tam 1985. Ici le signe négatif peut s'illustrer aussi comme :

$$cov(A_D P_{1h}\,, A_D P_{1l}) = E[P_{1h}, P_{1l}] - E[P_{1h}]E[P_{1l}] = 0 - E[P_{1h}]E[P_{1l}]$$

car du fait de leur complémentarité spatiale, $E[P_{1h}, P_{1l}] = 0$

De même, les variances des surfaces des strates sont binomiales, on retrouve bien le fait que
$var(A_h) = var(A_D P_{1h}) = A_D^2\ var(P_{1h}) = A_D^2 \frac{P_{1h}(1-P_{1h})}{n_1} = cov(A_h, A_h)$

On retrouve ces mêmes covariances chez Little 1993, et Valentine et al. 2009 toujours bien négatives (intuitivement, cela se comprend bien, une augmentation d'effectif dans une strate *h* ne pouvant correspondre en espérance qu'à une baisse dans une autre modalité *l*, à effort d'échantillonnage fixé).

En conclusion l'estimateur de la variance (Eq. 2) vaut :

$$\begin{aligned} var\left(\sum_{h=1}^{H} P_{1h}\ \bar{P}_{2hk}\right) & \\ &= \sum_{h=1}^{H}\left(P_{1h}^2 + var(P_{1h})\right)\left(\frac{\bar{P}_{2hk}(1-\bar{P}_{2hk})}{n_{2h}-1}\right) \\ &+ \sum_{h=1}^{H}\sum_{l\neq h}^{H} \bar{P}_{2hk}\,\bar{P}_{2lk}\,cov(P_{1h}, P_{1l}) \end{aligned} \qquad (3)$$

La variance de l'estimation d'une surface est ainsi la somme de deux termes, le premier étant nommé **variance de sous-domaine**, le deuxième la **variance de stratification**. Cet estimateur intègre ainsi les deux sources de variance que sont i) la variance spatiale d'échantillonnage du sous-domaine, et ii) l'erreur d'estimation des surfaces des strates qui est due à la post-stratification, et qui est liée à la première phase d'échantillonnage.

6.3 Variance d'un total d'une variable arbre niveau point

Le total de *y*, attribut d'intérêt dans $D \supset k$ est le produit de trois termes :

$$\hat{\tau}_y = A_D \sum_{h=1}^{H} P_{1h}\ \bar{P}_{2hk}\ \bar{M}_{hk}(y)$$

Par rapport à l'estimation d'une surface il fait intervenir en plus la moyenne de l'attribut *y* dans la catégorie *k* de sous-domaine (portion du domaine, sous-élément de la strate *h*), $\bar{M}_{hk}(y)$, estimée à base des points *i* de l'échantillon S_2 dans *k* et 0pour lesquels *y* est non nul.
La moyenne M_{hk} n'est a priori et en général pas issue du même nombre de points que la proportion du sous-domaine, et est par construction nulle en dehors du sous-domaine.

En parfaite analogie avec l'estimation de la variance de la surface d'un sous-domaine, la variance du total dans sous-domaine est estimée comme :

$$var(P_{1h}\ \bar{P}_{2hk}\bar{M}_{hk}|\ n_{1h}) = \left(\bar{P}_{1h} + var(\bar{P}_{1h})\right) var(\bar{P}_{2hk}\bar{M}_{hk})$$

ici $\bar{P}_{2hk}$ est la proportion de sous-domaine, qui remplace la proportion de la strate dans le domaine.
L'objet est ici d'estimer la variance de la moyenne $var(\bar{P}_{2hk}\bar{M}_{hk})$. Son estimation fait appel à un **conditionnement par les effectifs de deuxième phase** :

$$var(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h}) = E\{var(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})\} + var\{E(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})\} = V_{21} + V_{22}$$

Pour le premier terme, comme $\bar{P}_{2hk} = \frac{n_{2hk}}{n_{2h}}$

$$V_{21} = E\{var(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})\} = E\left\{\bar{P}_{2hk}{}^2\ var(\bar{M}_{hk}|\ n_{2h})\right\} = \bar{P}_{2hk}{}^2\ var(\bar{M}_{hk}|\ n_{2h})$$

