Les stratégies d'échantillonnage

Ouvrages de référence:

Frontier, S. 1983. *Stratégies d'échantillonnage en écologie*. Masson, Paris. 494 pp.

Scherrer, B. 1984. *Biostatistique*. Gaëtan Morin Editeur, Boucherville. 850 pp.

1. QU'EST CE QU'UN ECHANTILLONNAGE?

1.1. Définition

En forcant quelque peu le trait, deux démarches se complètent en science:

- l'une purement descriptive qui vise à donner une image la plus fidèle et la plus juste possible d'un système biologique naturel, accepté dans sa complexité, donc à mesurer la valeur d'un certain nombre de variables, et à mettre en évidence les corrélations qu'on observe entre elles.

- l'autre qualifiée d'expérimentale, qui crée des conditions artificielles contrôlées telles que l'effet d'une variable unique soit mesuré, pour pouvoir réfuter (et à défaut de réfutation confirmer...) son action sur un phénomène biologique.

Dans les deux cas, ces démarches ont en commun l'élaboration préalable d'une hypothèse à tester: hypothèse sur une structure ou une dynamique biologique dans le premier cas, hypothèse sur une causalité biologique dans le second. Dans le premier cas, on mettra en oeuvre un plan d'échantillonnage, dans le second un plan expérimental.

Gardons en mémoire tout au long de cet exposé que la démarche d'échantillonnage, n'a de sens qu'en fonction de cette hypothèse préalablement exprimée (par écrit SVP, ce qui oblige à la rigueur!). On ne mesure donc pas “tout” un système biologique (ce ne serait plus un échantillonnage, et ce serait, de plus, techniquement et conceptuellement impossible), mais bien *un fragment de l’ensemble, prélevé pour juger de certaines propriétés de ce tout*. Il faut donc clairement exprimer de quelle propriété on veut juger avant de pouvoir concevoir un plan d'échantillonnage. La notion d’échantillon *représentatif* est des plus difficile en écologie (Frontier, 1983). L’échantillonnage est, au mieux, *adapté à tester l'hypothèse que l'on a fait, à une échelle spatiale et temporelle donnée, sur la structure ou la dynamique du système biologique étudiée*.

L'échantillonnage est la procédure par laquelle les échantillons (fragment d'un ensemble concret ou abstrait) sont prélevés. De nombreuses procédures sont possibles, chacune ayant des avantages et des inconvénients. Discuter des méthodes possibles, et arrêter des choix, est donc un élément stratégique essentiel dans une démarche scientifique descriptive, c'est pourquoi la notion de *stratégie d'échantillonnage* est si importante.

