

Avis technique

SSRF-34

Direction de la recherche forestière

Titre :	Stratégie d'échantillonnage lors du suivi d'efficacité des procédés de régénération en forêt
Auteurs :	François Guillemette, ing.f., M. Sc. et Gaétan Daigle, M. Sc., P. Stat. (Service de consultation statistique, Université Laval)
Collaborateurs :	Charles Vigeant-Langlois, ing.f., M. Sc. (Unité de gestion 03-12 — Capitale-Nationale et Chaudière-Appalaches) et Geneviève Picher, B. Sc. stat. (DRF)
Date :	Juin 2024

Des ingénieurs forestiers et ingénieures forestières chargés de suivre les effets des procédés de régénération sur l'établissement de celle-ci souhaitent savoir si l'on peut optimiser l'échantillonnage pour effectuer une meilleure reddition de comptes. Pour trouver un meilleur équilibre entre les ressources disponibles et répondre adéquatement au besoin, nous avons relevé des principes d'échantillonnage, montré des exemples et proposé une façon d'analyser les données.

1. Introduction

Le ministère des Ressources naturelles et des Forêts est responsable du suivi et du contrôle des interventions en forêt publique au Québec (MFFP 2021). Entre autres, il réalise des suivis d'efficacité des traitements sylvicoles. Ces suivis visent notamment à vérifier si la régénération s'installe comme prévu à la suite d'un procédé de régénération. Ils servent aussi à réaliser une reddition de comptes pour vérifier si toutes les superficies traitées sont régénérées, à détecter certains problèmes, puis à prescrire les traitements sylvicoles nécessaires pour l'atteinte des objectifs sylvicoles et des stratégies d'aménagement préétablies.

C'est dans ce contexte que la Direction de la recherche forestière a été approchée par des aménagistes forestiers du Secteur des opérations régionales du ministère des Ressources naturelles et des Forêts responsables de réaliser les suivis d'efficacité des procédés de régénération. La pratique habituelle consiste à recueillir un échantillon (c'est-à-dire à prendre des mesures dans une grappe de 10 microplacettes dont la taille peut varier de 4 à 25 m² selon le cas) par 3 à 4 ha de superficie traitée. Conséquemment, les populations les plus grandes sont davantage échantillonnées que les plus petites.

Ce rapport vise à répondre aux 5 questions suivantes :

- 1) Est-il acceptable d'appliquer la compilation des échantillons pris dans une population (strate, strate regroupée) à tous les peuplements (polygones sur la carte) de cette population, même à ceux non échantillonnés ? En effet, certains secteurs devenus inaccessibles et certains petits peuplements dispersés sur un grand territoire peuvent être exclus du sondage.
- 2) Si oui, y a-t-il un seuil minimal de proportion de la surface à échantillonner pour que l'on puisse inférer le résultat du sondage à toute la population ?
- 3) Pour obtenir la précision recherchée, peut-on regrouper les populations sous-échantillonnées qui ont été traitées et sondées lors d'années différentes ?
- 4) Sinon, quel nombre minimal d'échantillons faut-il prendre dans ces populations chaque année ?
- 5) À partir de quel seuil peut-on réduire le nombre d'échantillons à prendre dans les populations suréchantillonnées sans compromettre l'atteinte de l'objectif du sondage, afin de rediriger les ressources libérées vers d'autres populations sous-échantillonnées ?

2. Principes de l'échantillonnage

Dans la pratique forestière, on cherche souvent à savoir si tous les peuplements écoforestiers traités dans une unité d'aménagement (UA) sont régénérés à un moment donné (5 ou 10 ans après une intervention) selon les objectifs fixés (tableau 1). Toutefois, le manque de ressources empêche souvent de recenser tous les peuplements traités dans une UA. C'est alors que le recours à l'échantillonnage prend tout son sens (Gregoire et Valentine 2007, p. 1). L'échantillonnage consiste d'abord à établir la base du sondage, soit la liste de toutes les unités de la population cible. Ensuite, certaines unités sont tirées au hasard dans cette liste, selon une certaine règle probabiliste communément appelée plan de sondage. Ces résultats obtenus dans les unités échantillonnées sont ensuite pondérés par leur poids de sondage respectif afin d'inférer plusieurs paramètres de la population. Le poids attribué à une unité correspond simplement à l'inverse de sa probabilité de sélection. À noter que la marge d'erreur autour de l'estimation d'un paramètre est fonction de la taille de l'échantillon et de la variabilité entre les unités sondées. Plus précisément, la marge d'erreur diminue lorsque la taille de l'échantillon augmente, mais elle augmente aussi avec la variabilité (variance, s^2) entre les unités sondées.