Pour le deuxième terme la loi de la variance totale donne à nouveau deux termes que l'on traite successivement :

$$\begin{aligned} V_{22} &= var\{E(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})\} \\ &= E\{var(E(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})|\ n_{2h})\} + var\{E(E(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h})|\ n_{2h})\} \\ &= v_{221} + v_{222} \end{aligned}$$

$$v_{221} = E\{var(\bar{P}_{2hk}E(\bar{M}_{hk}|\ n_{2h}))|\ n_{2h}\} = E\{\bar{M}_{hk}^2\ var(\bar{P}_{2hk})\} = \bar{M}_{hk}^2\ var(\bar{P}_{2hk})$$

car $E(\bar{M}_{hk}|\ n_{2h})$ est une constante conditionnellement à n_{2h}, donc $var(E(\bar{M}_{hk}|\ n_{2h})) = 0$

$$\begin{aligned} v_{222} &= var\{E(\bar{P}_{2hk}E(\bar{M}_{hk}|\ n_{2h}))|\ n_{2h}\} = var\{E(\bar{P}_{2hk})E(\bar{M}_{hk}|\ n_{2h})\} \\ &= var(\bar{P}_{2hk})\ var(\bar{M}_{hk}|\ n_{2h}) \end{aligned}$$

En sommant les trois termes, V_{21}, V_{221} et V_{222}, et en factorisant on obtient :

$$\begin{aligned} var(\bar{P}_{2hk}\bar{M}_{hk}|\ n_{2h}) &= \bar{P}_{2hk}{}^2\ var(\bar{M}_{hk}|\ n_{2h}) + \bar{M}_{hk}{}^2\ var(\bar{P}_{2hk}) + var(\bar{P}_{2hk})\ var(\bar{M}_{hk}|\ n_{2h}) \\ &= \left(\bar{P}_{2hk}{}^2 + var(\bar{P}_{2hk})\right) var(\bar{M}_{hk}|\ n_{2h}) + \bar{M}_{hk}{}^2\ var(\bar{P}_{2hk}) \end{aligned}$$

En remplaçant $var(\bar{P}_{2hk})$ par $var(\bar{P}_{2hk}\bar{M}_{hk}|\, n_{2h})$ dans (3) et en adaptant les termes de covariance ($\bar{P}_{2hk}$ devient $\bar{P}_{2hk}\bar{M}_{hk}$) on a :

$$\begin{aligned} var\left(\sum_{h=1}^{H} P_{1h}\,\bar{P}_{2hk}\bar{M}_{hk}\right) \\ = \sum_{h=1}^{H}\left(P_{1h}^2 + var(P_{1h})\right)\left(\left(\bar{P}_{2hk}{}^2 + var(\bar{P}_{2hk})\right) var(\bar{M}_{hk}|\, n_{2h})\right. \\ \left. + \bar{M}_{hk}{}^2\, var(\bar{P}_{2hk})\right) \\ + \sum_{h=1}^{H}\sum_{l\neq h}^{H} \bar{P}_{2hk}\bar{M}_{hk}\,\bar{P}_{2lk}\bar{M}_{lk}\, cov(P_{1h}, P_{1l}) \end{aligned} \quad (4)$$

On peut réécrire (4) en trois termes additifs qui correspondent, dans l'ordre, à la variance d'hétérogénéité (de *y* dans le sous-domaine et dans la strate), de sous-domaine (variance de la proportion du sous-domaine dans la strate) et de stratification (variance d'estimation de la taille de la strate) qui sont élégamment détaillées dans la note de spécification (Hervé 2006, 2007) - à la différence qu'ici les formules s'appuient sur les proportions de première phase et non sur les surfaces, donc les trois termes doivent être multipliés par A_D^2 pour retrouver la valeur de la variance du total, ou, simplement en remplaçant P_{1h} par A_h) :

Terme 1, variance d'hétérogénéité

$$\sum_{h=1}^{H}\left(P_{1h}^2 + var(P_{1h})\right)\left(\left(\bar{P}_{2hk}{}^2 + var(\bar{P}_{2hk})\right) var(\bar{M}_{hk}|\, n_{2h})\right)$$