1.2. Place de l'échantillonnage dans la démarche de recherche

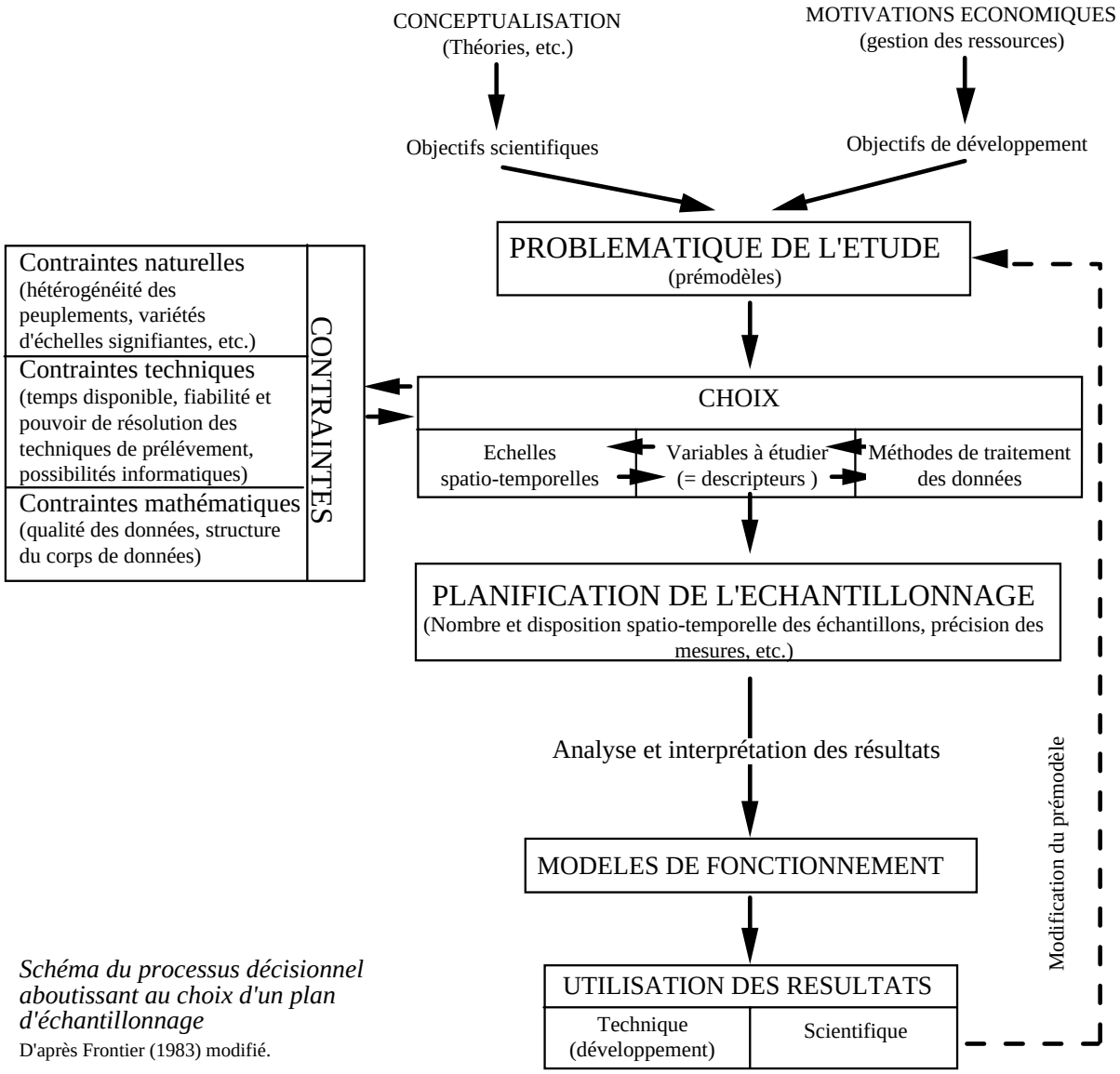


Schéma du processus décisionnel aboutissant au choix d'un plan d'échantillonnage
D'après Frontier (1983) modifié.

De fait, on échantillonne correctement uniquement, (1) si l'on sait ce que l'on va faire des données (d'où la nécessité de réfléchir à l'exploitation statistique des résultats avant de commencer l'étude), et (2) si on a compris en quoi consiste l'interaction, nommée échantillonnage, entre l'objet analysé et l'acte d'analyse (Frontier, 1983).

La notion d'échantillonnage est donc liée à celle de stratégie, qui doit assurer le meilleur compromis entre :

- l'objectif de l'étude (question/hypothèse préalablement correctement posée)
- les contraintes naturelles (hétérogénéité spatiale, variété d'échelles signifiantes, etc.)
- les contraintes techniques (temps disponible, fiabilité des mesures, etc.) financières
- les contraintes mathématiques (qualité des données et des instruments mathématiques, etc.)

Le compromis trouvé, écrit sous forme de mode opératoire, porte le nom de *plan d'échantillonnage.*

2. PROBLEMATIQUE DES CHOIX

2.1. Choix de la question biologique

- état des connaissances (théories) > problématique > hypothèse (prémodèle)

2.2. Choix du matériel biologique

- peut être une donnée de la problématique, mais si ce n'est pas le cas, toujours se poser la question de son adéquation au test de l'hypothèse...

2.3. Choix de l'élément, de l'unité d'échantillonnage et de la population statistique

2.3.1.Position du problème

- la définition de l'élément est évidente quand il s'agit d'estimer le poids moyen d'une population: élément = individu. Mais ce n'est pas toujours le cas: par exemple, si on étudie une espèce animale, et qu'on décide que l'unité d'échantillonnage sera définie par les unités de peuplement végétaux, il y a risque de référer les éléments à tel ou tel peuplement végétal mais pas à l'interface entre deux peuplements... Donc on néglige les effets de lisière, les écotones, etc... Un retour à l'hypothèse à tester permet seul de savoir si ce choix est pertinent ou non.