Tableau 1. Quelques exemples de seuils dérivés de critères forestiers (coefficients de distribution [CD] recherchés après une période donnée) selon la famille de procédés de régénération et le gradient d'intensité de la sylviculture applicable à une population donnée.

Famille de procédés de régénération	Intensité de la sylviculture	Âge limite pour l'atteinte du seuil de régénération	Seuil visé*
Coupes de régénération	Extensive	10 ans après la récolte	CD $\geq$ 50 % en essences commerciales
Coupes de régénération	De base	5 ans après la récolte	CD $\geq$ 60 % en essences désirées
Coupes de régénération	Intensive ou d'élite	5 ans après la récolte	CD $\geq$ 75 % en essences désirées
Coupes partielles	De base	5 ans après la dernière intervention de régénération	CD $\geq$ 60 % en essences désirées
Coupes partielles	Intensive ou d'élite	5 ans après la dernière intervention de régénération	CD $\geq$ 75 % en essences désirées

* La taille des placettes peut varier. Dans cet avis, nous supposons qu'elle est de 6,25 m², ce qui correspond à un cercle de 1,41 m de rayon.

Réponse à la question 1 : oui, il est possible d'inférer les résultats provenant d'un échantillon de peuplements à tous les peuplements de la population, pourvu que l'échantillon soit représentatif de cette population. C'est l'essence même de l'échantillonnage. Pour ce faire, il faut toutefois que tous les peuplements de la population aient eu une chance d'être échantillonnés. Or, il arrive parfois en pratique que certains peuplements ne puissent pas être échantillonnés parce qu'ils sont soit trop petits, soit inaccessibles. En présence de ce phénomène marginal, il est courant d'inférer tout de même les résultats de l'échantillon à l'ensemble de la population. Ce processus n'est valable que si les peuplements exclus de l'échantillonnage n'ont pas de caractéristiques particulières et qu'ils sont semblables aux autres. En pareille circonstance, nous recommandons de dévoiler cet écart et de prendre les mesures nécessaires pour le minimiser.

Réponse à la question 3 : oui, il est possible de regrouper des populations traitées et sondées lors d'années différentes. Tout dépend de la définition initiale de la population. En effet, nous pouvons très bien définir la population comme l'ensemble des peuplements traités au cours des années 2021 et 2022. Ainsi, les résultats des échantillons pris en 2021 et en 2022 peuvent être utilisés pour faire des inférences au sujet de cette population. À la section 5, nous proposons une méthode qui mériterait d'être explorée afin d'utiliser l'information provenant d'années différentes, tout en tirant profit des variations interannuelles.

Malgré ce qui précède, il faut être attentif et éviter de regrouper des années de traitements qui pourraient avoir été affectées par des conditions climatiques particulières. Par exemple, il pourrait être hasardeux de regrouper une année de précipitations anormalement abondantes avec une autre de sécheresse exceptionnelle. De même, il est plus risqué de regrouper deux années consécutives de strates destinées à une production de bouleau jaune (*Betula alleghaniensis* Britton), car il est reconnu que chez cette essence, il y a alternance entre les bonnes et les mauvaises années semencières.

3. Inférence des résultats de l'échantillonnage

Réponse à la question 2 : Oui, on peut très bien inférer les résultats d'un échantillonnage probabiliste à la population cible, dès que la taille d'échantillon est plus grande ou égale à 2 ($n \geq 2$). Toutefois, la précision de l'estimation dépend de la taille de l'échantillon. Celle-ci doit généralement être fixée en fonction de la marge d'erreur désirée autour de l'estimation du paramètre d'intérêt de la population, et de la variabilité attendue entre les unités de sondage. Le calcul de la taille d'échantillon est abordé à la section 4.