Terme 2, variance de sous-domaine

$$\sum_{h=1}^{H}\left(P_{1h}^2 + var(P_{1h})\right)\left(\bar{M}_{hk}{}^2\, var(\bar{P}_{2hk})\right)$$

Terme 3, variance de stratification

$$\sum_{h=1}^{H}\sum_{l\neq h}^{H} \bar{P}_{2hk}\bar{M}_{hk}\,\bar{P}_{2lk}\bar{M}_{lk}\, cov(P_{1h}, P_{1l})$$

Le calcul de ces termes nécessite d'estimer les différentes espérances et variance impliquées, à base des données mesurées ou estimées, ici présentées dans le tableau suivant (Tableau 2).

Lorsque le nombre de points de deuxième phase dans une strate est petit, l'estimation de la variance d'un total (Eq. 3 et 4) ne se fait plus à base de ces estimateurs, mais à base d'un calcul qui s'appuie sur l'hypothèse d'un coefficient de variation de 100%. Cette estimation alternative n'est pas décrite ici, devra faire l'objet d'une analyse sur son ampleur opérationnelle, ses effets et ses limites.

Tableau 2. Estimation des termes d'espérance et variance mis en œuvre par les estimateurs du total d'un paramètre y.

Terme	Estimateur	Equation
$var(P_{1h})$	$\widehat{var}(P_{1h}) = \frac{P_{1h}(1-P_{1h})}{n_h - 1}$	2
$var(\bar{P}_{2hk})$	$\widehat{var}(\bar{P}_{2hk}) = \frac{\bar{P}_{2hk}(1-\bar{P}_{2hk})}{n'_{2h} - 1}$	2, 3
$var(\bar{M}_{hk} \mid n_{2h})$	$\widehat{var}(\bar{M}_{hk} \mid n_{2h}) = \frac{S^2_{y,hk}}{n'_{2h}}$	3
	$S^2_{y,hk} = \left(1 - \frac{1}{n'_{2hk}}\right) \frac{\sum_{i=1}^{n_{2hk}} w_i \left(y_i - \hat{\bar{m}}_{y,hk}\right)^2}{\sum_{i=1}^{n_{2hk}} w_i}$ avec n_{2hk} taille de l'échantillon $(s_{2h} \cap k)$ de phase 2 et $\hat{\bar{m}}_{y,hk} = \frac{\sum_{i=1}^{n_{2hk}} w_i \, y_i}{\sum_{i=1}^{n_{2hk}} w_i}$	

6.4 Variance de la moyenne

La variance de la moyenne d'un attribut *y* sur un sous-domaine donné est calculée comme un ratio de moyennes. D'une manière générique, pour deux totaux X et Y, le ratio des moyennes est comme suit :

$$R = \frac{\tau_Y}{\tau_X} = \frac{A_D \bar{Y}}{A_D \bar{X}} = \frac{\bar{Y}}{\bar{X}}$$

Dans le cas du volume moyen (m^3 ha^{-1}), τ_Y est le volume total et τ_X est la surface du sous-domaine en question.

Il n'y a pas de solution analytique simple au calcul de la variance de R, mais à la différence de pratiquement tous les inventaires forestiers, ce n'est pas la linéarisation que Jean-Christophe Hervé a choisie pour approcher la variance du ratio.
Dans la linéarisation, le ratio empirique $\hat{R}$ sert à calculer une variable résiduelle sur chaque point, différence entre la valeur au point de l'attribut y (numérateur) et la prédiction de y basée sur le ratio. Cette résiduelle est ensuite traité comme une nouvelle variable aléatoire dont la variance, calculée comme celle d'une variable continue quelconque, sert d'estimation à la variance de $\hat{R}$:

$$\hat{R} = \frac{\bar{Y}}{\bar{X}} \rightarrow \hat{Y} = \hat{R}X \;\; et \;\; Y - \hat{Y} = Y - \hat{R}X$$

Ici au lieu de se baser sur la résiduelle issue de linéarisation $u_i = y_i - \hat{R}_k x_i$
c'est plutôt

$$u_i = y_i - \hat{R}_k \hat{\bar{X}} = y_i - \hat{\bar{m}}_{y,k}$$

qui est utilisé.