- la population statistique n'est pas toujours aisée à définir: cas de la mesure d'un paramètre physiologique qui nécessite de manipuler des oiseaux... La population statistique visée est-elle la population d'oiseaux OU la population d'oiseaux stressés par une manipulation? Cas des pièges d'interception ou d'attraction, etc.

2.3.2.Dépendance d'échelle

2.3.2.1. Problématique

On peut approcher la complexité du problème en utilisant ce texte de Robert Hainard :

"Imaginons un néophyte à qui on aurait attribué la tâche de réfléchir aux problèmes causés par le trafic automobile dense sur un réseau routier. D'un hélicoptère, il voit un serpent avancer, ralentir et s'arrêter longtemps à certaines heures. Il pense : ça bouge, c'est donc actif. Il s'approche et voit que le serpent est fait d'automobiles. Il pense à nouveau : chacune bouge, chacune est donc active. Il s'approche, puis il voit une caisse sur des roues et se dit : moi qui suis actif je peux la pousser. Mais il y a un moteur, ça, c'est actif. Il ouvre le capot et voit un ensemble de pièces inertes, qui se poussent passivement l'une, l'autre. Le physicien vient et dit que ce sont des corpuscules poussés par une force qui n'est qu'un nom et qu'un prochain progrès d'analyse réduira à un mécanisme mû par une force… à l'infini. Rationnellement -donc- la voiture n'a pas de moteur. Et pourtant elle se meut. Rationnellement, également, le trafic est un problème d'atomes et de forces, et pas de voitures !"

Nous laisserons la conclusion de cette réflexion sur la notion d'échelle à Frontier, qui souligne que "le choix des échelles d'observation est certainement l'un des plus

difficiles d'un plan d'échantillonnage. (…) L'appréciation des échelles signifiantes est généralement laissée à l'intuition du chercheur. Non qu'il faille sous-estimer cette dernière, qui est l'intégration (consciente ou non) de ses connaissances, de son expérience et de ses hypothèses de travail. Mais il faut malheureusement reconnaître que, dans ce qu'on appelle la "vision intuitive" d'un objet, entrent souvent pour une grande part des habitudes de pensée non remises en question et une imitation des travaux antérieurs, qui font que trop souvent un plan d'échantillonnage se ramène à l'application d'une routine."*

2.3.2.2. Sens du mot « échelle »

Echelle du géographe (*scale*)

Etendue (*range*)

Pixel, résolution

2.4. Choix des variables

- complétude: permet de décrire toutes les situations possibles...

- pertinence: ... en fonction du prémodèle

- indépendance: pas de variable redondante

- validité: ex. mesure de l'hétérogénéité, etc.

- manque à gagner: en budget limité, l'ajout de telle variable n'empêche-t-il pas la mesure d'une autre plus importante?

2.5. Choix des dispositif de mesure

1 - **justesse** : on dira qu'une méthode est juste si elle n'entraîne pas d'erreurs systématiques (les valeurs obtenues ne sont pas systématiquement surestimées ou sous-estimées).

2 - **sensibilité** : pouvoir de résolution de la méthode: la plus petite différence détectable entre deux valeurs

3 - **fidélité** : la multiplication des mesures dans des conditions sembables fournit des résultats identiques.

2.6. Choix du plan d'échantillonnage

(cf ci-dessous section 3)

2.7. Choix des estimateurs et des analyses statistiques

Estimateur: expression mathématique qui mesure, à partir des données de l'échantillon un paramètre de population statistique (cf cours biostats). La valeur obtenue est appelée estimation. En règle générale, à chaque estimation est associée un intervalle de confiance à un risque donné (sous-entendu "…que l'estimation réelle s'écarte de cet intervalle").