Nous avons identifié 3 aspects à considérer pour déterminer l'échantillonnage requis. Ces 3 aspects sont décrits dans les sous-sections suivantes.

3.1. Marge d'erreur désirée

Le premier aspect à considérer est la marge d'erreur d désirée autour d'une estimation. Par exemple, dans le cas d'une variable qualitative, supposons que nous cherchions une certaine marge d'erreur d sur l'estimation $\hat{p}$ de la proportion P de peuplements bien régénérés :

$$P \in \hat{p} \pm d \quad (\text{équation 1})$$

Au niveau de confiance de $1 - \alpha$, la marge d'erreur d est alors liée à la variance de $\hat{p}$ comme suit :

$$d = z_{\alpha/2} \sqrt{\text{var}(\hat{p})} = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (\text{équation 2})$$

Comme le produit $\hat{p}(1 - \hat{p})$ est maximal à $\hat{p} = 50 \%$ (figure 1), c'est dans cette situation que la taille d'échantillon nécessaire pour atteindre la marge d'erreur souhaitée est la plus grande. En fait, la taille d'échantillon requise diminue à mesure que la valeur de $\hat{p}$ approche de 0 % ou de 100 %. Ainsi, il faut moins d'échantillons pour obtenir une information précise si la régénération est très bien répartie ou si elle est rare. En revanche, il faudra plus d'échantillons si le CD de la régénération est d'environ 50 % dans une population. Généralement, on recommande de viser une marge d'erreur (d) de 10 % pour une estimation (p. ex., Méthot *et al.* 2014).

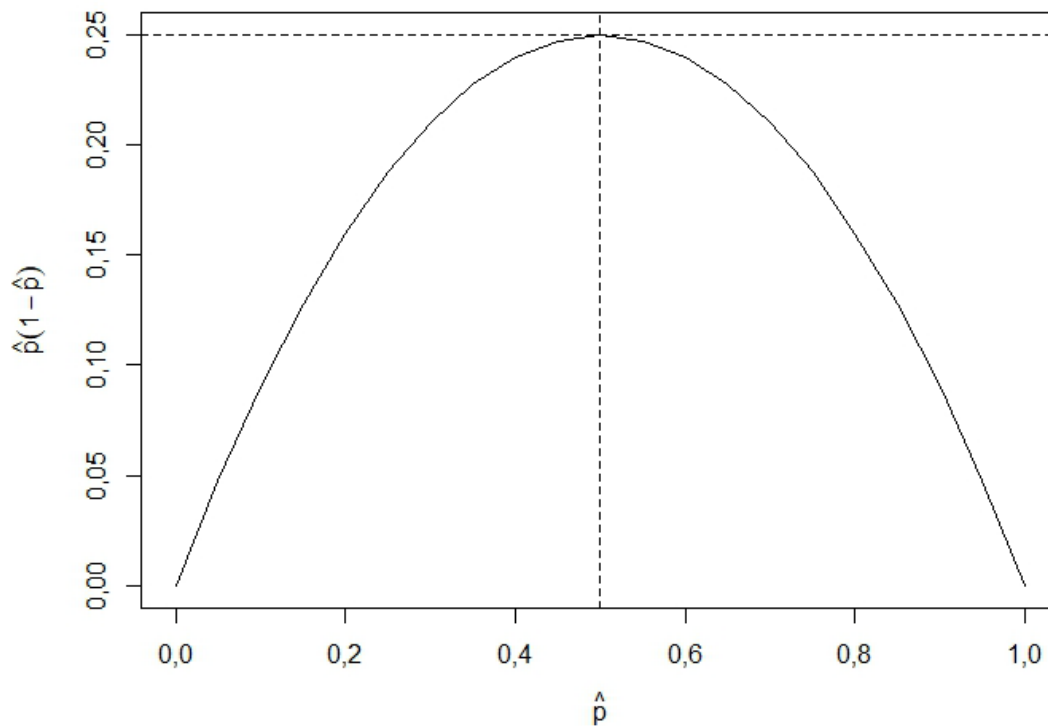


Figure 1. Relation entre le produit $\hat{p}(1 - \hat{p})$ et la valeur de $\hat{p}$ elle-même.