Les raisons de choix ne sont pas claires pour le moment, devront faire l'objet de plus de travaux, et la comparaison d'estimateurs alternatifs devrait apporter un éclairage sur les performances et propriétés de l'estimateur implémenté.

Les estimateurs de ratio ont donc une forme très proche de celle du total d'un attribut y quelconque, à la différence que :
- la moyenne de y dans h et k est remplacée par la résiduelle u_i calculée entre la valeur de y à chaque point de S_2, et la moyenne dans la catégorie et dans tout le domaine (pas dans les post-strates, dans D).
De fait cette résiduelle est calculée sur le même sous-ensemble de S_2 que la moyenne pondérée elle-même.
- ici les termes de variance ne sont pas multipliés par le carré de la surface, s'agissant d'un ratio et pas d'un total.

7 Discussion

Fondements principaux

L'IFN met en œuvre un échantillonnage annuel basé sur un réseau systématique spatialement, en deux phases, avec stratification à l'issue de la première phase, et post-stratification. Associer stratification (tirage d'unités d'échantillonnage à poids variables selon l'issue de la photo-interprétation) et post-stratification (groupement deux échantillons toujours selon la photo-interprétation) est très rare (et peut être même unique ?) dans une enquête.
La post-stratification est ici un outil de redressement sur la surface (méthode d'estimation au sens de la théorie des sondages). C'est a priori très efficace parce que l'échantillon de première phase, beaucoup plus volumineux, estime justement la surface des strates par proportions. Il y a à ce sujet un vrai consensus (ex. Saborowski et Cancino 2007, Pulkinnen et al. 2018, Westfall et al. 2019). L'échantillon de deuxième phase fournit les valeurs qui ne peuvent s'estimer qu'à base de mesures de terrain.
Les estimateurs de variance diffèrent clairement de ceux généralement utilisés pour un inventaire à deux phases pour stratification. La raison principale tient au fait que la moyenne d'un attribut est calculée exclusivement sur la fraction non nulle de l'échantillon, formant une sorte de double post-stratification. La conséquence n'est pas tant visible sur les estimateurs des totaux que sur ceux des variances. Il est pratiquement impossible d'en mesurer l'importance pratique sans tests spécifiques, mais l'avantage qu'apporte une variable mieux distribuée (car n'ayant pas une majorité de valeurs nulles donc probablement plus « normale ») est probablement souvent contrebalancé par des effectifs beaucoup plus faibles pour les estimations, notamment dans les strates.
L'obligation de manipuler des points de deuxième phase à poids inégaux fut une contrainte relativement beaucoup moins importante. S'agissant des totaux, on peut vérifier qu'on obtient les formules les plus classiques (ex. Cochran 1977) en posant $w_i=1$.

Aspect spatialement systématique des tirages, principe de dualité

a) **Tirage aléatoire systématique**
Dans l'examen des hypothèses et difficultés d'estimation des variances liées à l'échantillonnage, il faut découpler les problèmes liés à un tirage systématique de ceux liés à un tirage spatialement systématique.
La grille est le support d'un échantillonnage bidimensionnel systématique (Dunn et Harrison 1993) qui fait intervenir les problèmes dits de périodicité des variations (Cochran 1977). Mais les performances de l'échantillonnage systématique semblent dépendre de la taille des mailles en rapport avec les paramètres étudiés (Dunn et Harrison 1993).
Malgré les reproches que l'on peut émettre, en particulier des difficultés de calculer une variance non biaisée dans le cas d'un échantillon systématique, ici trois avantages très forts militent en faveur de cette méthode :
- la couverture spatiale plus efficace, en particulier l'absence de trous qu'un tirage vraiment aléatoire génèrerait presque sûrement
- tous les points ont le même poids, c'est extrêmement utile dans la formation des estimateurs
- la distance entre points (plutôt grande) réduit très fortement les risques de corrélation spatiale
Les aspects liés au caractère systématique ne sont donc pas a priori problématiques (avis apparemment partagé par Fattorini et al. 2009).
- une partition du domaine spatial permettant d'extraire, dans l'ensemble des sous-ensembles de l'ensemble des entités de surface forestière, un sous-ensemble d'unités d'échantillonnage pondérées autorisant la sommation, c'est-à-dire l'estimation de la surface forestière.