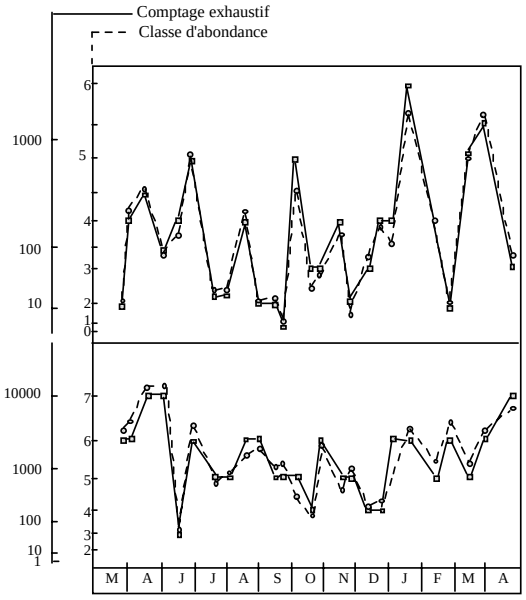
Penser à la nature des échelles de mesure utilisées (qualitative, ordinale, d'intervalle, métrique...) et aux traitements statistiques à prévoir en aval.

Toujours rationnaliser l'effort de prise de mesure à la lumière d'une analyse coût/bénéfice. Il faut se départir de l'idée que la sensibilité (précision) maximale est optimale. A quoi cela servirait-il de mesurer des effectifs de population à l'unité près si le phénomène qu'on doit mettre en évidence se manifeste par des variations d'effectif de plusieurs centaines d'individus? L'effort d'échantillonnage à consentir (au détriment d''autres activités plus utiles pour la problématique), outre qu'il est souvent hors de portée des techniques, est souvent sans commune mesure avec un illusoire gain d'information. Frontier (1969) l'illustre très bien dans une étude qu'il a réalisé sur le plancton marin à Madagascar.

Il a adopté dans une métrique log (progression géométrique de raison 4,3) une cotation par classes d'abondance illustrée dans le tableau suivant:

Classes	Effectifs par récolte	Classes	Effectifs par récolte
1	1 à 4	5	350 à 1500
2	4 à 18	6	1500 à 6500
3	18 à 80	7	6500 à 27000
4	80 à 350	…	…

"L'attribution d'un échantillon à une classe d'abondance est quasi-immmédiate au vu de l'échantillon" alors qu'un comptage exact est incomparablement plus dispendieux en temps.



Outre que les graphes ci-dessus sont déjà très parlants, "les analyse factorielles exécutées à partir des comptages exacts et à partir des cotes d'abondance aboutissent à des résultats pratiquement identiques, avec des coefficients de corrélation de 0,976 à 0,998".

3. LES PRINCIPAUX PLANS D'ECHANTILLONNAGE

3.1. Plan d'échantillonnage aléatoire simple

Consiste à prélever au hasard et de façon indépendante n unités d'échantillonnage d'une population (statistique) de N éléments". Chaque unité d'échantillonnage doit avoir la même probabilité que les autres d'être tirée.

3.2. Plan d'échantillonnage systématique

Consiste à tirer au sort le premier élément d'une série d'unité d'échantillonnage, puis à prélever systématiquement les éléments suivants selon un intervalle (ou période) convenu d'avance. Les unités d'échantillonnage ne sont donc pas prélevées de façon indépendantes.

3.3. Plan d'échantillonnage stratifié

Consiste à utiliser des sous-ensembles (= strates), mutuellement exclusifs et collectivement exhaustifs. Un échantillon indépendant est par la suite prélevé au sein de chacune des strates en appliquant un plan d'échantillonnage au choix de l'écologue. Les résultats obtenus dans chaque strate sont ensuites pondérés par la représentation de la strate dans l'échantillon.

On peut ici s'arrêter un instant pour montrer que ce plan peut permettre l'optimisation de l'effort d'échantillonnage, dans la détermination de l'effectif d'un échantillon. On peut à ce stade pratiquer de edux manières:

- **allocation proportionnelle**: on conserve la même fraction d'échantillonnage dans chaque strate. On parle d'*allocation proportionnelle* parce que l'effectif n_h de l'échantillon est proportionnel à l'effectif N_h de la strate