3.2. Écart entre les mesures et le seuil dérivé du critère forestier

Le second aspect à considérer est l'ampleur de l'écart entre les mesures (X) et le seuil dérivé du critère forestier (c), comme l'atteinte d'un certain CD (tableau 1). Par exemple, il pourrait être souhaitable d'estimer la probabilité que X soit supérieur (ou inférieur) à ce critère. Prenons le cas d'une variable aléatoire X distribuée selon une loi normale, dont la moyenne est $\mu = 0,80$ et l'écart-type est $s = 0,2$, puis considérons que nous souhaitons ainsi estimer la probabilité que X soit plus grand que le seuil forestier fixé à $c = 0,75$. En termes mathématiques, cela revient à calculer la probabilité suivante :

$$P(X > c) = P\left(\frac{X - \mu}{s} > \frac{c - \mu}{s}\right) \quad (\text{équation 3})$$

Nous pouvons estimer cette probabilité comme suit :

$$P(\widehat{X} > c) = 1 - \Phi\left(\frac{c - \bar{x}}{s}\right) = 1 - \Phi\left(\frac{0,75 - 0,80}{0,2}\right) = 1 - \Phi(-0,25) = 0,5987 \quad (\text{équation 4})$$

où $\Phi()$ est la fonction de répartition d'une loi normale centrée réduite. Ainsi, nous pouvons nous attendre à ce qu'en moyenne, 60 % des mesures de l'échantillon soient plus grandes que le seuil dérivé du critère forestier c .

3.3. Habitudes professionnelles

Finalement, dans le cas des inventaires d'intervention ou de suivi des opérations, nous observons que la communauté forestière au Québec est habituée à échantillonner une placette ou une grappe de microplacettes par 4 ha de forêt environ. À notre avis, cette manière de procéder découle plus de considérations budgétaires qu'elle ne s'appuie sur un fondement statistique. En effet, il est tout à fait possible d'atteindre la marge d'erreur désirée en échantillonnant en forêt des placettes espacées de plus ou de moins de 4 ha. Cela dépend non pas de la taille de la population dans ce contexte forestier, mais bien de la variabilité du phénomène à l'étude et de la marge d'erreur désirée (Blais 2014).

4. Taille d'échantillon

Dans le cas d'une variable binomiale dont le CD (c'est-à-dire p) est connu, on peut calculer la taille de l'échantillon (n^*) requis pour obtenir une estimation à l'intérieur d'une marge d'erreur donnée (d) au niveau de confiance $1 - \alpha$ (où en général, $\alpha = 5\%$) à partir de l'équation 5 ci-dessous (Cochran 1977) :

$$n^* = \left(\frac{z_{\alpha/2}}{d} \right)^2 p(1 - p) \quad (\text{équation 5})$$

Dans les cas où p n'est pas connu, on obtient une estimation conservatrice de la taille d'échantillon en fixant $p = 0,5$ dans l'équation précédente. On obtient alors que $n^* = z_{\alpha/2}^2 / (4d^2)$. Ce calcul n'est valable que dans le contexte d'un plan aléatoire simple.

En foresterie, la coutume est d'utiliser un plan d'échantillonnage par grappes avec $m = 10$ microplacettes par grappe. Avec ce type de plan, il faut appliquer une correction pour l'effet de plan. Ainsi, le nombre de grappes (n) à échantillonner s'obtient à partir de l'équation 6 :

$$n = \frac{n^*[1+(m-1)r]}{m} \quad (\text{équation 6})$$

où r est la corrélation intraclasse, c'est-à-dire la corrélation entre les microplacettes à l'intérieur d'une même grappe. La quantité $\frac{[1+(m-1)r]}{m}$ est le facteur de correction de l'effet de ce plan. Lorsque $r = 1$, on a alors $n = n^*$. À l'opposé, lorsque $r = 0$, nous avons alors $n = n^*/m$.