b) **Principe de dualité, ou comment passer d'une population continue à une population discrète**.
Cette dualité (ou équivalence) est très peu discutée dans la littérature (Mandallaz 1991, 2013). Elle est décrite comme une forme tirage Monte Carlo brut (*Crude Monte Carlo*). Au fond, elle suppose que la densité moyenne peut être approchée par la moyenne empirique estimée sur un échantillon. Elle suppose aussi une indépendance des points utilisés pour le calcul de la moyenne, ce qui n'est pas le cas compte tenu du caractère spatialement systématique. Aussi la randomisation spatiale contribue-t-elle un peu à rendre un caractère plus indépendant.
Quoi qu'il en soit, il ne faut pas conclure qu'il s'agit d'une approche basée sur les modèles des estimateurs, ceux-ci restant bien basés sur l'échantillonnage, donc sur l'aspect aléatoire du tirage des points.
On notera bien qu'on a fait l'hypothèse d'une indépendance des points, mais pas seulement : l'estimation de la variance d'un attribut arbre ou peuplement *y* repose sur l'hypothèse que tous les points ont une même variance $S^2_{y,hk}$ sur tout le domaine, dans la strate. Ce n'est pas si évident compte tenu des regroupements des catégories de couverture au sol (UCS).

Les questions les plus difficiles vis-à-vis des méthodes déployées sont plutôt :

a) **la censure des 9/10 de la surface du territoire lié à l'utilisation d'un découpage en années de la grille**.
L'objectif originel était de s'assurer que tout le territoire était parcouru à l'issue de 10 années. Chaque année prend donc une fraction, à l'image des cycles d'inventaires pour les inventaires périodiques. Mais ici les estimations et les erreurs d'échantillonnage à 5 ans sont calculées comme des moyennes arithmétiques, donc faisant complètement abstraction de l'augmentation du taux de couverture spatiale. Mais on ne pouvait pas tenir compte de cette augmentation si on n'avait pas aussi négligé en même temps les aspects des tirages liés aux mailles et aux points. On arrive donc à une contradiction : on néglige l'organisation spatiale annuelle et on crée une dépendance temporelle des échantillons.
Une réflexion sur la construction des échantillons annuels et la combinaison de plusieurs échantillons s'impose.

b) **le conditionnement fait l'objet de questions dans la littérature**. Si quelques auteurs préconisent l'utilisation du conditionnement (Gregoire et al. 2016, Westfall et al. 2019) mais certains détracteurs (ex. Scott et al. 2005; Köhl et al. 2006) ont déjà fait valoir que l'hypothèse implicite d'équilibre de distribution des effectifs de points entre strates que suppose le conditionnement n'était pas toujours satisfaite. Dans l'inventaire français, compte tenu du double conditionnement et de la stratification très poussée, ces questions restent ouvertes : impossible de se positionner sans une analyse approfondie.

Une grosse hypothèse ? (simplification implicite d'un terme de variance)

Le développement de l'estimateur de la moyenne dans une strate fait apparaître un terme présentant la variance de l'espérance d'un produit (le terme v222 dans l'équation 4). Pour converger sur les estimateurs mis en œuvre, il faut passer de la variance de l'espérance d'un produit à un produit de variances :
$v_{222} = var\{E(\bar{F}_{hk}E(\bar{M}_{hk}|\, n_{2h}))|\, n_{2h}\} = var(\bar{F}_{hk})\, var(\bar{M}_{hk}|\, n_{2h})$

Mais la variance d'un produit devrait donner plus de termes :

$$var(XY) = var(X)\ var(Y) + \ E(X)^2 var(Y) + E(Y)^2 var(X) + 2E(X)E(Y)\ cov(X,Y)$$