- **allocation optimale**: on module l'effort d'échantillonnage afin de minimiser le coût total de l'opération pour une précision donnée, ou on maximise la précision pour un coût total fixé. L'effectif n_h est d'autant plus élevé que la variance de la strate est grande, que son effectif N_h est élevé et que le coût unitaire d'échantillonnage d'une strate est faible. Si les éléments sont prélevés de façon aléatoire, l'effectif optimal de l'échantillon de la strate *h* est égal à:

$$n = \frac{n.(w_h \cdot s_h/\sqrt{c_h})}{\sum^{k}(w_h . s_h/\sqrt{c_h})}$$

avec:

$$n = \frac{(C - c_0)\sum_1^k (n_h s_h/\sqrt{c_h})}{\sum_1^k (n_h s_h \sqrt{c_h})} \quad \text{ou} \quad n = \frac{(\sum_1^k W_h s_h \sqrt{c_h})\sum_1^k (W_h s_h/\sqrt{c_h})}{V + \frac{1}{N}\sum_1^k W_h s_h^2}$$

pour un coût total C donné (formule de gauche), ou pour une variance V (précision) donnée (formule de droite). La signification des symboles donnés dans ces formules est la suivante:

n_h: effectif de l'échantillon de la strate h

c_h: coût relatif au prélévement et à la mesure d'une unité de la strate h

c_0: frais généraux ou frais fixes indépendants de l'effectif n

C: coût total de l'opération ou budget disponible ($C = c_o + \sum c_h.n_h$)

N_h: effectif de la strate h

s_h: écart-type de la variable étudiée au niveau de la strate h

W_h: poids de la strate h ($W_h = N_h/N$)

N: effectif de la population

V: variance désirée de la variable étudiée.

Un cas particulier, parfois appelé *allocation de Neyman* apparaît lorsque tous les coûts unitaires sont égaux (pour tout h, $c = c_1 = c_2 = ... = c_k$). La précision optimale est alors obtenue par la formule simplifiée:

$$n_h = \frac{(C-c_0)\ W_h.s_h}{c(\sum_k W_h.s_h)}$$

La stratgéie d'allocation optimale fournit des précisions supérieures ou au pire égales (mais à moindre coût) à l'aléatoire simple ou à l'allocation proportionnelle. Toutefois, le champ d'application des statistiques multidimensionnelles se réduit quelque peu, car il revent à un type d'échantillonnage avec propabilité proportionnelle à la taille (les strates de plus grande taille sont mieux représentées dans l'échantillon que les autres).

3.4. Exemples d'autres plans plus complexes

3.4.1.Plan d'échantillonnage par degrés

C'est une batterie de plans d'échantillonnage caractérisés par une organisation ramifiée et hiérachisée des unités. Chaque unité définie dans la population statistique de base est appellée unité primaire ou (= unité du premier degré) grappe. Chaque grappe se compose de sous-unités plus petites, aussi appelées unités secondaires (unités du deuxième degré), qui peuvent elles-même comporter des unités tertiaires, et ainsi de suite.

Les unités de chaque degré font l'objet d'un plan d'échantillonnage selon les techniques de sondage appropriées.

3.4.2.Plan d'échantillonnage avec régression

Consiste à corriger l'estimation de la moyenne d'un échantillon aléatoire d'une variable y, en fonction de la valeur d'une variable auxiliaire x. Cet ajustement repose sur la corrélation existant entre x et y.

Exemple : Scherrer (1972) a voulu estimer les réserves adipeuses de la Mésange noire (Parus ater) durant ses invasions. En 1969, année d'invasion, l'indice d'adiposité (nombre variant de 0 à 21) a été estimé sur 1529 des 9138 mésanges noires capturées au col de la Golèze (Savoie, France). L'adiposité moyenne s'élevait alors à 8,62. Parallèlement à ce travail, 51 Mésanges noires prélevées de façon aléatoire des 9138, ont été sacrifiées pour extraire et mesurer leur quantité y de lipides (en grammes) et observer leur indice x d'adiposité.La corrélation entre les deux ayant été étudié, il a pu en être déduit statistiquement sans biais la quantité moyenne de lipides et son intervalle de confiance correspondant à l'adiposité moyenne de 8,62.

4. CONCLUSION

Bien garder à l'esprit qu'il n'y a pas de "recette" applicable en toute occasion, mais qu'une bonne stratégie d'échantillonnage est avant tout un compromis optimisé entre une question et un ensemble de contraintes...