En combinant les équations 5 et 6, on obtient un lien direct entre le nombre de grappes (n) à échantillonner et la précision souhaitée (d) :

$$n = \left(\frac{z_{\alpha/2}}{d} \right)^2 p(1 - p) \left[\frac{1+(m-1)r}{m} \right] \quad (\text{équation 7})$$

En guise d'exemple, prenons le cas où l'on estime la proportion de grappes régénérées dans une population avec une marge d'erreur de $d = 10\%$ au niveau de confiance de 95% ($\alpha = 5\%$), sans connaître p . Une estimation conservatrice du nombre de grappes s'obtient alors en fixant $p = 0,5$ dans l'équation 7. Conformément à l'hypothèse selon laquelle la corrélation intraclasse est de $r = 0,2$ ¹, et comme $z_{\alpha/2} = 1,96$ au seuil $\alpha = 5\%$, on déduit que le nombre de grappes (n) nécessaires pour atteindre la marge d'erreur souhaitée est d'environ 27.

$$n = \left(\frac{1,96}{0,10}\right)^2 0,5(1 - 0,5) \left[\frac{1+(10-1)\times 0,2}{10}\right] \approx 27 \quad (\text{équation 8})$$

Ce nombre de grappes est assez conservateur du fait que p n'est pas connu et qu'il est fixé à son seuil maximal d'après la figure 1. De plus, nous avons utilisé un coefficient de corrélation intraclasse relativement élevé pour ce type de données (voir la note 1). Toutefois, en reprenant le même calcul, mais en fixant $p = 0,80$ ou sinon $r = 0,09$ dans l'équation 7, on obtient alors un nombre de grappes (n) d'environ 17 dans les deux cas.

Réponse aux questions 4 et 5 : il faudrait donc échantillonner au moins 15 à 20 grappes dans les populations ayant un CD de près de 80% (ou, à l'opposé, dans celles ayant un CD de près de 20%); ce nombre augmente jusqu'à 25 à 30 grappes si le CD est près de 50% et si la corrélation intraclasse est relativement grande, soit près de $0,2$ avec ce type de données. Cependant, chaque cas est unique et peut être ajusté selon le CD obtenu (p) et selon la force de la corrélation intraclasse r .

5. Autre approche proposée

À notre avis, l'analyse des données de sondage sur une base annuelle dans des populations semblables ne permet pas d'optimiser l'utilisation des ressources. Nous proposons une façon de valoriser l'accumulation annuelle de l'information, puis d'utiliser ce suivi pour optimiser dans le temps l'allocation des ressources consacrées à l'échantillonnage. Cette proposition s'inspire de la maîtrise statistique des procédés (*statistical process control*), une méthode qui a fait ses preuves dans de nombreux processus de fabrication industrielle (Oakland et Oakland 2019).

Dans le contexte de cet avis technique, cette méthode consiste à utiliser les données recueillies dans les premières années avec une précision acceptable pour en dresser un état de référence (figure 2). Cet état présente la valeur moyenne et une mesure de dispersion (soit l'intervalle de confiance à 95% sur la moyenne, dans ce cas-ci). Au cours des années suivantes, il suffit de vérifier comment la moyenne annuelle évolue par rapport à cet intervalle de confiance. Tant qu'elle demeure à l'intérieur de cet intervalle et qu'elle reste nettement supérieure au seuil dérivé du critère forestier, le sondeur ou la sondeuse peut avoir confiance que les nouveaux peuplements traités dans cette population sont encore bien régénérés, comme

¹ Avec un jeu de données réelles de 45 strates contenant de 10 à 48 grappes de 10 microplacettes dont le coefficient de distribution moyen varie de 29% à 91% , nous avons calculé des valeurs de r allant de moins de $0,00$ jusqu'à $0,45$, mais une seule strate avait un $r > 0,19$.

par le passé (scénario 1 de la figure 2). Pour respecter la marge d'erreur recherchée, il n'est donc pas nécessaire d'échantillonner autant de grappes annuellement que lors des premières années.