Ici les estimateurs déployés ne gardent que le premier terme, le produit des variances. L'importance relative des termes négligés ici est une bonne question. Si l'on part du point de vue que X et Y sont tout à fait indépendantes, alors la covariance est nulle (probablement le choix implicite), et donc il reste un produit de carré d'espérance et de variance. Si au contraire on suppose qu'il existe une covariance, celle-ci pourrait être négative et compenser partiellement les autres termes. **Cela peut avoir une incidence sur les estimations de variance dans l'inventaire, qu'il faut explorer.**

Remerciements

- A Jean-Daniel Bontemps, Nicolas Paparoditis et la DRE de l'IGN, pour leur confiance, et le recrutement d'une compétence en inventaire et sondage statistique, à un moment où cette dernière n'avait désormais plus aucune assise à l'inventaire.

- Guillaume Chauvet (ENSAI) puis Philippe Brion (IRMAR), qui ont accepté d'entrer dans une phase d'échanges approfondis puis de collaboration sur une enquête à caractère singulier, avec des approches originales mais jusqu'ici très mal documentées.

- A Jean-Christophe Hervé, pour avoir accepté de partager, toujours exclusivement dans une forme de tradition orale qui était sa façon de dispenser les connaissances, quelques fragments de raisonnement sur ces estimateurs qui m'ont servi de points d'ancrage dans cette analyse

8 Bibliographie

8.1 Documents à disposition présentant les estimateurs

Hervé JC (2005) Estimateurs du sondage systématique. 23 mai 2005.
Hervé JC (2006) Variance des estimateurs trois fois post-stratifiés. 10 janvier 2006.
Hervé JC (2007) Variance des estimateurs stratifiés. 5 juin 2007.
Pesty B (2014) Projet : Service de calcul. Architecture Technique. Version : 3.1.0. 21 Janvier 2014.
Wolsack J (2002) Estimateurs statistiques de l'inventaire général. 23 avril 2002.

IFN (collectif). Dossier de présentation au Comité du label de la statistique publique. Mai 2017.
IFN (collectif). Réponse aux remarques du pré-label du 3 mai 2017.

8.2 Littérature de spécialité

Cochran W. (1977). Sampling Techniques. Wiley

Dunn, R., & Harrison, A. R. (1993). Two-dimensional systematic sampling of land use. Journal of the Royal Statistical Society: Series C (Applied Statistics), 42(4), 585-601.

Fattorini, L., Franceschi, S., & Pisani, C. (2009). A two-phase sampling strategy for large-scale forest carbon budgets. Journal of statistical planning and inference, *139*(3), 1045-1055.

Gregoire, T. G., Ringvall, A. H., Ståhl, G., & Næsset, E. (2016). Conditioning post-stratified inference following two-stage, equal-probability sampling. Environmental and ecological statistics, 23(1), 141-154.

Köhl, M., Magnussen, S. S., & Marchetti, M. (2006). Sampling methods, remote sensing and GIS multiresource forest inventory. Springer Science & Business Media.

Little, R.J. (1993). Post-stratification: a modeler's perspective. Journal of the American Statistical Association, 88(423), 1001-1012.

Mandallaz, D. (1991). A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models (Thèse de doctorat, ETH Zurich).

Mandallaz, D. (2007). Sampling techniques for forest inventories. CRC Press.

Mandallaz, D. (2013). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. Canadian Journal of Forest Research, 43(5), 441-449.

Pulkkinen, M., Ginzler, C., Traub, B., & Lanz, A. (2018). Stereo-imagery-based post-stratification by regression-tree modelling in Swiss National Forest Inventory. Remote Sensing of Environment, 213, 182-194.

Saborowski J, Cancino J (2007) About the benefits of poststratification in forest inventories. J For Sci 53(4):139–148

Tam, S.M. (1985). On covariance in finite population sampling. The Statistician, 429-433.Westfall, J. A., Lister, A. J., Scott, C. T., & Weber, T. A. (2019). Double sampling for post-stratification in forest inventory. European Journal of Forest Research, 138(3), 375-382.

Valentine, H. T., Affleck, D.L., & Gregoire, T.G. (2009). Systematic sampling of discrete and continuous populations: sample selection and the choice of estimator. Canadian Journal of Forest Research, 39(6), 1061-1068.