Lorsque ces conditions sont respectées dans une population pendant au moins 2 années d'échantillonnage consécutives, nous proposons de faire passer de 10 à 20 % la marge d'erreur tolérée. En doublant cette marge d'erreur, le nombre de grappes requises est largement diminué. En fait, toujours en supposant que la corrélation intragrappe est de $r = 0,2$ et que le CD (c'est-à-dire p) est inconnu, l'équation 7 indique qu'il suffirait d'échantillonner 7 grappes au lieu de 27.

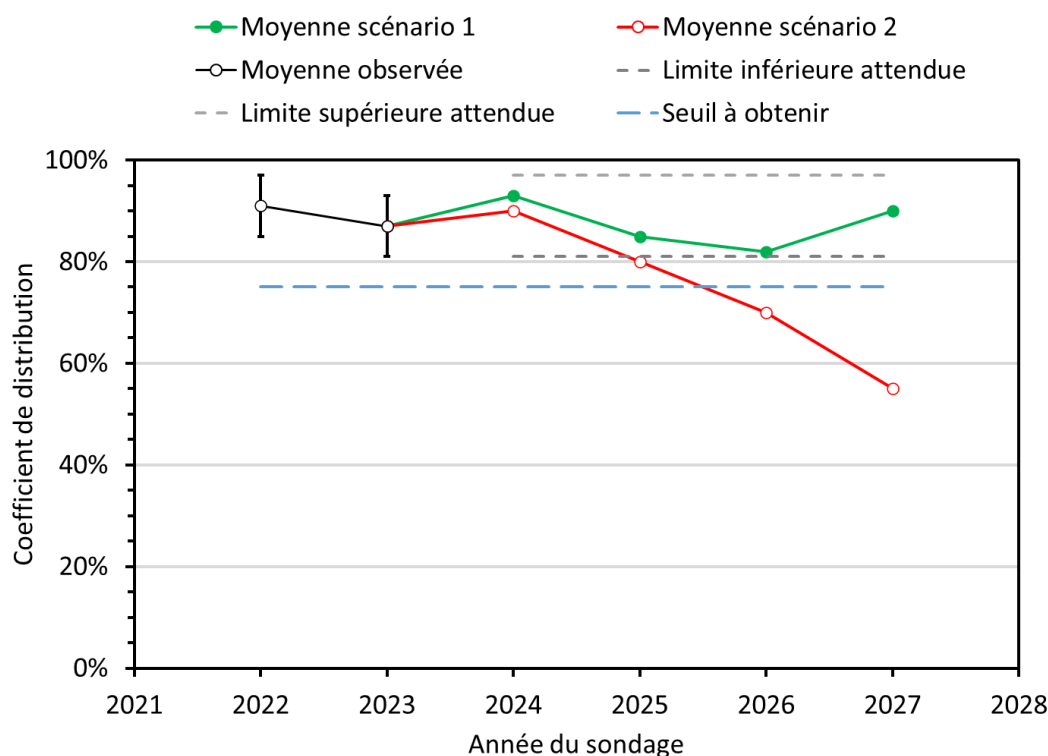


Figure 2. Exemple de 2 scénarios possibles d'évolution d'un coefficient de distribution. Les limites inférieures et supérieures des intervalles de confiance à 95 % (barres d'erreur) des sondages de 2022 et 2023 sont utilisées pour définir les limites attendues pour les années à venir (lignes tiretées grises). Le scénario 1 (ligne verte) montre de nouveaux sondages ayant une moyenne à l'intérieur de ces limites, ce qui indique que le peuplement évolue conformément aux attentes. Le scénario 2 montre l'apparition graduelle d'un problème de régénération : en 2025, le coefficient de distribution sort des limites de l'intervalle de confiance, ce qui peut soulever une inquiétude. En 2026, le coefficient de distribution passe sous le seuil dérivé du critère forestier (ligne tiretée bleue), ce qui indique que le problème devient plus sérieux.

Cette approche permettrait de libérer des ressources en vue d'augmenter l'échantillonnage dans des populations qui en ont davantage besoin, soit celles dont les moyennes sont historiquement près du seuil dérivé du critère forestier, ou celles ayant une faible superficie. L'idée générale est d'attribuer moins de ressources aux populations les plus prévisibles, tout en assurant une surveillance permettant de réagir si un changement apparaît dans la dynamique d'une population (scénario 2 de la figure 2).

Advenant que la moyenne d'une année sorte de l'intervalle de confiance attendu et menace de descendre sous le seuil à obtenir (p. ex., en 2024 pour le scénario 2 de la figure 2), alors il faudrait d'abord y ajouter des échantillons (reprenre une marge d'erreur tolérée de 10 % au lieu de 20 %) pour vérifier si l'écart est dû au hasard de l'échantillonnage. Si cet échantillonnage supplémentaire confirme la tendance, alors il faudrait investiguer pour voir si des changements ont été apportés dans la pratique et si des correctifs peuvent être apportés. Dans la négative, il faudrait définir un nouvel intervalle de confiance ou revoir le seuil dérivé du critère forestier, si ces changements sont structurels (comme ceux attribuables aux changements climatiques, par exemple).

6. Conclusion

Il n'est pas réaliste de mesurer l'entièreté de la forêt aménagée pour en assurer le suivi. Des stratégies d'échantillonnage peuvent être mises au point pour trouver un point d'équilibre entre les ressources disponibles et le degré de confiance requis. Nous avons résumé quelques aspects importants à considérer dans le contexte d'un suivi de la régénération forestière.

7. Références bibliographiques

- Blais, L., 2014. Simplifions la statistique — Fiche 1. De quoi la taille d'un échantillon dépend-elle ? Gouvernement du Québec, ministère des Forêts, de la Faune et des Parcs, Direction de l'aménagement et de l'environnement forestiers, 4 p. <https://mffp.gouv.qc.ca/nos-publications/simplifions-statistique-taille-echantillon/>
- Cochran, W. G. 1977. Sampling techniques, 3rd ed. John Wiley & Sons, New York. 428 p.
- Gregoire, T. G. et H.T. Valentine, 2007. Sampling strategies for natural resources and the environment (1st ed.). Chapman & Hall/CRC. New York. 496 p. <https://doi.org/10.1201/9780203498880>
- Méthot, S., L. Blais, J. Gravel, I. Latrémouille, S. St-Pierre et S. Vézeau, 2014. Guide d'inventaire et d'échantillonnage en milieu forestier. Gouvernement du Québec, ministère des Ressources naturelles, Direction de l'aménagement et de l'environnement forestiers, 237 p. <https://mffp.gouv.qc.ca/documents/forets/connaissances/Guide-Inventaire-Echantillonnage.pdf>

[MFFP] Ministère des Forêts, de la Faune et des Parcs, 2021. Suivis forestiers des traitements sylvicoles réalisés au Québec et rentabilité des investissements. Gouvernement du Québec, ministère des Forêts, de la Faune et des Parcs, 14 p.

<https://diffusion.mern.gouv.qc.ca/public/Biblio/Mono/2023/04/1288207.pdf>

Oakland, J. et J.S. Oakland, 2019. Statistical process control, 7th edition. Routledge, New York.
xvi + 429 p.

François Guillemette, ing.f., M. Sc.

Service de la sylviculture et du rendement des forêts

Gaétan Daigle, M. Sc., P. Stat.

Service de consultation statistique, Université Laval

Correspondance :

François Guillemette

Ministère des Ressources naturelles et des Forêts

Direction de la recherche forestière

2700, rue Einstein, bureau C.1.345.11

Québec (Québec) G1P 3W8

Tél. : 418 643-7994 poste 706629

Courriel : françois.guillemette@mrnf.gouv.qc.ca

On peut citer tout ou partie de ce texte en indiquant la référence. Citation recommandée :

Guillemette, F. et G. Daigle, 2024. *Stratégie d'échantillonnage lors du suivi d'efficacité des procédés de régénération en forêt*. Gouvernement du Québec, ministère des Ressources naturelles et des Forêts, Direction de la recherche forestière. Avis technique SSRF-34. 10 p.